Optimal Transport for Machine Learners

Gabriel Peyré
CNRS and ENS, PSL Université
gabriel.peyre@ens.fr

July 31, 2026

Abstract

Modern machine learning repeatedly manipulates probability measures: empirical datasets, generated samples, latent distributions, class-conditional laws, particle systems, parameter distributions of wide networks and attention patterns. Optimal transport is useful in this setting because it compares such objects by asking how mass should move. It combines a statistically meaningful discrepancy with a geometry of interpolation, dual certificates and variational dynamics, making OT a common language for losses, generative modeling, domain adaptation, robust learning, barycenters, gradient flows and mean-field descriptions of learning algorithms.

This book uses optimal transport as a meeting point between probability, PDEs, optimization and statistics, with modern machine learning as the organizing pressure. In current learning systems, probability distributions are no longer peripheral objects: datasets are empirical laws, generators define push-forward laws, samplers solve evolution equations, and large models move information through populations of particles, parameters and tokens. Comparing and evolving such distributions has become a central mathematical problem, with particular urgency in generative AI. The resulting tensions are both conceptual and computational: empirical measures are singular, dimensions are large, parametrizations are non-convex, and numerical approximations cannot be separated from statistical error. The goal is to expose the OT tools that organize these tensions, while keeping their connection to the training and deployment of large models in view.

Several books already cover optimal transport from complementary viewpoints. The two-volume monograph of Rachev and Rüschendorf [445, 446] gives a broad probabilistic treatment of mass transportation and its applications. Villani's books [538, 539] are the standard references for the modern mathematical theory, from Kantorovich duality to curvature, concentration and geometric analysis. Santambrogio's text [469] offers a concise applied-mathematics route through the same foundations, with a strong emphasis on PDEs and variational arguments, and Ambrosio, Gigli and Savaré [12] develop the metric-space theory of gradient flows that underlies the dynamical part of the subject. On the computational side, Peyré and Cuturi [430] provide the reference account of numerical OT, entropic regularization and applications in data sciences; Galichon's book [225] explains the economic and matching-theoretic viewpoint; and the statistical theory of OT is developed in the recent lecture notes of Chewi, Niles-Weed and Rigollet [135]. Recent surveys complement these books by emphasizing scalable algorithms and machine-learning applications [301, 388], as well as the role of OT in imaging and graphics [75]. These references remain the natural places to find exhaustive proofs, historical details and specialized variants.

All material for this book, including the code used to reproduce the figures, is available at `gpeyre/ot4ml`. Most computational figures were produced with the Python Optimal Transport (POT) library [217]. The author warmly thanks the POT team and contributors for their important and sustained effort in making reliable optimal-transport algorithms available to the community.

Acknowledgements

This work was supported by the European Research Council through the ERC project WOLF, and by the French government under the management of Agence Nationale de la Recherche as part of the "France 2030" program, reference ANR-23-IACL-0008 (PRAIRIE-PSAI).

These notes grew out of many conversations, blackboard detours, failed proofs, numerical surprises and shared mathematical obsessions. I am deeply grateful to the students, collaborators and friends. I thank Pierre Ablin; Francisco Andrade; Raphaël Barboni; Jean-David Benamou; Mathieu Blondel; Léa Bohbot; Nicolas Bonneel; Laura Cantini; Guillaume Carlier; José Antonio Carrillo; Valérie Castin; Lénaïc Chizat; David Coeurjolly; Marco Cuturi; Gwendoline De Bie; Antonin De Courcel; Jalal Fadili; Takashi Furuya; Alexandre Gramfort; Rémi Gribonval; Matthieu Heitz; Maarten V. de Hoop; Geert-Jan Huizing; Samuel Hurault; Hicham Janati; Jérémie Kalfon; Cyril Letrouit; Flavien Leger; Sibylle Marcotte; Arthur Mensch; Thomas Moreau; Boris Muzellec; Luca Nenna; Clarice Poon; Julien Rabin; Zakaria Ramzi; Pierre Roussillon; Jules Samaran; Michael Sander; Meyer Scetbon; Othmane Sebbouh; Justin Solomon; Thibault Séjourné; Franccois-Xavier Vialard; Zhenzhang Ye; and Kelvin S. Zhang. Their ideas, taste, questions and generosity have shaped the mathematical atmosphere in which this book was written.

Finally, my deepest and most personal thanks go to my three children, Mael, Camille and Arthur, and to my love, Gwenn Guichaoua. I am lucky to share with her a passion for mathematics, computer science and biology.

Contents

Abstract i

Acknowledgements ii

1 Optimal Matching between Point Clouds 1
1.1 Monge Problem for Discrete Points . . . 1
1.2 Hungarian Algorithm . . . 7

2 Monge Problem between Measures 11
2.1 Measures . . . 11
2.2 Push-Forward . . . 14
2.3 Monge's Formulation . . . 16
2.4 Existence and Uniqueness of the Monge Map . . . 19
2.5 Beyond the Quadratic Euclidean Cost . . . 25
2.6 One-Dimensional Transport and Quantiles . . . 28
2.7 Gaussian Measures and the Bures Metric . . . 34

3 Kantorovich Relaxation 41
3.1 Discrete Relaxation . . . 41
3.2 Linear-Programming Algorithms . . . 52
3.3 Relaxation for Arbitrary Measures . . . 53
3.4 Cone of Convex Functions and Monge Gap . . . 59
3.5 c-Cyclical Monotonicity . . . 61

4 Wasserstein Space 63
4.1 Wasserstein Distances . . . 63
4.2 Topology and Applications . . . 69
4.3 Distributional Robustness and $\mathcal{W}_\infty$ . . . 73
4.4 Measure-to-Vector Maps on Wasserstein Space . . . 75
4.5 Measure-to-Measure Maps on Wasserstein Space . . . 77

5 Dual Problem 82
5.1 Discrete dual . . . 82
5.2 General formulation . . . 83
5.3 c-transforms . . . 85

6 Semi-discrete and W1 91
6.1 Semi-dual . . . 91
6.2 Auction Algorithm . . . 92
6.3 Semi-discrete . . . 96
6.4 Optimal Quantization . . . 99
6.5 W1 . . . 106

7 Divergences and Dual Norms 110
7.1 Dual norms (Integral Probability Metrics) . . . 110
7.2 Dual RKHS Norms and Maximum Mean Discrepancies . . . 112
7.3 phi-divergences . . . 114
7.4 GANs via Duality . . . 118

8 Entropic Regularization: Sinkhorn Algorithm **120**
8.1 Entropic Regularization for Discrete Measures . . . 120
8.2 Sinkhorn's Algorithm . . . 122
8.3 Reformulation Using Relative Entropy . . . 126
8.4 General Formulation . . . 129
8.5 Dual of Sinkhorn . . . 134
8.6 Path-Space Schrödinger Problem . . . 138
8.7 Marginal-Dependent Problems . . . 142
8.8 Heat Kernels and Hopf–Cole Transforms . . . 144
8.9 Other Convex Regularizers . . . 148
8.10 Sinkhorn Divergences . . . 152
8.11 Complex ε . . . 155

9 Entropic Regularization: Convergence **158**
9.1 Sinkhorn Convergence: Bregman Point of View . . . 158
9.2 Sinkhorn Convergence: Monotone Point of View . . . 162
9.3 Sinkhorn Convergence: Sublinear Robust Rate . . . 163
9.4 Sinkhorn Convergence: Linear Hilbert Metric Rate . . . 165
9.5 Local Convergence Analysis and Acceleration . . . 169
9.6 Entropic Optimal Transport between Gaussians . . . 174
9.7 Continuous ε-Sinkhorn Flow . . . 177
9.8 Monotone Clearing Beyond Variational Sinkhorn . . . 179

10 Statistical Optimal Transport **183**
10.1 Law of Large Numbers and Central Limit Theorem . . . 183
10.2 Sample Complexity . . . 187
10.3 Bias and Variance of OT . . . 197
10.4 Sketching Sinkhorn in Linear Time . . . 200

11 Generalized Wasserstein Distances **207**
11.1 Unbalanced OT . . . 207
11.2 Sliced Wasserstein Distances . . . 221
11.3 Quotient Wasserstein and Wasserstein–Procrustes . . . 231
11.4 Vector Quantiles and Linear Optimal Transport . . . 234
11.5 Spectral and Robust Wasserstein Distances . . . 236
11.6 Conditional Wasserstein Distances . . . 240

12 Generalized OT Problems **243**
12.1 OT Barycenters . . . 243
12.2 Multimarginal OT . . . 256
12.3 Low-Rank Optimal Transport . . . 263
12.4 Capacity-Constrained Optimal Transport . . . 265
12.5 Metric Learning and Inverse OT . . . 268
12.6 Weak Optimal Transport . . . 273

13 Beyond Comparing Measures **279**
13.1 Vector and Matrix-Valued Measures . . . 279
13.2 Wasserstein over Wasserstein . . . 283
13.3 Gromov–Wasserstein . . . 285
13.4 Quantum Optimal Transport . . . 307
13.5 Dynamic Time Warping . . . 311

14 Dynamic Optimal Transport **315**
14.1 Evolutions over the Space of Measures . . . 315
14.2 Benamou–Brenier dynamic formulation of OT . . . 317
14.3 Generalized Dynamic Wasserstein Distances . . . 321
14.4 Nonlocal Wasserstein Distances . . . 328
14.5 Dynamic Unbalanced Wasserstein Distances . . . 333

14.6 Variational Mean Field Games . . . 335

15 Wasserstein Gradient Flows **338**
15.1 Minimizing Movements and Wasserstein Gradients . . . 338
15.2 Geodesic Convexity and Convergence . . . 349
15.3 Functional Inequalities via Optimal Transport . . . 361
15.4 Training Two-Layer MLPs as Wasserstein Flows . . . 368
15.5 Generalized Dynamic Wasserstein Flows . . . 372
15.6 Nonlocal Wasserstein Flows . . . 377
15.7 Dynamic Unbalanced OT and WFR Flows . . . 379
15.8 Conditional Wasserstein Training of Infinite ResNets . . . 381
15.9 Second-Order Momentum Flows . . . 384

16 Generative Models via Transportation **389**
16.1 Generative Models via Flow Matching . . . 389
16.2 One-Step Generative Models . . . 397
16.3 Moment Measures . . . 403
16.4 Evolution in Depth of Transformers . . . 406
16.5 Flows over the Gaussian Manifold . . . 412

Conclusion **422**

A Notation Table **453**

B Index **461**

CHAPTER 1

Optimal Matching between Point Clouds

This opening chapter isolates the simplest form of optimal transport: pairing two finite, equally weighted point clouds of the same cardinality. The stakes are algorithmic and geometric at once: one sees the combinatorial nature of transport, the special simplicity of the line, and the limitations of permutations once cardinalities or weights differ. Classical assignment algorithms such as the Hungarian and auction methods [318, 50] provide the computational backdrop, while the restriction to equal cardinalities and weights motivates the Kantorovich relaxation.

1.1 Monge Problem for Discrete Points

This section formulates matching as Monge's deterministic transport problem on two equally weighted clouds. The one-dimensional case is a transparent reference case where the optimal map can be read off by sorting.

Assignment problem. The finite matching problem searches over bijections rather than over general mass splittings.

Definition 1.1 (Optimal assignment problem)**.** For a cost matrix $\mathrm{C} \in \mathbb{R}^{n\times n}$ and the set $\mathrm{Perm}(n)$ of bijections of $[\![n]\!] = \{1, \dots, n\}$, the optimal assignment value is

$$\min_{\sigma\in\mathrm{Perm}(n)} \frac{1}{n}\sum_{i=1}^{n} \mathrm{C}_{i,\sigma(i)}. \tag{1.1}$$

A minimizer is an optimal assignment, or optimal permutation.

Here $\mathrm{C}_{i,j}$ is the cost of pairing source i with target j. When $\mathrm{C}_{i,j} = c(x_i, y_j)$, this is the Monge problem between the empirical measures $n^{-1}\sum_i \delta_{x_i}$ and $n^{-1}\sum_j \delta_{y_j}$. The normalization by $1/n$ records the mass of each atom but does not change the minimizing permutation. Exhaustive search evaluates all $n!$ permutations and is therefore impractical; efficient methods exploit geometry or linear-programming duality. Unless additional assumptions are imposed on C, the optimizer need not be unique.

Convex costs on the line. On the line, convexity turns the order of the points into a complete optimality certificate.

Proposition 1.2 (Monotone matching on the line)**.** *Assume that $x_1, \dots, x_n$ are pairwise distinct and that $y_1, \dots, y_n$ are pairwise distinct. If $\mathrm{C}_{i,j} = h(x_i - y_j)$ for a strictly convex function $h : \mathbb{R} \to \mathbb{R}$, then the unique optimizer is the order-preserving permutation, characterized by*

$$(x_i - x_{i'})(y_{\sigma(i)} - y_{\sigma(i')}) > 0 \qquad \text{for every } i \neq i'.$$

In particular, the result applies to $h(s) = |s|^p$ for $p > 1$.

Proof. A permutation that does not preserve order contains an inversion: after relabeling, $x < x'$ are matched to $y' > y$. Set $d = y' - y > 0$ and $D(s) = \frac{h(s)-h(s-d)}{d}$. Strict convexity of h makes D strictly increasing. Therefore

$$h(x - y') + h(x' - y) - h(x - y) - h(x' - y') = d\big(D(x' - y) - D(x - y)\big) > 0.$$

Swapping the two inverted targets strictly lowers the cost. Repeating this exchange eliminates every inversion; the only order-preserving bijection between the two sorted clouds pairs equal ranks. □

If h is convex but not strictly convex, the same exchange inequality is non-strict. The equal-rank matching remains optimal, but other, non-monotone optimizers can coexist. To construct a monotone optimizer, choose sorting permutations σ_X, σ_Y such that

$$x_{\sigma_X(1)} \leqslant x_{\sigma_X(2)} \leqslant \dots \quad \text{and} \quad y_{\sigma_Y(1)} \leqslant y_{\sigma_Y(2)} \leqslant \dots$$

and match $x_{\sigma_X(k)}$ with $y_{\sigma_Y(k)}$. Thus $\sigma = \sigma_Y \circ \sigma_X^{-1}$ is optimal. A comparison sort gives a worst-case complexity $O(n \log n)$, for instance with mergesort or heapsort; standard quicksort has the same expected complexity but a quadratic worst case.

Algorithm 1.1: One-dimensional sorting assignment

Input: Equal-weight point clouds $(x_i)_{i=1}^n$, $(y_j)_{j=1}^n$ on $\mathbb{R}$; convex cost $h(x-y)$.
Output: Optimal permutation σ.
Sort source and target points: $x_{\sigma_X(1)} \leqslant \cdots \leqslant x_{\sigma_X(n)}$, $\quad y_{\sigma_Y(1)} \leqslant \cdots \leqslant y_{\sigma_Y(n)}$.
For $k = 1, \dots, n$ **do**:
  Match $x_{\sigma_X(k)}$ with $y_{\sigma_Y(k)}$.
Return $\sigma = \sigma_Y \circ \sigma_X^{-1}$.

Concave costs on the line. Concavity reverses the exchange preference: one long and one short displacement can cost less than two displacements of intermediate length. Thus costs $c(x,y) = g(|x-y|)$ with g strictly concave and nondecreasing, such as $g(r) = r^p$ for $0 < p < 1$, favor nested or crossing assignments rather than equal ranks; see Figure 1.1. This is the strictly concave regime described by Gangbo and McCann [231].

The order of the line still provides an exact algorithmic structure. For a source point, its right neighbor is the nearest target to its right such that the intervening open interval contains equally many sources and targets; left neighbors and target-to-source neighbors are defined symmetrically. Iterating this balanced-neighbor relation partitions a unit-mass problem into independent alternating chains [183]. On a balanced chain

$$p_1 < q_1 < p_2 < q_2 < \cdots < p_N < q_N,$$

with the opposite orientation handled by exchanging the roles of p and q, Delon, Salomon and Sobolevski define the local indicators

$$I_k^p(i) = c(p_i, q_{i+k}) + \sum_{r=0}^{k-1} c(p_{i+r+1}, q_{i+r}) - \sum_{r=0}^{k} c(p_{i+r}, q_{i+r}),$$
$$I_k^q(i) = c(p_{i+k+1}, q_i) + \sum_{r=1}^{k} c(p_{i+r}, q_{i+r}) - \sum_{r=0}^{k} c(p_{i+r+1}, q_{i+r}),$$

whenever the displayed indices exist. If all relevant lower-order indicators are nonnegative, a negative $I_k^p(i)$ certifies the k pairs $p_{i+r} \leftrightarrow q_{i+r-1}$, whereas a negative $I_k^q(i)$ certifies $p_{i+r} \leftrightarrow q_{i+r}$, for $r = 1, \dots, k$. Removing these certified pairs and restarting on the remaining chains gives an exact algorithm for equal unit masses. With a table that reuses previously computed indicators, its worst-case complexity is $O(n^2)$ [183]. The hypotheses that g be concave and nondecreasing are essential to this certificate.

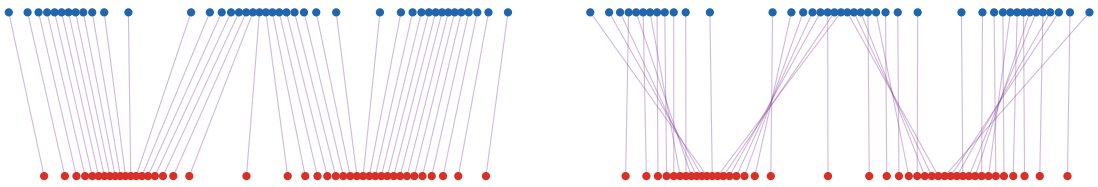
For arbitrary real masses, the stratification argument in the same work gives a larger $O(n^3)$ worst-case bound. Very concave costs also motivate the simpler greedy rule that repeatedly matches the closest red–blue pair. This rule is generally heuristic, although quantitative guarantees are known for $g(r) = r^p$ when $0 < p < 1/2$ [413]. These methods are not generic linear-programming solvers: they exploit the order and concavity specific to the line.

The resulting constructive rule is summarized as Algorithm 1.2.

Figure 1.1 contrasts this alternating-chain behavior with monotone rank matching for a convex cost on the same ordered clouds.

Algorithm 1.2: Concave line matching by local indicators

Input: Two n-point unit-mass clouds on $\mathbb{R}$; cost $c(x, y) = g(|x - y|)$ with g strictly concave and nondecreasing.
Output: Optimal concave-cost matching M.
Sort the combined source-target sequence.
Construct the independent alternating chains induced by the balanced-neighbor relation.
Initialize: Set $M = \emptyset$.
While an active chain remains **do**:
  Select the leftmost active chain and set the indicator order $k = 1$.
  While the chain is nonempty **do**:
    Retrieve or compute all admissible order-k indicators $I_k^p(i)$ and $I_k^q(i)$.
    If a negative indicator is found **then**:
      Select the negative indicator with smallest site index i; prefer $I_k^p(i)$ to break a tie.
      Add its certified block of k neighboring pairs to M.
      Remove their endpoints, invalidate affected cached indicators, relabel the chains, and set $k = 1$.
    Else if k is below the maximal admissible order **then**:
      Set $k \leftarrow k + 1$.
    Else:
      Match the remaining chain by equal indices in its current orientation and remove it.
Return M.



convex cost, $p = 2$ concave cost, $p = 1/2$

Figure 1.1: One-dimensional assignments for convex and concave costs. The same ordered source and target clouds are matched with $c_p(x, y) = |x - y|^p$. For the convex quadratic cost, equal ranks are matched and the segments do not cross. For the concave cost, the optimum instead creates long crossing exchanges governed by the alternating-chain structure.

Figure 1.2 isolates the rank-matching mechanism for both unimodal and multimodal laws: sampling the two quantile functions at the same levels produces the non-crossing optimal assignment. Section 6.4, and specifically Proposition 6.13, revisits this inverse-CDF placement for one-dimensional equal-weight quantization; there the exact quadratic optimum uses quantile-bin averages, while midpoint quantiles provide an asymptotically equivalent sampling rule.

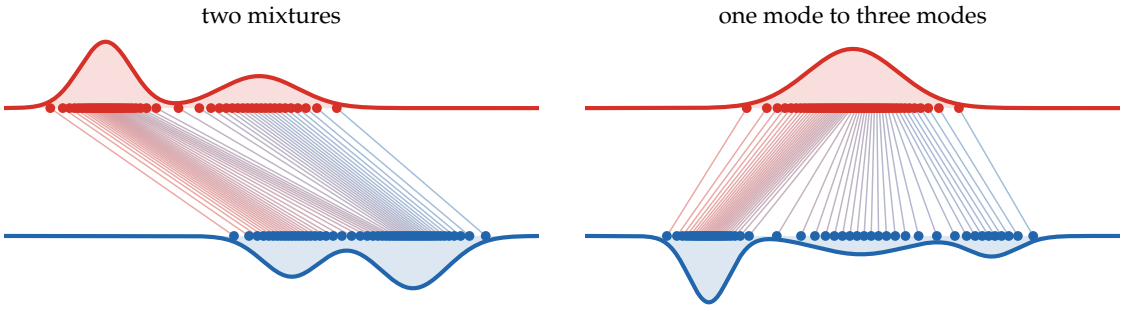


Figure 1.2: One-dimensional optimal matching by quantile sorting. The red and blue curves are smooth laws used to generate equal-weight empirical measures; the dots are inverse-CDF samples at common quantile levels. The monotone assignment connects equal ranks, both for two Gaussian mixtures and for the transport from one central Gaussian toward a three-mode target law.

Histogram equalization. Order preservation survives any strictly increasing change of coordinates. In particular, if $\varphi : \mathbb{R} \to \mathbb{R}$ is strictly increasing and $h : \mathbb{R}_+ \to \mathbb{R}$ is convex and nondecreasing, sorting solves

the assignment problem for the cost $h(|\varphi(x) - \varphi(y)|)$ after applying φ to both clouds.

A typical application is grayscale histogram equalization. The empirical distribution of image luminances is transported to a prescribed target histogram, for instance a concentrated law or the histogram of a reference image. For equal-size samples with distinct ranks, monotone rearrangement gives the exact assignment. Repeated intensity values require a consistent tie-breaking or mass-splitting convention, but the same quantile construction remains the canonical one-dimensional solution. The operation matches distributions of intensities rather than spatial pixel locations.

Figure 1.3 applies this construction to image intensities and shows simultaneously the interpolated images and the monotone evolution of their histograms.

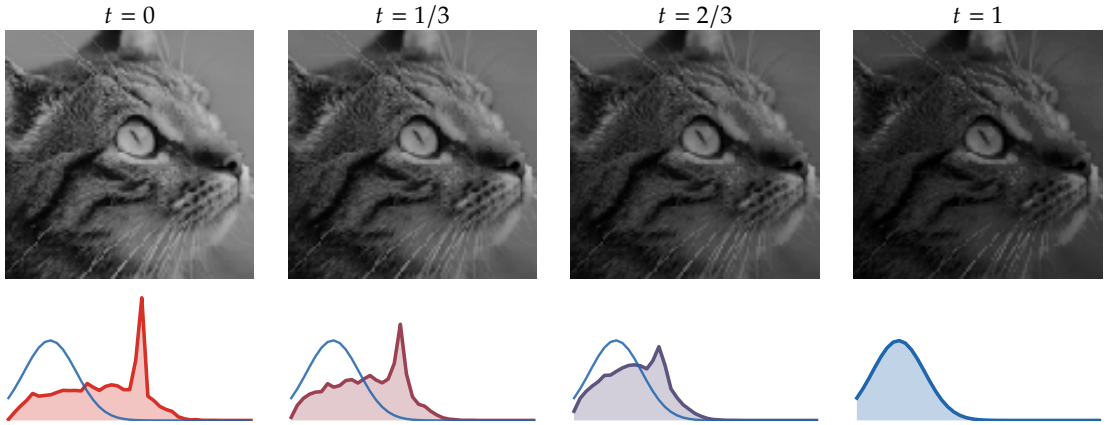


Figure 1.3: Histogram equalization as one-dimensional Monge transport on pixel intensities. The map is the monotone rearrangement $T = Q_\beta \circ F_\alpha$, using the cumulative and quantile functions defined later in Definition 2.39; here β is a truncated Gaussian concentrated near dark intensities. The images are interpolated pointwise by $I_t = (1-t)I + tT(I)$, and all histograms share the same vertical scale to make the displacement of intensity mass comparable across time.

Flat directions for the linear cost. Strict convexity makes every optimizer increasing and, for distinct points, makes the optimizer unique. For a merely convex cost such as $|x - y|$, non-increasing optimal assignments can also exist. The next example shows that this is not a discretization artifact but a genuine flat direction of the linear cost.

Example 1.3 (Discrete book-shifting and non-uniqueness for W_1). Fix $m \geqslant 1$ and consider the two equal-cardinality point clouds $X = \{1, \dots, 2m\}$ and $Y = \{m+1, \dots, 3m\}$ with uniform weights. The monotone assignment sends i to $i + m$ and has average cost m for $c(x, y) = |x - y|$. It is not unique. The discrete book-shifting assignment

$$T_{\text{book}}(i) = \begin{cases} i + 2m, & 1 \leqslant i \leqslant m, \\ i, & m < i \leqslant 2m, \end{cases}$$

is also a bijection from X to Y and has the same average cost:

$$\frac{1}{2m} \sum_{i=1}^{2m} |T_{\text{book}}(i) - i| = \frac{1}{2m} \sum_{i=1}^{m} 2m = m.$$

Optimality follows from the lower bound

$$\frac{1}{2m} \sum_{i=1}^{2m} |y_{\sigma(i)} - i| \geqslant \frac{1}{2m} \sum_{i=1}^{2m} (y_{\sigma(i)} - i) = m,$$

which holds for every assignment because the target sum is fixed. The continuous Monge version of the same book-shifting phenomenon is given later in Example 2.21.

For $|x - y|^p$ with $p > 1$, this degeneracy disappears: the monotone assignment has average cost m^p, whereas the book-shifting assignment has average cost $2^{p-1}m^p > m^p$. Strict convexity penalizes concentrating all displacement on half of the points.

Optimal transport on the circle. The sorting rule on the line has a periodic analogue. Identify the circle with $\mathbb{S}^1 = \mathbb{R}/\mathbb{Z}$, let

$$d_{\mathbb{S}^1}(x,y) := \min_{k\in\mathbb{Z}} |x-y+k|, \qquad c_p(x,y) := d_{\mathbb{S}^1}(x,y)^p, \qquad p>1.$$

The only extra datum, compared with the line, is where one opens the circle. Once a cut has been chosen, the circle is unfolded into an interval and the one-dimensional monotone assignment can be used. In the discrete case, changing the cut is the same as applying a cyclic shift to one of the two circular orderings.

Proposition 1.4 (Discrete circle transport by a cut)**.** *Assume that the $2n$ points $x_1,\dots,x_n,y_1,\dots,y_n$ are pairwise distinct. Let $x_{(1)},\dots,x_{(n)}$ and $y_{(1)},\dots,y_{(n)}$ be fixed cyclic orderings, with indices understood modulo n. For $p>1$, every optimal assignment for the cost c_p is a cyclic shift*

$$x_{(k)} \longmapsto y_{(k+s)}, \qquad k\in[\![n]\!], \quad s\in\{0,\dots,n-1\}.$$

Consequently, its cost is obtained by minimizing

$$E_s := \frac{1}{n}\sum_{k=1}^{n} d_{\mathbb{S}^1}\big(x_{(k)},y_{(k+s)}\big)^p$$

over the n shifts. For each optimal assignment there is a cut $\theta\in\mathbb{S}^1\setminus(\{x_i\}_i\cup\{y_j\}_j)$, crossed by none of its shortest transport arcs, such that the lifted assignment on $(\theta,\theta+1)$ is the equal-rank matching on the line.

Proof. Fix an optimal assignment σ and let γ_i be an open shortest arc from x_i to $y_{\sigma(i)}$, with either orientation fixed in the antipodal case. The path-uncrossing lemma of Delon, Rabin and Gousseau states that if two arcs of an optimal assignment intersect, then they have the same orientation and neither is strictly contained in the other [443]. Its proof is a two-edge exchange: opposite orientations or strict containment would, by monotonicity and strict convexity of $r\mapsto r^p$, produce a strictly cheaper pairing of the four endpoints.

Suppose that the arcs cover the circle. Since each γ_i is open, every source point lies in another transport arc. Order the sources cyclically. The path-uncrossing lemma then propagates one common orientation: either $x_{(k+1)}\in\gamma_{(k)}$ for every k, or $x_{(k-1)}\in\gamma_{(k)}$ for every k. In the first case, define a new assignment by giving $x_{(k+1)}$ the target previously assigned to $x_{(k)}$; every transport arc becomes strictly shorter. The opposite orientation gives the analogous backward reassignment. Both contradict optimality. Hence some θ lies outside every transport arc.

Cut the circle at such a θ and lift it to $(\theta,\theta+1)$. Every matched geodesic avoids the cut, so its circular length equals the ordinary distance between the corresponding lifts. Proposition 1.2 now implies that the lifted matching preserves order. Returning to the circle, an order-preserving bijection between two cyclically ordered n-point sets is a cyclic shift. Minimizing over these n shifts therefore recovers the optimum. This cut argument, including extensions to convex costs and nonuniform masses, is developed in [443, 182]. □

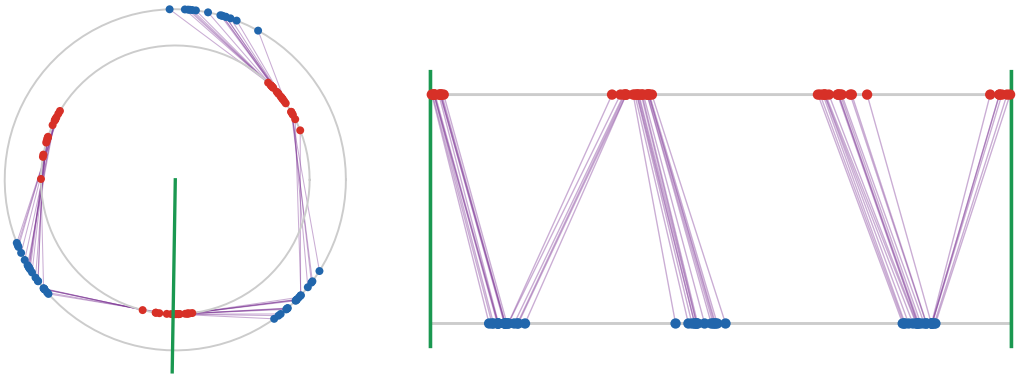
Algorithm 1.3: Circle assignment by cutting

Input: Equal-weight points $(x_i)_{i=1}^n$, $(y_j)_{j=1}^n$ on $\mathbb{S}^1$; exponent $p>1$.
Output: Optimal cyclic assignment and a compatible cut θ_{cut}.
Let $x_{(1)},\dots,x_{(n)}$ and $y_{(1)},\dots,y_{(n)}$ be the points sorted by increasing angle from a fixed origin.
For $s=0,\dots,n-1$ **do**:
  $E_s = n^{-1}\sum_{k=1}^n d_{\mathbb{S}^1}\big(x_{(k)},y_{(k+s)}\big)^p$, $\quad y_{(k+n)}=y_{(k)}$.

Set $s^\star$ to the smallest minimizer of $(E_s)_{s=0}^{n-1}$.
For each component I of $\mathbb{S}^1\setminus(\{x_i\}_i\cup\{y_j\}_j)$ in cyclic order **do**:
  Choose a test angle $\theta\in I$ and orient any antipodal matched arc to avoid θ.
  If no shortest matched arc crosses θ **then**:
    Set $\theta_{\mathrm{cut}}=\theta$ and leave the loop.

Lift every point to its representative in $(\theta_{\mathrm{cut}},\theta_{\mathrm{cut}}+1)$.
Return $x_{(k)}\mapsto y_{(k+s^\star)}$ and θ_{cut}.

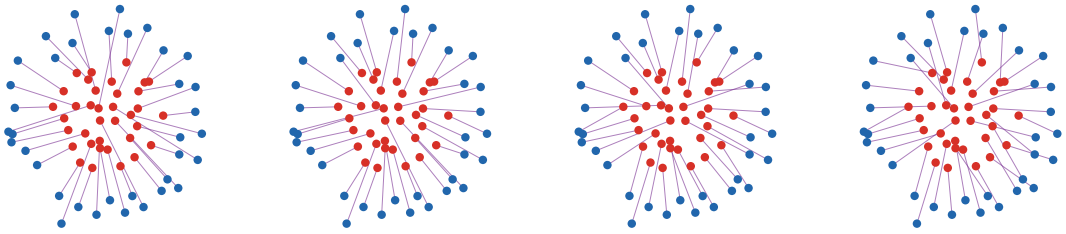
After sorting, this direct enumeration costs $O(n^2)$. The convex dependence of the circular transport cost on a continuous shift parameter leads to faster algorithms for general weighted histograms [182]. Figure 1.4 shows how the selected cut turns the circular assignment into an ordinary ordered matching on an interval.



circular problem unfolded interval

Figure 1.4: Optimal transport on the circle by cutting and unfolding. The red and blue atoms live on two copies of the circle; the denser point clouds make the cyclic ordering visible. Purple segments show the optimal matching and the green radius marks the chosen cut. Once the circle is opened at this angle, the same matching appears as a monotone one-dimensional assignment on the interval, with the two green endpoints identified.

Figure 1.5 returns to planar assignments and shows that, even for fixed point clouds, changing the exponent of the Euclidean cost can reorganize the optimal permutation globally.



$c(x,y) = \|x-y\|^{1/2}$ $c(x,y) = \|x-y\|$ $c(x,y) = \|x-y\|^2$ $c(x,y) = \|x-y\|^6$

Figure 1.5: Optimal assignments between the same two point clouds for four powers of the Euclidean distance. The source atoms are semi-regular samples in a central disk, while the target atoms are semi-regular samples on a thin annulus; this canonical geometry is reused in later coupling and regularization figures. The feasible set is unchanged, but changing p changes the global organization of the permutation: the concave case $p = 1/2$ penalizes long edges only sublinearly and therefore permits longer exchanges, whereas larger powers increasingly suppress the longest edges.

Two-dimensional geometry. Sorting no longer orders a planar cloud, but a simple local exchange still rules out proper crossings for the Euclidean-distance cost. This observation, already present in Monge's geometric reasoning, is necessary but far from sufficient for computing an optimum.

Proposition 1.5 (No proper crossings in Euclidean-distance matchings)**.** *Let $x_1, \dots, x_n, y_1, \dots, y_n \in \mathbb{R}^2$ and let $c(x,y) = \|x-y\|$. If σ is optimal, then two matched segments cannot cross properly: their relative interiors cannot meet at a point where their supporting lines are distinct.*

Proof. Suppose that $[x_i, y_{\sigma(i)}]$ and $[x_j, y_{\sigma(j)}]$ cross properly at z. The triangle inequality along the two reconnected paths gives

$$\|x_i - y_{\sigma(j)}\| \leqslant \|x_i - z\| + \|z - y_{\sigma(j)}\|, \qquad \|x_j - y_{\sigma(i)}\| \leqslant \|x_j - z\| + \|z - y_{\sigma(i)}\|.$$

The sum of the two right-hand sides is exactly the cost of the two original segments. At least one inequality is strict because a proper crossing is non-collinear. Swapping $\sigma(i)$ and $\sigma(j)$ therefore strictly

decreases the assignment cost, contradicting optimality. Collinear overlaps are excluded from the statement precisely because the exchange can then have equal cost. □

This property alone is, however, not enough to lead to an efficient algorithm. Non-crossing is only a necessary local test, not a compact certificate of optimality. For instance, if n sources and n targets are placed alternately on the boundary of a convex polygon, the number of non-crossing perfect matchings is the Catalan number

$$C_n = \frac{1}{n+1}\binom{2n}{n} \sim \frac{4^n}{\sqrt{\pi} n^{3/2}}.$$

> **Remark 1.6 (Catalan count of alternating non-crossing matchings).** The count follows from the standard Catalan recurrence. Fix one red vertex r. In a non-crossing perfect matching, if r is matched to a blue vertex b, the chord $[r, b]$ splits the polygon into two smaller polygons. Since the boundary colors alternate, each side contains the same number of red and blue vertices. If one side contains k red and k blue vertices, the other contains $n-1-k$ red and $n-1-k$ blue vertices. Non-crossing matchings on the two sides are independent, because no segment can cross the chord $[r, b]$. Thus, denoting by M_n the number of such matchings, one has $M_0 = 1$ and $M_n = \sum_{k=0}^{n-1} M_k M_{n-1-k}$. This recurrence characterizes the Catalan numbers, hence $M_n = C_n$.

Thus even after forbidding proper crossings, exhaustive search remains exponential. The two-segment swap explains why a transverse crossing cannot be optimal, but it does not select among the exponentially many planar matchings that survive this local test.

1.2 Hungarian Algorithm

We now present a first algorithm for solving the optimal matching problem (1.1). The Hungarian method [318, 90] provides a gentle, self-contained preview of duality: it computes a permutation together with a certificate of its optimality. The auction algorithm [49, 50] is postponed to Section 6.2, after discrete duality and the semi-dual formulation have been fully developed.

Hungarian primal-dual method. The method constructs an assignment and its optimality certificate simultaneously. Its derivation below uses only the elementary lower bound displayed next; Section 5.1 later embeds this certificate in general discrete Kantorovich duality.

Since the factor $1/n$ in (1.1) does not affect the optimizer, work with the unnormalized total cost. Its dual problem is

$$\max_{(\mathrm{f},\mathrm{g})\in\mathbb{R}^n\times\mathbb{R}^n} \sum_{i=1}^n \mathrm{f}_i + \sum_{j=1}^n \mathrm{g}_j \qquad \text{subject to} \qquad \mathrm{f}_i + \mathrm{g}_j \leqslant \mathrm{C}_{i,j} \quad \forall i, j. \tag{1.2}$$

The elementary certificate used by both Hungarian and auction methods is worth isolating.

Proposition 1.7 (Dual certificate for an assignment)**.** *Let* (f, g) *be feasible for* (1.2)*. Then every permutation* $\tau \in \mathrm{Perm}(n)$ *satisfies*

$$\sum_i \mathrm{C}_{i,\tau(i)} \geqslant \sum_i \mathrm{f}_i + \sum_j \mathrm{g}_j. \tag{1.3}$$

If a permutation σ *uses only tight constraints,*

$$\mathrm{f}_i + \mathrm{g}_{\sigma(i)} = \mathrm{C}_{i,\sigma(i)} \qquad \textit{for every } i, \tag{1.4}$$

then σ *is an optimal assignment and* (f, g) *is dual optimal. Equivalently, for a feasible pair, equality in* (1.3) *at* $\tau = \sigma$ *holds if and only if every edge* $(i, \sigma(i))$ *is tight.*

Proof. Dual feasibility gives $\mathrm{C}_{i,\tau(i)} \geqslant \mathrm{f}_i + \mathrm{g}_{\tau(i)}$ for every i. Summing and using that τ is a permutation yields

$$\sum_i \mathrm{C}_{i,\tau(i)} \geqslant \sum_i \mathrm{f}_i + \sum_i \mathrm{g}_{\tau(i)} = \sum_i \mathrm{f}_i + \sum_j \mathrm{g}_j,$$

which proves (1.3). If (1.4) holds, the assignment cost of σ equals the dual value of (f, g). The lower bound shows that no permutation can have smaller cost. Applied to σ and any other feasible dual pair, it also shows that no feasible dual value can exceed this cost. Thus σ and (f, g) are respectively primal and dual optimal. Finally, the difference between the two sides of (1.3) at $\tau = \sigma$ is the sum of the nonnegative slacks $\mathrm{C}_{i,\sigma(i)} - \mathrm{f}_i - \mathrm{g}_{\sigma(i)}$. It vanishes exactly when every one of them vanishes. □

Formula (1.2) is precisely the uniform-mass specialization of the discrete Kantorovich dual (5.2), multiplied by n. The method maintains such feasible potentials together with a partial matching M, namely a set of source–target pairs in which no source or target occurs twice. Define the slack, also called the reduced cost, by

$$s_{i,j} \coloneqq \mathrm{C}_{i,j} - \mathrm{f}_i - \mathrm{g}_j \geqslant 0. \tag{1.5}$$

The equality graph is the bipartite graph of zero-slack edges,

$$E(\mathrm{f},\mathrm{g}) \coloneqq \big\{(i,j)\,;\; s_{i,j} = 0\big\} = \big\{(i,j)\,;\; \mathrm{f}_i + \mathrm{g}_j = \mathrm{C}_{i,j}\big\}\,.$$

The invariant $M \subset E(\mathrm{f},\mathrm{g})$ means that every currently matched pair is already certified by equality in the dual constraint. If M is perfect, its associated permutation σ satisfies (1.4); Proposition 1.7 then certifies that σ is optimal. The algorithmic task is therefore to enlarge M while preserving dual feasibility and the inclusion $M \subset E(\mathrm{f},\mathrm{g})$.

To increase the matching, choose an unmatched source i_0 as a root. The algorithm grows a tree in the bipartite equality graph whose reached source and target vertices are denoted by S and T. From a reached source, it follows a zero-slack edge to a new target; from a reached target that is already matched, it follows the unique matched edge back to its source. Write $p(j)$ for the source preceding a reached target j, and $q(i)$ for the matched target preceding a reached non-root source i. The resulting edges alternate between unmatched and matched edges, hence the name *alternating tree*. If the tree reaches an unmatched target, these parent pointers describe an augmenting path: exchanging matched and unmatched edges along this path increases $|M|$ by one.

It remains to explain how the tree grows when no zero-slack edge currently leaves S. For a source set S, write $N_E(S) = \{j : \exists i \in S,\ (i,j) \in E(\mathrm{f},\mathrm{g})\}$ for its target neighborhood. If $N_E(S) \setminus T$ is empty, shift the potentials by the smallest slack from a reached source to an unreached target,

$$\delta = \min_{i\in S,\ j\notin T} \big(\mathrm{C}_{i,j} - \mathrm{f}_i - \mathrm{g}_j\big), \qquad \mathrm{f}_i \leftarrow \mathrm{f}_i + \delta\ (i \in S), \qquad \mathrm{g}_j \leftarrow \mathrm{g}_j - \delta\ (j \in T). \tag{1.6}$$

Here $\delta > 0$ because no such edge is currently tight. The update leaves the slack unchanged on $S \times T$, decreases it by δ on $S \times T^c$, and increases it by δ on $S^c \times T$. Feasibility is therefore preserved, all existing tree edges remain at zero slack, and at least one new zero-slack edge appears from S to T^c. Before an unmatched target is reached, the tree satisfies $|S| = |T| + 1$; hence the dual objective in (1.2) increases by $\delta(|S| - |T|) = \delta$. This is the primal-dual ascent step of the method.

For an efficient implementation, one does not recompute all the minima in (1.6). For each unreached target, store

$$\ell_j = \min_{i\in S} s_{i,j} \tag{1.7}$$

and a parent source $p(j) \in S$ attaining this minimum. When a source i' enters S, each unreached target is updated by comparing ℓ_j with the single new slack $s_{i',j}$. A dual shift uses $\delta = \min_{j\notin T} \ell_j$ and replaces ℓ_j by $\ell_j - \delta$ for $j \notin T$. Thus one tree expansion costs $O(n)$ rather than rescanning $S \times T^c$. This is what is meant below by *maintaining the slacks*. Algorithm 1.4 summarizes the primal-dual loop. Both Hungarian and auction methods are driven by the same discrete potentials and reduced costs, but they use them differently: Hungarian grows an exact zero-slack augmenting tree, whereas the auction method of Section 6.2 uses ε-relaxed contacts, price bids, and ownership changes. Figure 1.6 shows this mechanism geometrically on the planar clouds of Figure 1.5: thick violet edges form the current matching M, while thin translucent edges mark unmatched pairs with the lowest reduced costs (1.5).

Algorithm 1.4: Hungarian primal-dual augmentation

Input: Square cost matrix $C \in \mathbb{R}^{n \times n}$.
Output: Optimal permutation σ and feasible potentials (f, g) certifying its optimality.
Initialize: Set $f_i = \min_j C_{i,j}$, $g_j = 0$ for all i, j, and $M = \emptyset$.
While $|M| < n$ **do**:
  Set $i_0 = \min\{i : i \text{ is unmatched in } M\}$, $S = \{i_0\}$, and $T = \emptyset$; leave j_0 and $q(i_0)$ undefined.
  For $j = 1, \dots, n$ **do**:
    Set $\ell_j = C_{i_0,j} - f_{i_0} - g_j$ and $p(j) = i_0$.
  While j_0 is undefined **do**:
    Set $\delta = \min_{j \notin T} \ell_j$.
    Update $f_i \leftarrow f_i + \delta$ for $i \in S$, $g_j \leftarrow g_j - \delta$ for $j \in T$, and $\ell_j \leftarrow \ell_j - \delta$ for $j \notin T$.
    Select $j = \min\{k \notin T : \ell_k = 0\}$ and set $T \leftarrow T \cup \{j\}$.
    If j is unmatched in M **then**:
      Set $j_0 = j$ and exit the inner while loop.
    Otherwise set i' to the source matched to j, $S \leftarrow S \cup \{i'\}$, and $q(i') = j$.
    For $k \notin T$ **do**:
      Set $r = C_{i',k} - f_{i'} - g_k$.
      If $r < \ell_k$ **then set** $\ell_k = r$ and $p(k) = i'$.
  Set $j = j_0$.
  While j is defined **do**:
    Set $i = p(j)$ and $j_{\text{old}} = q(i)$.
    Delete (i, j_{old}) from M when j_{old} is defined.
    Set $M \leftarrow M \cup \{(i, j)\}$ and $j \leftarrow j_{\text{old}}$.
Return the permutation σ defined by $M = \{(i, \sigma(i))\}_i$ and the potentials (f, g).

The snapshots in Figure 1.6 are taken at initialization and after 9, 18, 27, and 36 augmentations. The faint candidate graph changes with the dual potentials, whereas every thick matching edge remains exactly dual-tight.

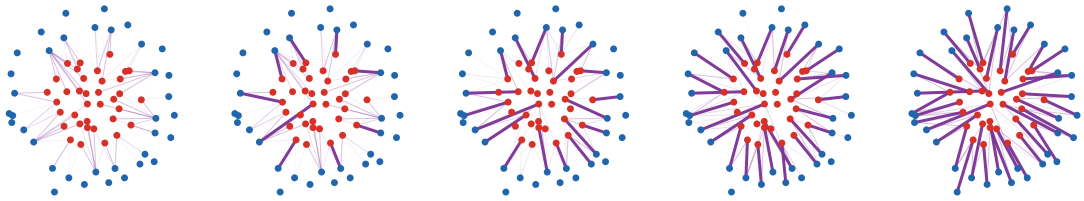


Figure 1.6: Growth of the Hungarian matching on the planar point clouds of Figure 1.5. Thick violet segments show the current partial matching M, whose cardinality is reported below each panel. Thin translucent segments show the $2n$ lowest-slack unmatched edges, ranked by $s_{i,j} = C_{i,j} - f_i - g_j$; exact zero-slack candidates are slightly stronger. This overlay summarizes the near-tight dual constraints, while the algorithm selects its next edge only across the current alternating-tree cut. In the final panel, all thick edges are dual-tight and form an optimal squared-distance assignment.

Proposition 1.8 (Correctness and complexity of the Hungarian primal-dual method)**.** *For every cost matrix $C \in \mathbb{R}^{n \times n}$, Algorithm 1.4 terminates after n augmentation phases. It returns a permutation $\sigma \in \mathrm{Perm}(n)$ and feasible potentials (f, g) such that*

$$f_i + g_{\sigma(i)} = C_{i,\sigma(i)} \qquad \forall i.$$

Consequently, σ solves (1.1). With the maintained slacks (1.7), the algorithm uses $O(n^3)$ arithmetic and comparison operations and $O(n^2)$ storage, including the cost matrix.

Proof. Initialization is dual feasible because $f_i = \min_j C_{i,j}$ and $g = 0$, and the empty matching is contained in the equality graph. We prove inductively that feasibility and $M \subset E(f, g)$ hold throughout the algorithm.

At the start of a phase, the initialization of ℓ_j and $p(j)$ gives

$$\ell_j = \min_{i \in S} s_{i,j} = s_{p(j),j} \qquad (j \notin T).$$

This identity remains true when a matched source i' enters S, because the algorithm compares every ℓ_j with the one new candidate $s_{i',j}$.

Consider an inner iteration before a free target is reached. Every target in T is matched to a source in S, and every source in S except the unmatched root is reached through its matched target. Hence $|S| = |T| + 1$, so $T \neq [\![n]\!]$. Dual feasibility gives $\ell_j \geqslant 0$, and therefore

$$\delta = \min_{j \notin T} \ell_j = \min_{i \in S,\, j \notin T} s_{i,j} \geqslant 0.$$

After the potential update, the slack becomes $s^+_{i,j} = s_{i,j} - \delta$ on $S \times T^c$, $s^+_{i,j} = s_{i,j} + \delta$ on $S^c \times T$, and remains $s^+_{i,j} = s_{i,j}$ on the other two blocks. The definition of δ makes every entry nonnegative. For $j \notin T$, all slacks from S decrease by the same amount δ, exactly as does ℓ_j; hence the maintained-minimum identity remains valid. At least one maintained slack becomes zero, and its parent $p(j)$ identifies a corresponding zero-slack edge from S to an unreached target. Thus dual feasibility is preserved and the next target selected by the algorithm is connected to the alternating tree. Moreover, every matched edge either has both endpoints in $S \times T$ or both endpoints outside the tree, so all matched edges remain tight.

Each inner iteration adds a new target to T. If that target is matched, its unique partner is added to S, and the slack update restores the maintained-minimum identity above. If it is unmatched, the pointers p and q trace an alternating path from i_0 to this target. Flipping the status of the edges on that path preserves the matching property, increases $|M|$ by one, and uses only equality edges. Since a target enters T at most once, every phase reaches a free target after at most n inner iterations. Starting from $M = \emptyset$, exactly n phases therefore produce a perfect matching $M = \{(i, \sigma(i))\}_i$.

At termination, every matched edge is tight, so Proposition 1.7 directly proves that σ is optimal and (f, g) is dual optimal. This is the finite assignment instance of the complementary-slackness mechanism proved in general in Proposition 5.3. Finally, initializing the slack and parent arrays costs $O(n)$ per phase. Each inner iteration scans the unreached targets to find δ, updates the potentials and maintained slacks, and, when necessary, compares those slacks with one newly reached source row. These operations cost $O(n)$ per iteration. There are at most n iterations per phase and n phases, giving $O(n^3)$ arithmetic and comparison operations. The cost matrix uses $O(n^2)$ storage; the matching, reached sets, parents, and maintained slacks require only $O(n)$ additional storage. □

The chapter has thus exposed two complementary routes to finite matching: geometric structure can reduce the problem to sorting, while primal-dual labels provide a global certificate for an arbitrary cost matrix. The next chapter passes from finite permutations to transport maps between general measures.

CHAPTER 2

Monge Problem between Measures

The goal of this chapter is to pass from finite matching to transport between arbitrary probability laws. The central stakes are to define measures, push-forwards and Monge maps carefully enough that the discrete picture survives, while exposing why deterministic maps can fail to exist. Monge's original formulation [387] and modern treatments [538, 539, 469, 445] are the conceptual background for this transition.

The strict assignment model of the previous chapter handles uniform point clouds of equal cardinality. Passing to arbitrary measures allows unequal weights, continuous densities and singular distributions to be treated within one language.

2.1 Measures

Measures are the language that lets point clouds, densities and singular objects be handled uniformly. We only recall the facts needed later: integration, total variation, densities and probabilistic laws.

Histograms The finite-dimensional model for a probability law is a nonnegative vector with unit total mass.

Definition 2.1 (Probability simplex)**.** The probability simplex of length n is

$$\Sigma_n := \left\{ \mathrm{a} \in \mathbb{R}_+^n \; ; \; \sum_{i=1}^n \mathrm{a}_i = 1 \right\}.$$

Its elements are also called probability vectors or histograms.

Discrete measure, empirical measure Probability vectors become measures once their masses are attached to locations.

Definition 2.2 (Discrete measure)**.** A discrete measure with weights a and locations $x_1, \dots, x_n \in \mathcal{X}$ is

$$\alpha = \sum_{i=1}^n \mathrm{a}_i \delta_{x_i}, \tag{2.1}$$

where δ_x is the Dirac mass at position x. It is a probability measure if $\mathrm{a} \in \Sigma_n$, and a positive measure if all weights a_i are nonnegative.

The Dirac mass should be thought of as a unit of mass infinitely concentrated at one location. The empirical probability distribution of a point cloud is $\alpha = \frac{1}{n} \sum_i \delta_{x_i}$. Many applications manipulate both the positions x_i (a Lagrangian discretization) and the weights a_i (an Eulerian discretization). Moving the positions gives an adaptive representation, but usually destroys convexity; changing only the weights often preserves it.

General measures We consider finite signed Borel measures $\alpha \in \mathcal{M}(\mathcal{X})$ on a metric space $(\mathcal{X}, d)$. The Borel sets form the smallest σ-algebra containing all open subsets of $\mathcal{X}$, obtained by closing the open sets under complements and countable unions. Unless otherwise stated, all measures are finite. A Dirac measure δ_x is then defined as $\delta_x(A) = 1$ if $x \in A$ and 0 otherwise, and this extends by linearity for discrete measures of the form (2.1) as

$$\alpha(A) = \sum_{x_i \in A} \mathrm{a}_i$$

Positive and probability measures are formalized below together with the Radon regularity used throughout the book.

Polish metric spaces. Many measure-theoretic statements used later require a mild regularity assumption on the underlying space. The point is not to restrict applications, since Euclidean spaces, complete separable manifolds and separable Hilbert spaces are covered, but to exclude pathological measurable spaces where disintegration, tightness or weak convergence can fail to behave properly.

Definition 2.3 (Polish metric space)**.** A metric space $(\mathcal{X}, d)$ is Polish if it is complete and separable: every Cauchy sequence converges in $\mathcal{X}$, and $\mathcal{X}$ contains a countable dense subset. More generally, a topological space is called Polish if its topology can be induced by some complete separable metric.

This assumption already includes genuinely infinite-dimensional spaces. In the Euclidean dynamic OT and Schrödinger formulations of Sections 14.2 and 8.6, the path space

$$\mathcal{S} = C([0,1]; \mathbb{R}^d), \qquad d_\infty(\gamma, \eta) := \sup_{t\in[0,1]} \|\gamma(t) - \eta(t)\|$$

is Polish. Completeness follows because a uniformly Cauchy sequence of continuous paths converges uniformly to a continuous path, and separability follows by approximating paths uniformly by piecewise-linear paths with rational breakpoints and rational values. This path space supports laws of trajectories: its endpoint evaluation maps are continuous, which makes endpoint constraints well behaved for dynamical optimal plans and Schrödinger bridges. Polish spaces are the natural ambient category for probability measures. Borel probability measures on them are regular, tightness gives compactness criteria, regular conditional probabilities and disintegrations exist under standard assumptions, and Wasserstein spaces remain Polish; see Proposition 13.8.

Definition 2.4 (Support of a measure)**.** Let α be a positive Borel measure on a metric space $\mathcal{X}$. Its support is

$$\operatorname{supp}(\alpha) := \left\{x \in \mathcal{X} \;;\; \alpha(B(x,r)) > 0 \text{ for every } r > 0\right\}.$$

If $\mathcal{X}$ is separable, this is the smallest closed set of full α-mass.

Radon measures Radon regularity ensures that Borel measures can be approximated from within by compact sets and paired reliably with test functions.

Definition 2.5 (Radon and probability measures)**.** A locally finite positive Borel measure α on $\mathcal{X}$ is Radon if

$$\alpha(A) = \sup\left\{\alpha(K) \;;\; K \subset A \text{ compact}\right\}$$

for every Borel set A. We write $\mathcal{M}_+(\mathcal{X})$ for finite positive Radon measures and $\mathcal{M}_+^1(\mathcal{X}) = \{\alpha \in \mathcal{M}_+(\mathcal{X}) \;;\; \alpha(\mathcal{X}) = 1\}$ for probability measures.

Every finite Borel measure on a Polish space is Radon. For an integrable function f, we use the pairing notation

$$\langle f, \alpha\rangle := \int_{\mathcal{X}} f(x)\mathrm{d}\alpha(x).$$

Integration of such a measurable f against a discrete measure α computes a sum

$$\int_{\mathcal{X}} f(x)\mathrm{d}\alpha(x) = \sum_{i=1}^{n} \mathrm{a}_i f(x_i).$$

This applies in particular to continuous test functions, which are Borel measurable. Integration against a finite measure on a compact space thus defines a continuous linear form $f \mapsto \int f\mathrm{d}\alpha$ on the Banach space of continuous functions $(C(\mathcal{X}), \|\cdot\|_\infty)$, indeed $|\int f\mathrm{d}\alpha| \leqslant \|f\|_\infty |\alpha|(\mathcal{X})$. On compact spaces, the converse is true: every continuous linear form on $C(\mathcal{X})$ is represented by integration against a finite signed Radon measure. This is the Riesz–Markov–Kakutani representation theorem [459, 69]. It identifies $\mathcal{M}(\mathcal{X})$ with the Banach dual of $C(\mathcal{X})$, and this duality pairing $\langle f, \alpha\rangle$ will be crucial for the convex duality arguments used later.

Relative densities. Many formulas below compare measures through densities with respect to a reference measure.

Definition 2.6 (Relative density)**.** If α is absolutely continuous with respect to a positive reference measure λ, its relative density is the Radon–Nikodym derivative

$$\rho_\alpha := \frac{\mathrm{d}\alpha}{\mathrm{d}\lambda}, \qquad \mathrm{d}\alpha(x) = \rho_\alpha(x)\mathrm{d}\lambda(x).$$

Equivalently, for every nonnegative measurable function h, and hence for every integrable h,

$$\int_{\mathcal{X}} h(x)\mathrm{d}\alpha(x) = \int_{\mathcal{X}} h(x)\rho_\alpha(x)\mathrm{d}\lambda(x).$$

On $\mathbb{R}^d$ the reference λ is often Lebesgue measure $\mathrm{d}x$.

Total variation norm. The norm inherited from the duality $\mathcal{M}(\mathcal{X}) = C(\mathcal{X})^*$ is the total variation norm. We use the notation

Definition 2.7 (Total variation)**.** For a finite signed Radon measure α on a compact space $\mathcal{X}$,

$$\|\alpha\|_{\mathrm{TV}} := \sup_{f \in C(\mathcal{X})} \{\langle f, \alpha\rangle \,;\, \|f\|_\infty \leqslant 1\}.$$

This formula is useful because it computes the norm of a measure as the operator norm of the corresponding linear form on $C(\mathcal{X})$.

The same norm also has a direct measure-theoretic expression. The absolute value of a signed measure is

$$|\alpha|(A) := \sup_{A=\cup_i B_i} \sum_i |\alpha(B_i)|,$$

where the supremum is over finite or countable measurable partitions of A. Thus, if $\alpha = \sum_i \mathrm{a}_i \delta_{x_i}$ with distinct atoms, $|\alpha| = \sum_i |\mathrm{a}_i|\delta_{x_i}$; if $\mathrm{d}\alpha(x) = \rho(x)\mathrm{d}\lambda(x)$, then $\mathrm{d}|\alpha|(x) = |\rho(x)|\mathrm{d}\lambda(x)$.

Proposition 2.8 (Dual and measure definitions of total variation)**.** *For a finite signed Radon measure α on a compact space $\mathcal{X}$, one has $\|\alpha\|_{\mathrm{TV}} = |\alpha|(\mathcal{X})$. Consequently $(\mathcal{M}(\mathcal{X}), \|\cdot\|_{\mathrm{TV}})$ is isometrically the Banach dual of $(C(\mathcal{X}), \|\cdot\|_\infty)$.*

Proof. The inequality $\|\alpha\|_{\mathrm{TV}} \leqslant |\alpha|(\mathcal{X})$ follows from

$$\left|\int f \,\mathrm{d}\alpha\right| \leqslant \int |f|\,\mathrm{d}|\alpha| \leqslant \|f\|_\infty |\alpha|(\mathcal{X}).$$

For the reverse inequality, write the Jordan decomposition $\alpha = \alpha^+ - \alpha^-$, so that $|\alpha| = \alpha^+ + \alpha^-$. The measurable sign $s = \frac{\mathrm{d}\alpha}{\mathrm{d}|\alpha|}$ takes values in $\{-1, 1\}$ outside a $|\alpha|$-null set and satisfies $\mathrm{d}\alpha = s\,\mathrm{d}|\alpha|$. By regularity of Radon measures on compact spaces, s can be approximated in $L^1(|\alpha|)$ by continuous functions f_k with $\|f_k\|_\infty \leqslant 1$. Hence $\int f_k \,\mathrm{d}\alpha \to \int s \,\mathrm{d}\alpha = |\alpha|(\mathcal{X})$, which proves the equality. The final statement is the Riesz–Markov–Kakutani representation theorem with this norm. □

For two absolutely continuous measures $\mathrm{d}\alpha = \rho_\alpha \mathrm{d}\lambda$ and $\mathrm{d}\beta = \rho_\beta \mathrm{d}\lambda$, this gives the concrete formula

$$\|\alpha - \beta\|_{\mathrm{TV}} = \int_{\mathcal{X}} |\rho_\alpha(x) - \rho_\beta(x)|\,\mathrm{d}\lambda(x).$$

For two discrete measures, one first recasts both measures as weight vectors on the same finite support. If

$$\alpha = \sum_{i=1}^n \tilde{a}_i \delta_{x_i}, \qquad \beta = \sum_{j=1}^m \tilde{b}_j \delta_{y_j},$$

let $z_1, \ldots, z_r$ be the distinct points in the union $\{x_i\}_i \cup \{y_j\}_j$ and set

$$a_k := \sum_{i : x_i = z_k} \tilde{a}_i, \qquad b_k := \sum_{j : y_j = z_k} \tilde{b}_j.$$

Then $\alpha = \sum_{k=1}^{r} a_k \delta_{z_k}$ and $\beta = \sum_{k=1}^{r} b_k \delta_{z_k}$. This convention merges masses located at the same point before comparing the two measures. If the two input supports are already enumerated without repetitions and are disjoint, this simply amounts to taking $(z_k)_k = (x_1, \dots, x_n, y_1, \dots, y_m)$ and padding the weights as $a = (\tilde{a}, 0_m)$ and $b = (0_n, \tilde{b})$. With this common-support notation,

$$\|\alpha - \beta\|_{\mathrm{TV}} = \sum_{k=1}^{r} |a_k - b_k|.$$

Probabilistic interpretation. Probability measures represent random observations through their laws.

Definition 2.9 (Random variable and law)**.** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $\mathcal{X}$ be Polish. An $\mathcal{X}$-valued random variable is a measurable map $X : (\Omega, \mathcal{F}) \to (\mathcal{X}, \mathcal{B}(\mathcal{X}))$. Its law is

$$\mathrm{Law}(X) := X_\sharp \mathbb{P}, \qquad \mathrm{Law}(X)(A) = \mathbb{P}(X \in A)$$

for every Borel set $A \subset \mathcal{X}$.

For every integrable function f, integration with respect to this law is expectation:

$$\int_{\mathcal{X}} f(x) \, \mathrm{d}\,\mathrm{Law}(X)(x) = \mathbb{E}[f(X)].$$

2.2 Push-Forward

Push-forwards encode how maps move mass. This short section is the bridge between deterministic maps and linear operations on measures.

For a measurable map $T : \mathcal{X} \to \mathcal{Y}$, the push-forward operator $T_\sharp : \mathcal{M}(\mathcal{X}) \to \mathcal{M}(\mathcal{Y})$ records the distribution of the image points. For a Dirac mass, one has $T_\sharp \delta_x = \delta_{T(x)}$. The operation $\alpha \mapsto T_\sharp \alpha$ is linear in the measure, although its definition for a general measure uses inverse images rather than a decomposition into Dirac masses. In this sense, moving from T to $T_\sharp$ linearizes the action of a map at the price of passing from a possibly finite-dimensional space $\mathcal{X}$ to the infinite-dimensional space $\mathcal{M}(\mathcal{X})$, an idea that underlies many convex relaxations. For discrete measures (2.1), the push-forward operation consists simply in moving the positions of all the points in the support of the measure

$$T_\sharp \alpha := \sum_i \mathrm{a}_i \delta_{T(x_i)}.$$

For more general measures, for instance for those with a density, the notion of push-forward plays a fundamental role in describing spatial modifications of probability measures. The formal definition reads as follows.

Definition 2.10 (Push-forward)**.** For a measurable map $T : \mathcal{X} \to \mathcal{Y}$ and $\alpha \in \mathcal{M}(\mathcal{X})$, the push-forward measure $\beta = T_\sharp \alpha \in \mathcal{M}(\mathcal{Y})$ is defined, for every measurable set $B \subset \mathcal{Y}$, by

$$\beta(B) = \alpha\big(\{x \in \mathcal{X} \,;\, T(x) \in B\}\big). \tag{2.2}$$

Equivalently, for every bounded measurable function $h : \mathcal{Y} \to \mathbb{R}$,

$$\int_{\mathcal{Y}} h(y) \mathrm{d}\beta(y) = \int_{\mathcal{X}} h(T(x)) \mathrm{d}\alpha(x). \tag{2.3}$$

Note that $T_\sharp$ preserves positivity and total mass, so that if $\alpha \in \mathcal{M}_+^1(\mathcal{X})$ then $T_\sharp \alpha \in \mathcal{M}_+^1(\mathcal{Y})$.

Remark 2.11 (Pullback and push-forward). If $T : \mathcal{X} \to \mathcal{Y}$ is continuous, the pullback by T is the linear operator

$$T^\sharp : C(\mathcal{Y}) \to C(\mathcal{X}), \qquad T^\sharp g = g \circ T.$$

The definition of the push-forward is exactly the dual relation between this pullback on functions and

the action of $T_\sharp$ on measures:

$$\int_{\mathcal{X}} T^\sharp g(x)\mathrm{d}\alpha(x) = \int_{\mathcal{Y}} g(y)\mathrm{d}(T_\sharp\alpha)(y).$$

In pairing notation,

$$\left\langle T^\sharp g, \alpha \right\rangle_{C(\mathcal{X}),\mathcal{M}(\mathcal{X})} = \left\langle g, T_\sharp\alpha \right\rangle_{C(\mathcal{Y}),\mathcal{M}(\mathcal{Y})}.$$

Thus push-forward is the adjoint operation to pullback, with the direction reversed. The two arrows should not be confused: $T_\sharp$ transports mass from $\mathcal{X}$ to $\mathcal{Y}$, whereas $T^\sharp$ transports test functions from $\mathcal{Y}$ back to $\mathcal{X}$.

Proposition 2.12 (Push-forward formula for densities)**.** *Let $U, V \subset \mathbb{R}^d$ be open, let α and β have Lebesgue densities ρ_α and ρ_β on U and V, and let $T : U \to V$ be a C^1 diffeomorphism such that $\beta = T_\sharp\alpha$. Then, for Lebesgue-almost every $x \in U$,*

$$\rho_\alpha(x) = |\det(T'(x))|\rho_\beta(T(x)). \tag{2.4}$$

Equivalently, for Lebesgue-almost every $y \in V$,

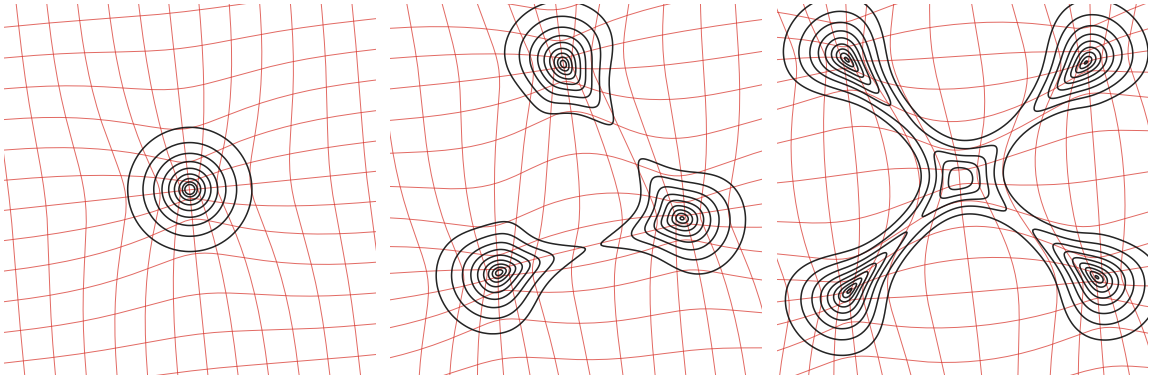
$$\rho_\beta(y) = \rho_\alpha(T^{-1}y)\,\frac{1}{|\det(T'(T^{-1}y))|}.$$

Proof. Apply (2.3) to an arbitrary compactly supported continuous test function h and use the change of variables $y = T(x)$, for which $\mathrm{d}y = |\det(T'(x))|\mathrm{d}x$. This gives

$$\begin{aligned}\int_V h(y)\rho_\beta(y)\mathrm{d}y = \int_V h(y)\mathrm{d}\beta(y) = \int_U h(T(x))\mathrm{d}\alpha(x) = \int_U h(T(x))\rho_\alpha(x)\mathrm{d}x \\ = \int_V h(y)\rho_\alpha(T^{-1}y)\frac{\mathrm{d}y}{|\det(T'(T^{-1}y))|},\end{aligned}$$

Since the identity holds for every such h, uniqueness of Lebesgue densities gives the target-variable formula in the statement. Substituting $y = T(x)$ then yields (2.4), where $T'(x) \in \mathbb{R}^{d\times d}$ is the Jacobian matrix of T. □

Figure 2.1 makes the determinant factor visible. Even when the source density is uniform, nonlinear diffeomorphisms can compress area around one or several centers and expand it elsewhere; the target density is therefore not obtained by the naive composition $\rho_\alpha \circ T^{-1}$ alone. The high-density bumps appear exactly where the deformed grid cells have small area.



one compression bump | three compression bumps | five compression bumps

Figure 2.1: Jacobian determinant in the density push-forward formula. The three panels show smooth diffeomorphisms that strongly compress a uniform source grid around one, three, and five centers. The deformed grid is drawn as a dense red mesh, and the pushed density is shown through black level sets of $\rho_\beta(y) = |\det T'(T^{-1}y)|^{-1}$. The tighter crop focuses on the bump region, where small grid cells coincide with large values of the pushed density.

Remark 2.13 (Probabilistic interpretation). Let $X : \Omega \to \mathcal{X}$ be a random variable defined on an abstract probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Its law, or probability distribution, is the push-forward of $\mathbb{P}$

by X, namely $\alpha = X_\sharp \mathbb{P}$. Applying another push-forward $\beta = T_\sharp \alpha$ for $T : \mathcal{X} \to \mathcal{Y}$, following (2.3), is equivalent to defining another random variable $Y = T(X)$, namely $\omega \in \Omega \mapsto T(X(\omega)) \in \mathcal{Y}$. Indeed

$$\beta = T_\sharp \alpha = T_\sharp (X_\sharp \mathbb{P}) = (T \circ X)_\sharp \mathbb{P},$$

so β is the law of Y. Thus, if x is sampled according to the law α, then $y = T(x)$ is sampled according to the law β.

2.3 Monge's Formulation

Monge's problem asks for a deterministic map transporting one law onto another while minimizing a prescribed cost. This is the original formulation introduced by Monge in his memoir on the "déblai et remblai" problem [387]. It is geometrically direct, because every source point is assigned one destination, but it is also analytically fragile: the feasible set is non-convex, it can be empty, and a map cannot split mass. These limitations are precisely what motivate Kantorovich's relaxation in the next chapter.

Monge problem. The deterministic transport constraint leads to the following optimization problem.

Definition 2.14 (Monge problem and Monge map)**.** For $\alpha \in \mathcal{M}_+^1(\mathcal{X})$, $\beta \in \mathcal{M}_+^1(\mathcal{Y})$ and measurable $c : \mathcal{X} \times \mathcal{Y} \to [0, +\infty]$, define

$$\mathcal{M}_c(\alpha, \beta) \coloneqq \inf_{T : \mathcal{X} \to \mathcal{Y} \text{ measurable}} \left\{ \int_{\mathcal{X}} c(x, T(x)) \mathrm{d}\alpha(x) \;;\; T_\sharp \alpha = \beta \right\}. \tag{2.5}$$

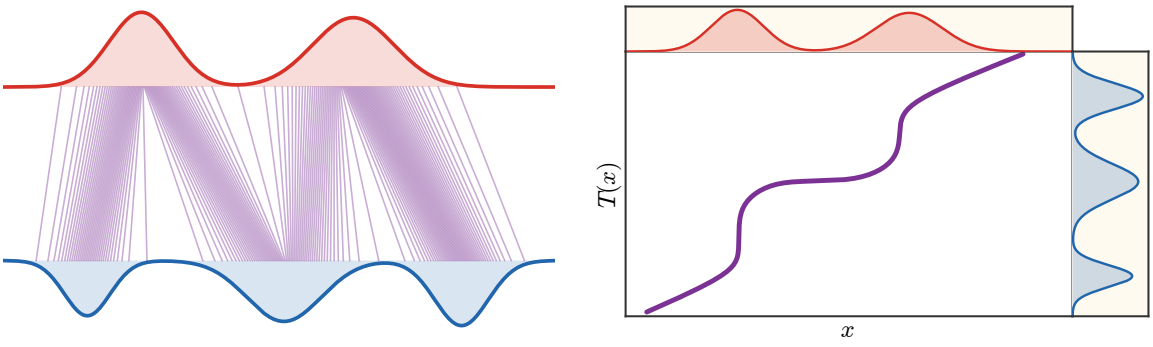
A feasible T is a Monge map; a minimizer is an optimal Monge map. The value is $+\infty$ if no feasible map exists.

The constraint $T_\sharp \alpha = \beta$ means that T pushes the mass of α onto β in the sense of Definition 2.10.

A diffuse one-dimensional example makes the deterministic constraint visible without reducing the measures to point clouds. For the quadratic cost, the optimal map between an atomless source and an arbitrary target is the monotone rearrangement

$$T = C_\beta^{-1} \circ C_\alpha,$$

as proved in Theorem 2.41. Figure 2.2 applies this formula to a two-component Gaussian mixture and a three-component Gaussian mixture. The left panel follows equal quantile levels from source to target; the right panel displays the resulting map directly.



equal-quantile transport Monge map $y = T(x)$

Figure 2.2: A one-dimensional Monge map between Gaussian mixtures. The red source is a two-component mixture and the blue target is a three-component mixture. The left panel draws a dense family of equal-quantile transport rays without empirical atoms. The right panel shows the same monotone rearrangement as the graph of $T = C_\beta^{-1} \circ C_\alpha$, framed by the source and target densities. Its changes of slope encode the density balance $\rho_\alpha(x) = T'(x)\rho_\beta(T(x))$ from Proposition 2.12.

Proposition 2.15 (Empirical Monge maps and matchings)**.** *Assume that the source atoms $x_1, \dots, x_n$ are distinct and that*

$$\alpha = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i}, \qquad \beta = \frac{1}{n}\sum_{j=1}^{n} \delta_{y_j}.$$

If $T_\sharp \alpha = \beta$, then for each distinct target value z in the support of β, exactly $n\beta(\{z\})$ source atoms are mapped to z. In particular, if the y_j are distinct, then there exists a permutation $\sigma \in \mathrm{Perm}(n)$ such that $T(x_i) = y_{\sigma(i)}$ for all i. Conversely, every such assignment of source atoms to target atoms with the correct masses defines a feasible Monge map on the support of α, and in the distinct-target case

$$\int_{\mathcal{X}} c(x, T(x)) \mathrm{d}\alpha(x) = \frac{1}{n}\sum_{i=1}^{n} c(x_i, y_{\sigma(i)}).$$

The values of a feasible map away from $\{x_1, \dots, x_n\}$ may be chosen arbitrarily, provided the resulting extension is measurable. If source locations are repeated, they should first be merged into atoms with larger masses; such atoms cannot be split by a Monge map.

Proof. Since $T_\sharp \alpha = \beta$, all images $T(x_i)$ must belong to the support of β; otherwise the push-forward would give positive mass to a point outside that support. For any target atom z,

$$\beta(\{z\}) = \alpha(T^{-1}(\{z\})) = \frac{1}{n}\#\{i \,;\, T(x_i) = z\}\,.$$

This proves the counting statement. If all target atoms have mass $1/n$, each target receives exactly one source atom, which is a permutation. The converse and the cost identity follow by direct substitution. □

Proposition 2.16 (Existence of transport maps from atomless sources)**.** *Let α and β be Borel probability measures on Polish spaces, and assume that α is atomless. Then there exists a measurable map T such that $T_\sharp \alpha = \beta$.*

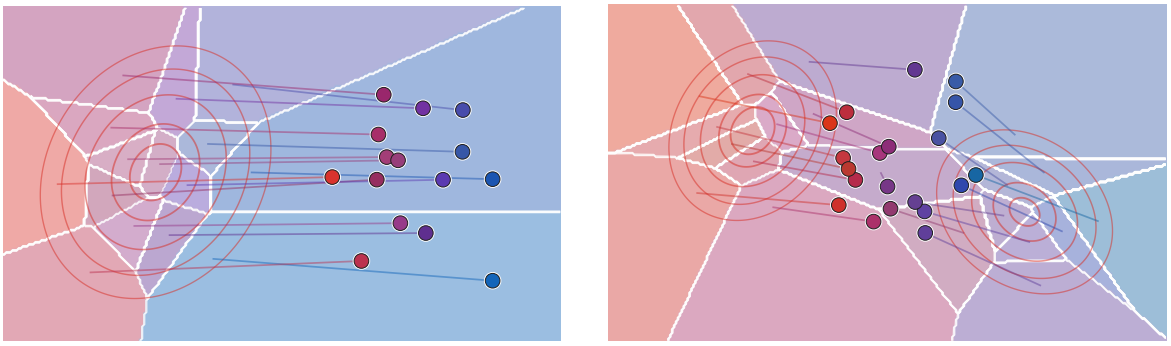
Proof. A standard measure-isomorphism theorem identifies the atomless standard probability space $(\mathcal{X}, \alpha)$ with $([0,1], \mathrm{Leb})$, modulo null sets [69]. It is therefore enough to construct a map from $[0,1]$ to the target law β. Since the target is Polish, it is a standard Borel space; choose a Borel isomorphism i from a full-β Borel subset of $\mathcal{Y}$ onto a Borel subset of $[0,1]$. The generalized quantile of $i_\sharp \beta$ pushes Lebesgue measure to $i_\sharp \beta$. Composing it with i^{-1} and with the source isomorphism gives the desired measurable map, after arbitrary definitions on the discarded null sets. □

Remark 2.17 (Feasibility versus optimality). An optimal map solving (2.5) might fail to exist for two distinct reasons. First, the constraint set can be empty, for instance if $\alpha = \delta_x$ and β is not a single Dirac mass. Proposition 2.16 shows that non-atomicity of the source removes this feasibility obstruction. Second, even when feasible maps exist, the infimum may fail to be attained because the class of maps is not closed under weak limits.

Example 2.18 (A splitting obstruction). Let α be uniform on $\{0\} \times [0,1]$ and let β be the average of the uniform laws on $\{-1\} \times [0,1]$ and $\{1\} \times [0,1]$. For the squared Euclidean cost, the relaxed plan that sends $(0,t)$ equally to $(-1,t)$ and $(1,t)$ has cost 1, which is the smallest possible value because every target point is at horizontal distance 1 from the source segment.
No deterministic map can attain this value. Indeed, equality would force the vertical coordinate to remain equal to t almost everywhere. If $A \subset [0,1]$ were the set of levels sent to the right segment, the target marginal would then require $|A \cap B| = |B|/2$ for every Borel set $B \subset [0,1]$, which is impossible by taking $B = A$. Feasible maps nevertheless approach cost 1: partition $[0,1]$ into n equal cells, send the lower half of each cell affinely onto the corresponding full cell on the left segment, and send the upper half similarly to the right segment. The vertical displacement is at most $1/(2n)$, so the costs converge to 1. Thus the Monge infimum equals the relaxed value but is not attained [469].

Example 2.19 (Semi-discrete Monge maps). The Monge formulation is not symmetric in α and β. It makes sense, for instance, when α has a density with respect to Lebesgue measure and β is discrete. On $\mathcal{X} = \mathcal{Y} = \mathbb{R}^d$, let $\beta = \sum_j b_j \delta_{y_j}$ be supported on $\{y_1, \dots, y_m\}$. A map T such that $T_\sharp \alpha = \beta$ defines a segmentation of the space into cells $C_j := T^{-1}(y_j)$ satisfying $\alpha(C_j) = b_j$. This is the semi-discrete setting. Chapter 6 explains how the cells become Laguerre cells for prescribed masses and ordinary

Voronoi cells when the masses are free. Figure 2.3 shows two such equal-mass piecewise-constant maps: the color of each target atom matches the Laguerre cell that is sent to it. If one exchanges the roles of α and β so that α is discrete, then no valid T exists in general: it is not possible to push forward a discrete measure to a measure with density.



Gaussian source to 15 atoms two-mode source to 20 atoms

Figure 2.3: Semi-discrete Monge maps. The red contours show the continuous source density α. The colored regions extend the numerical Laguerre cells over the whole displayed domain, while their masses are computed with respect to α. The circular atoms form the discrete target β. Their colors are tied to horizontal position, with a small random perturbation, and are reused for the cells. Faint segments connect cell barycenters to their images under the piecewise-constant Monge map.

Figure 2.4 shows a finite-dimensional instance of this deterministic viewpoint. The source and target measures are empirical color clouds in RGB space, and the map transports colors while leaving pixel positions fixed. Grayscale equalization is one-dimensional, but transferring a full color palette requires transporting empirical measures in a three-dimensional color space, for instance RGB or Lab. Early color-transfer methods used affine statistics or iterated one-dimensional projections [450, 436]; replacing these projections by a genuine three-dimensional OT map gives a more intrinsic way to match palettes while preserving their geometry [444].

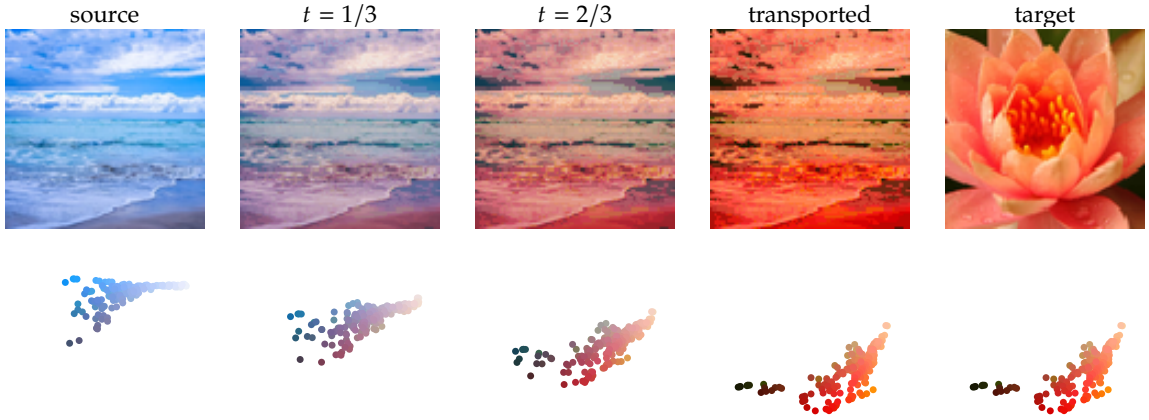


Figure 2.4: Color transfer as a Monge map in RGB space, from a beach photograph to a flower photograph. The top row applies the palette map to the source image; the bottom row shows the empirical color clouds in the RGB cube. Only colors are transported here, not pixel locations.

Monge distance. Specializing the cost to a power of a metric produces a directed transport distance.

Definition 2.20 (Directed Monge distance)**.** Let $(\mathcal{X}, d)$ be a metric space and $1 \leqslant p < +\infty$. For probability measures with finite pth moments, define

$$\tilde{\mathcal{W}}_p(\alpha,\beta)^p := \inf_{T:\mathcal{X}\to\mathcal{X} \text{ measurable}} \left\{ \int_{\mathcal{X}} d(x,T(x))^p \,\mathrm{d}\alpha(x) \;;\; T_\sharp \alpha = \beta \right\}. \tag{2.6}$$

The value is $+\infty$ when the constraint set is empty.

Example 2.21 (Book-shifting in the Monge problem). This is the continuous counterpart of the discrete book-shifting construction in Example 1.3; it exhibits the same non-uniqueness mechanism for the linear cost. Let $\alpha = \frac{1}{2}\mathbb{1}_{[0,2]}(x)\mathrm{d}x$ and $\beta = \frac{1}{2}\mathbb{1}_{[1,3]}(y)\mathrm{d}y$. The monotone translation $T_{\mathrm{tr}}(x) = x + 1$ pushes α to β. For the cost $|x - y|$, it is not the only optimal Monge map. The book-shifting map $T_{\mathrm{book}}(x) = x + 2$ for $0 \leqslant x \leqslant 1$ and $T_{\mathrm{book}}(x) = x$ for $1 < x \leqslant 2$ also satisfies $(T_{\mathrm{book}})_\sharp \alpha = \beta$: it keeps the overlapping interval $[1, 2]$ fixed and sends the left interval $[0, 1]$ to the uncovered interval $[2, 3]$. For any admissible map $T_\sharp \alpha = \beta$,

$$\int |T(x) - x| \mathrm{d}\alpha(x) \geqslant \int (T(x) - x)\mathrm{d}\alpha(x) = \int y \mathrm{d}\beta(y) - \int x \mathrm{d}\alpha(x) = 1.$$

Both T_{tr} and T_{book} satisfy $T(x) \geqslant x$ for α-almost every x, hence both saturate this lower bound and are optimal for the Monge value $\tilde{\mathcal{W}}_1$. More generally, one may replace the map $x \mapsto x + 2$ on $[0, 1]$ by any measure-preserving map from $[0, 1]$ to $[2, 3]$, while keeping the identity on $[1, 2]$. The flatness is specific to the linear cost: for $p > 1$, strict convexity selects the monotone translation.

Proposition 2.22 (Directed Monge distance)**.** *Let $\mathcal{X} = \mathcal{Y}$ be a Polish metric space and $1 \leqslant p < +\infty$. On probability measures with finite p-th moment, the quantity $\tilde{\mathcal{W}}_p$ is nonnegative, vanishes only on the diagonal, and satisfies the triangle inequality. It is therefore a directed extended distance: it need not be symmetric and may take the value $+\infty$.*

Proof. Nonnegativity is immediate. If $\tilde{\mathcal{W}}_p(\alpha, \beta) = 0$, choose feasible maps T_k with $\int d(x, T_k(x))^p \mathrm{d}\alpha(x) \to 0$. For every bounded 1-Lipschitz function h,

$$\left| \int h \mathrm{d}\beta - \int h \mathrm{d}\alpha \right| = \left| \int h(T_k(x)) - h(x) \mathrm{d}\alpha(x) \right| \leqslant \left(\int d(x, T_k(x))^p \mathrm{d}\alpha(x) \right)^{1/p} \longrightarrow 0.$$

Since bounded Lipschitz functions separate probability measures on Polish metric spaces, $\alpha = \beta$. Conversely, the identity map shows that $\tilde{\mathcal{W}}_p(\alpha, \alpha) = 0$.

We prove the triangle inequality. If either $\tilde{\mathcal{W}}_p(\alpha, \gamma)$ or $\tilde{\mathcal{W}}_p(\gamma, \beta)$ is infinite, the claim is automatic. If both are finite, feasible maps $S_\sharp \alpha = \gamma$ and $T_\sharp \gamma = \beta$ exist, so their composition is feasible from α to β; in particular, the left-hand side is finite. Fix $\varepsilon > 0$ and choose ε-minimizers satisfying

$$\mathcal{E}_\alpha(S)^{1/p} \leqslant \tilde{\mathcal{W}}_p(\alpha, \gamma) + \varepsilon, \qquad \mathcal{E}_\gamma(T)^{1/p} \leqslant \tilde{\mathcal{W}}_p(\gamma, \beta) + \varepsilon.$$

The composed map is feasible from α to β, and Minkowski's inequality gives

$$\begin{aligned} \tilde{\mathcal{W}}_p(\alpha, \beta) &\leqslant \left(\int d(x, T(S(x)))^p \mathrm{d}\alpha(x) \right)^{1/p} \\ &\leqslant \left(\int d(x, S(x))^p \mathrm{d}\alpha(x) \right)^{1/p} + \left(\int d(S(x), T(S(x)))^p \mathrm{d}\alpha(x) \right)^{1/p} \\ &= \mathcal{E}_\alpha(S)^{1/p} + \mathcal{E}_\gamma(T)^{1/p} \leqslant \tilde{\mathcal{W}}_p(\alpha, \gamma) + \tilde{\mathcal{W}}_p(\gamma, \beta) + 2\varepsilon. \end{aligned}$$

Letting $\varepsilon \to 0$ gives the result. □

The directed value $\tilde{\mathcal{W}}_p$ is useful conceptually, but it is too rigid to be the main distance between measures: it can be infinite and asymmetric. The Kantorovich formulation remedies both issues by replacing maps with couplings.

2.4 Existence and Uniqueness of the Monge Map

This section records the main regimes where Monge's deterministic formulation becomes well posed. Brenier's theorem is the central result: for the squared Euclidean cost, absolute continuity of the source restores existence, uniqueness and convex-potential structure.

Brenier's theorem. Brenier's theorem [81, 82] ensures that, in $\mathbb{R}^d$ for the quadratic cost, absolute continuity of the source is enough for Monge's problem to have a unique solution. It also gives the decisive structural description of this solution: the optimal map is not an arbitrary map, but the gradient of a convex potential.

Theorem 2.23 (Brenier). *Let $\alpha, \beta \in \mathcal{M}_+^1(\mathbb{R}^d)$ have finite second moments, and assume that α is absolutely continuous with respect to Lebesgue measure. For the quadratic cost $c(x,y) = \|x-y\|^2$, there exists a convex function $\varphi : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ such that*

$$T = \nabla\varphi, \qquad T_\sharp \alpha = \beta,$$

and T is the unique optimal Monge map α-almost everywhere. The optimal Kantorovich plan is $(\mathrm{Id}, T)_\sharp \alpha$.

Proof. The proof uses the Kantorovich relaxation and duality developed later in Chapter 5. Choose optimal c-concave potentials (f, g) for the cost $\|x-y\|^2$ and define

$$\varphi(x) := \frac{1}{2}\|x\|^2 - \frac{1}{2}f(x), \qquad \psi(y) := \frac{1}{2}\|y\|^2 - \frac{1}{2}g(y).$$

The c-concavity of f makes φ convex. Dual feasibility is equivalent to $\varphi(x) + \psi(y) \geqslant \langle x, y\rangle$, and equality holds for almost every pair under any optimal plan. Replacing ψ by the convex conjugate φ^* therefore yields the Fenchel equality $\varphi(x) + \varphi^*(y) = \langle x, y\rangle$ on the optimal relation, hence $y \in \partial\varphi(x)$. Since α has a density and convex functions are differentiable Lebesgue-almost everywhere, $\partial\varphi(x) = \{\nabla\varphi(x)\}$ for α-almost every x. Every optimal plan is therefore concentrated on the same graph, which proves existence and uniqueness of both the optimal plan and the Monge map. Related results for more general strictly convex costs are developed in [231, 94, 539]. □

Brenier's theorem is the higher-dimensional analogue of the one-dimensional monotone rearrangement theorem. In one dimension, the derivative of a convex function is an increasing map; in several dimensions, the corresponding object is the gradient of a convex function. Such gradients are monotone fields:

$$\langle \nabla\varphi(x) - \nabla\varphi(x'),\, x - x'\rangle \geqslant 0.$$

Remark 2.24 (Monotone fields need not be gradients). In dimensions larger than one, not all monotone fields are gradients of convex functions. Consider in $\mathbb{R}^2$ the rotation matrix $R_\theta = \left(\begin{smallmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{smallmatrix}\right)$. The linear map $x \mapsto R_\theta x$ is monotone as soon as $|\theta| \leqslant \pi/2$, because

$$\langle R_\theta x - R_\theta x',\, x - x'\rangle = \langle R_\theta(x-x'),\, x - x'\rangle = \cos(\theta)\|x-x'\|^2 \geqslant 0.$$

However, for $\theta \neq 0$, R_θ is not symmetric and therefore cannot be the gradient of a scalar potential. Indeed, if a linear field Ax equals $\nabla\varphi(x)$, then its Jacobian A must be symmetric; equivalently, a quadratic potential $\varphi(x) = \langle Bx, x\rangle/2$ has gradient $((B + B^\top)/2)x$. Thus monotonicity is weaker than Brenier optimality in dimension $d \geqslant 2$.

Definition 2.25 (Measures not charging hypersurfaces). A Borel measure α on $\mathbb{R}^d$ does not charge hypersurfaces if $\alpha(S) = 0$ for every countably $(d-1)$-rectifiable set S, i.e. every set that can be covered, up to an $\mathcal{H}^{d-1}$-null set, by countably many C^1 hypersurfaces.

Remark 2.26 (A sharp source hypothesis). The absolute-continuity assumption in Brenier's theorem can be weakened: for the quadratic cost, it is enough that α does not charge hypersurfaces [231, 539, 469]. The reason is that the set where a convex potential has a genuinely multi-valued subdifferential is contained in a countable union of lower-dimensional pieces. This condition is close to sharp. If the source gives positive mass to a segment or a hypersurface, the subdifferential may be multi-valued on a set of positive α-mass, and the optimal relaxed plan may need to split mass, as in Example 2.18.

Radial measures. Radial symmetry gives a useful higher-dimensional case where the Brenier map reduces to a one-dimensional monotone rearrangement. A measure α on $\mathbb{R}^d$ is radial if $Q_\sharp \alpha = \alpha$ for every orthogonal map Q. Such a measure is determined by the law of the radius $\|x\|$, and the optimal map transports this radius while keeping the angular direction fixed.

Proposition 2.27 (Optimal transport between radial measures). *Let $\alpha, \beta \in \mathcal{M}_+^1(\mathbb{R}^d)$ be radial probability measures with finite second moments, and assume that α is absolutely continuous with respect to Lebesgue measure. Define their radial distribution functions*

$$F_\alpha(r) := \alpha(\{x : \|x\| \leqslant r\}), \qquad F_\beta(r) := \beta(\{y : \|y\| \leqslant r\}), \qquad r \geqslant 0,$$

and let $F_\beta^{-1}(u) := \inf\{r \geqslant 0 : F_\beta(r) \geqslant u\}$ be the generalized inverse. Set $\tau(r) := F_\beta^{-1}(F_\alpha(r))$. Then the quadratic Brenier map from α to β is the radial map

$$T(0) = 0, \qquad T(x) = \frac{\tau(\|x\|)}{\|x\|} x \quad \text{for } x \neq 0.$$

Moreover, if $\alpha_R = (x \mapsto \|x\|)_\sharp \alpha$ is the law of the source radius, then

$$\mathcal{W}_2^2(\alpha, \beta) = \int_0^\infty |r - \tau(r)|^2 \, \mathrm{d}\alpha_R(r).$$

Proof. Let $\alpha_R = (\|\cdot\|)_\sharp \alpha$ and $\beta_R = (\|\cdot\|)_\sharp \beta$ be the radial laws. Since α is absolutely continuous, α_R has no atoms. The map $\tau = F_\beta^{-1} \circ F_\alpha$ is therefore the monotone one-dimensional transport from α_R to β_R. If $X \sim \alpha$ and $R = \|X\|$, then $\tau(R)$ has law β_R. Moreover, by radiality, the conditional direction $X/\|X\|$ given $R = r$ is uniform on the sphere for α_R-a.e. $r > 0$. Thus T changes only the radius, from R to $\tau(R)$, and keeps the uniform angular distribution. It follows that $T_\sharp \alpha$ is radial with radial law β_R, hence $T_\sharp \alpha = \beta$.

It remains to check optimality. The function τ is nondecreasing and nonnegative. Define $\psi(r) := \int_0^r \tau(s) \mathrm{d}s$ on $\mathbb{R}_+$. Then ψ is convex and nondecreasing, so $\varphi(x) := \psi(\|x\|)$ is convex on $\mathbb{R}^d$. Since α is absolutely continuous, it does not charge the origin nor the radii where the monotone function τ is discontinuous. Thus $\nabla\varphi(x) = \tau(\|x\|)x/\|x\| = T(x)$ for α-a.e. x. The map T is therefore a gradient of a convex potential and pushes α to β. Brenier's theorem identifies it as the unique quadratic optimal Monge map. The displayed cost formula follows by substituting $T(x)$ in $\int \|x - T(x)\|^2 \mathrm{d}\alpha(x)$. □

If α is not absolutely continuous, the same radial idea is still useful but has to be interpreted at the level of couplings of the radial variables: atoms on spheres may have to be split, so a Monge map need not exist.

Polar factorization. Brenier's theorem does more than solve one transport problem: it provides a canonical way to extract the "monotone part" of a nondegenerate map. Suppose one starts from a square-integrable deformation $u : \Omega \to \mathbb{R}^d$, for instance a velocity snapshot, an image deformation or a generative map. The law $\beta = u_\sharp \lambda$ records where the mass ends up, but it forgets how the points of Ω were labelled before being sent there. Brenier's polar factorization [81, 82] separates these two effects. A measure-preserving rearrangement s first changes the labels without changing the source law; the unique convex-gradient map $\nabla\varphi$ then sends the uniform source to the output law. Thus the Brenier factor is the canonical optimal-transport representative among all maps with the same image distribution.

Proposition 2.28 (Polar factorization). *Let $\Omega \subset \mathbb{R}^d$ be a bounded domain endowed with normalized Lebesgue measure λ, and let $u \in L^2(\Omega; \mathbb{R}^d)$. Assume that the image law $\beta = u_\sharp \lambda$ is absolutely continuous with respect to Lebesgue measure. Then there exist a measure-preserving map $s : \Omega \to \Omega$ and a convex function φ such that*

$$u = \nabla\varphi \circ s \qquad \lambda\text{-a.e.}$$

The Brenier factor $\nabla\varphi$ and the rearrangement s are unique λ-almost everywhere.

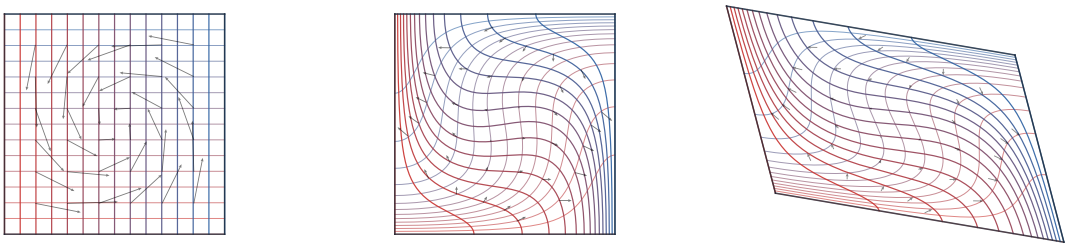
Proof. By Brenier's theorem there is a unique convex-gradient map $T = \nabla\varphi$ transporting λ to β. Since β is also absolutely continuous, the reverse Brenier map S transporting β to λ exists, and the two maps are inverse almost everywhere: $T \circ S = \mathrm{Id}$ β-almost everywhere and $S \circ T = \mathrm{Id}$ λ-almost everywhere. Set $s := S \circ u$. Then

$$s_\sharp \lambda = S_\sharp(u_\sharp \lambda) = S_\sharp \beta = \lambda, \qquad T \circ s = T \circ S \circ u = u \quad \lambda\text{-a.e.}$$

Thus s is measure preserving and gives the announced factorization. Uniqueness of T follows from Brenier's theorem, and then $s = S \circ u$ is unique almost everywhere. □

The absolute-continuity assumption is a transparent sufficient form of Brenier's nondegeneracy hypothesis: $u^{-1}(N)$ must be λ-null whenever $N \subset \mathbb{R}^d$ is Lebesgue-null. Without such a hypothesis, a deterministic factorization can fail or the rearrangement can be nonunique; the always-available relaxation replaces s by a measure-preserving plan.

Figure 2.5 separates these two factors on a colored grid. Its three panels display the initial coordinates x, the relabeled coordinates $s(x)$, and the factorized deformation $u(x) = (\nabla\varphi \circ s)(x)$. The map s swirls the grid inside the square while preserving area, whereas $\nabla\varphi$ is a symmetric positive definite affine map and hence the gradient of a convex quadratic potential.



x $s(x)$ $u(x) = (\nabla\varphi \circ s)(x)$

Figure 2.5: Polar factorization as a relabeling followed by a Brenier map. From left to right, the panels show x, $s(x)$ with $s_\sharp\lambda = \lambda$, and $u(x) = \nabla\varphi(s(x))$. Here s is generated by the area-preserving flow of a divergence-free Hamiltonian vector field, while the Brenier factor is the symmetric positive definite affine map $\nabla\varphi(z) = Bz$. Faint arrows indicate the successive maps.

The name is meant to echo the usual polar decomposition of matrices. This analogy becomes literal for linear maps under the Gaussian reference measure. If $X \sim \mathcal{N}(0, \mathrm{Id})$ and $u(x) = Ax$, then $u_\sharp\mathcal{N}(0, \mathrm{Id}) = \mathcal{N}(0, AA^\top)$. The Brenier map from $\mathcal{N}(0, \mathrm{Id})$ to this Gaussian is $x \mapsto Sx$, where $S = (AA^\top)^{1/2}$ is symmetric positive semidefinite. Hence the decomposition $u = \nabla\varphi \circ s$ reads $A = SO$, where $O = S^{-1}A$ is orthogonal when A is invertible. If A is singular, $S^\dagger A$ is the canonical partial isometry on $(\ker A)^\perp$; because A is square, it can be extended arbitrarily to an orthogonal matrix O on all of $\mathbb{R}^d$ while preserving $A = SO$. This orthogonal extension, rather than the unextended partial isometry, preserves the standard Gaussian law. The factor Sx is the convex-gradient transport part, while Ox is the measure-preserving relabeling.

Displacement interpolation. An optimal map does not only match two endpoint measures; it also tells how to draw a path between them. The construction is Lagrangian and geometric: each particle keeps its identity and travels at constant speed from its initial position to its image under the transport map.

Definition 2.29 (Monge and McCann displacement interpolation)**.** If $T_\sharp\alpha = \beta$, the Monge interpolation generated by T is the curve

$$T_t(x) := (1-t)x + tT(x), \qquad \alpha_t := (T_t)_\sharp\alpha, \qquad t \in [0,1].$$

For the quadratic cost, when T is the optimal Brenier map, this curve is called McCann's displacement interpolation.

McCann's displacement convexity theory [371] clarifies the geometric meaning of interpolating along Brenier maps; this is the language of Wasserstein geodesics used later for barycenters in Section 12.1 and for gradient flows in Chapter 15. If no Monge map exists, or if the optimal object splits mass, the same straight-line idea is applied to each active pair of a coupling, as explained in the Kantorovich interpolation of Section 4.1. The following proposition makes the constant-speed property precise for the directed Monge value.

Proposition 2.30 (Directed Monge displacement geodesics)**.** *Let $1 \leqslant p < +\infty$, let $\alpha, \beta \in \mathcal{M}_+^1(\mathbb{R}^d)$ have finite p-th moments, and let T be an optimal map for $\tilde{\mathcal{W}}_p(\alpha,\beta)$ in (2.6). Set $\alpha_t = (T_t)_\sharp\alpha$ with $T_t = (1-t)\mathrm{Id} + tT$. Assume that, for every $t < 1$, T_t is one-to-one on a full α-measure Borel set. Its inverse on the image of that set is*

then measurable and defined α_t-almost everywhere. For $0 \leqslant s \leqslant t \leqslant 1$,

$$\tilde{\mathcal{W}}_p(\alpha_s, \alpha_t) = (t-s)\tilde{\mathcal{W}}_p(\alpha, \beta).$$

Thus $t \mapsto \alpha_t$ is an oriented constant-speed geodesic for the directed Monge distance. In particular, for $p = 2$, this applies to the Brenier map under the hypotheses of Theorem 2.23.

Proof. The case $s = t$ is trivial, so assume $s < t$. Since $s < 1$, the inverse T_s^{-1} is defined α_s-almost everywhere along the transported particles. Define $S_{s,t} := T_t \circ T_s^{-1}$ α_s-a.e. Then $(S_{s,t})_\sharp \alpha_s = \alpha_t$, and, using the optimality of T,

$$\tilde{\mathcal{W}}_p(\alpha_s, \alpha_t)^p \leqslant \int \|S_{s,t}(z) - z\|^p \mathrm{d}\alpha_s(z) = \int \|T_t(x) - T_s(x)\|^p \mathrm{d}\alpha(x) = (t-s)^p \tilde{\mathcal{W}}_p(\alpha, \beta)^p.$$

The same particle construction gives $\tilde{\mathcal{W}}_p(\alpha, \alpha_s) \leqslant s\tilde{\mathcal{W}}_p(\alpha, \beta)$. If $t < 1$, the map $T \circ T_t^{-1}$ sends α_t to β and gives $\tilde{\mathcal{W}}_p(\alpha_t, \beta) \leqslant (1-t)\tilde{\mathcal{W}}_p(\alpha, \beta)$; if $t = 1$, this latter distance is zero. Using the triangle inequality from Proposition 2.22,

$$\tilde{\mathcal{W}}_p(\alpha, \beta) \leqslant \tilde{\mathcal{W}}_p(\alpha, \alpha_s) + \tilde{\mathcal{W}}_p(\alpha_s, \alpha_t) + \tilde{\mathcal{W}}_p(\alpha_t, \beta) \leqslant s\tilde{\mathcal{W}}_p(\alpha, \beta) + \tilde{\mathcal{W}}_p(\alpha_s, \alpha_t) + (1-t)\tilde{\mathcal{W}}_p(\alpha, \beta).$$

Hence $\tilde{\mathcal{W}}_p(\alpha_s, \alpha_t) \geqslant (t-s)\tilde{\mathcal{W}}_p(\alpha, \beta)$, which proves equality. For a Brenier map $T = \nabla\varphi$, the map T_t is the gradient of $x \mapsto (1-t)\frac{\|x\|^2}{2} + t\varphi(x)$, which is $(1-t)$-strongly convex for every $t < 1$. On the full-measure set where φ is differentiable, this gives $\langle T_t(x) - T_t(y), x - y\rangle \geqslant (1-t)\|x-y\|^2$, so T_t is injective there. This gives the last claim. □

Figure 2.6 illustrates this displacement geodesic on two non-convex silhouettes, both through representative particle paths and through the evolving transported density.

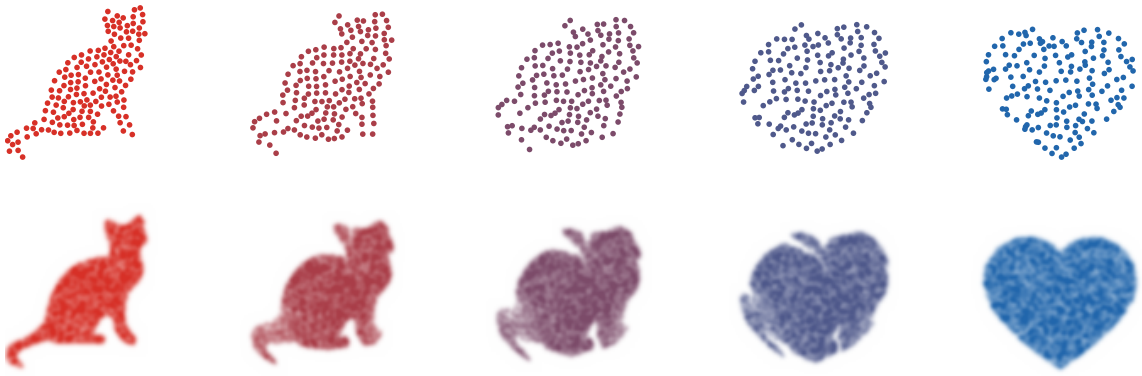


Figure 2.6: McCann displacement interpolation between a cat silhouette and a heart silhouette. The first row displays a small farthest-point subset of transported particles along $T_t(x) = (1-t)x + tT(x)$. The second row renders kernel-smoothed densities from a denser transported cloud as color images: white means zero density, while high density saturates in the red-to-blue interpolation color of the corresponding time.

Caffarelli's regularity theory, discussed next, explains when the convex potential is actually smooth enough to define a classical deformation.

Regularity and the Monge–Ampère equation. The previous results identify the optimal map; regularity theory asks when this map is a classical smooth deformation rather than only an almost-everywhere gradient. For quadratic costs this question becomes the regularity theory of the Monge–Ampère equation.

Proposition 2.31 (Caffarelli regularity)**.** *Let $\Omega, \Lambda \subset \mathbb{R}^d$ be bounded uniformly convex domains with C^2 boundaries. Let $\alpha = \rho_\alpha(x)\mathrm{d}x$ and $\beta = \rho_\beta(y)\mathrm{d}y$ be supported on Ω and Λ, respectively, with*

$$0 < m \leqslant \rho_\alpha, \rho_\beta \leqslant M < +\infty.$$

If $\rho_\alpha \in C^\alpha(\overline{\Omega})$ and $\rho_\beta \in C^\alpha(\overline{\Lambda})$ for some $\alpha \in (0,1)$, then the Brenier potential φ is strictly convex in Ω and belongs to $C^{2,\alpha}_{\mathrm{loc}}(\Omega)$. Consequently, $\nabla\varphi \in C^{1,\alpha}_{\mathrm{loc}}(\Omega)$. Under the standard compatibility and higher boundary-regularity assumptions, the same regularity extends to the boundary.

Proof. The potential solves the Monge–Ampère equation

$$\det(\nabla^2\varphi(x)) = \frac{\rho_\alpha(x)}{\rho_\beta(\nabla\varphi(x))}$$

in the Alexandrov sense, with second boundary condition $\nabla\varphi(\Omega) = \Lambda$. The density bounds and convexity of the domains give strict convexity and localization of sections. Caffarelli's interior theory then yields $C^{2,\alpha}_{\mathrm{loc}}$ estimates for φ; the boundary statement follows from the boundary regularity theory under uniform convexity and compatibility assumptions [93, 539]. □

Figure 2.7 illustrates the geometric role of the convexity assumption through a finite-sample stress test rather than a counterexample. A quadratic assignment transports a dense farthest-point sample of the disk to a dense farthest-point sample of a connected non-convex target made of two disks joined by a thin rectangle. The panels show the corresponding empirical McCann interpolation, so the loss of convex target geometry is visible directly along the transported particles.

$t = 0$ $t = 1/4$ $t = 1/2$ $t = 3/4$ $t = 1$

Figure 2.7: Empirical quadratic OT interpolation from a disk to a connected non-convex two-disk domain. A 5200-point farthest-point sample of the disk is matched to a 5200-point farthest-point sample of two disks connected by a thin rectangle. The panels display $(1-t)x_i + tT_N(x_i)$ for the empirical optimal assignment T_N. Colors are inherited from the horizontal coordinate of x_i in the initial disk, making the transported material regions visible throughout the interpolation.

Remark 2.32 (Regularity, weak maps, and splitting). Caffarelli's theorem should be read as a warning as well as a theorem. Brenier's theorem gives existence and uniqueness under mild assumptions, but smoothness requires density bounds, smoothness and convex geometry that are rarely satisfied by empirical, manifold-supported or neural generative distributions. In such applications, the exact OT map is often only weakly defined, possibly unstable, and better represented by a coupling, an entropic approximation or a learned parametric surrogate.
Even without smoothness, the convex potential is locally Lipschitz on the interior of its domain, so $\nabla\varphi$ is defined Lebesgue-almost everywhere. If the source measure does not satisfy the non-splitting hypotheses of Brenier's theorem, the correct relaxed object is instead an optimal Kantorovich plan concentrated on the graph of the set-valued map $\partial\varphi$. At points where $\partial\varphi(x)$ contains several target locations, the plan may split the mass starting from x. Thus the subdifferential still describes the geometry of optimality, but the transport object is a coupling rather than a single-valued map.

For sufficiently smooth positive densities and a classical Brenier map, the change-of-variables formula (2.4) gives the Monge–Ampère equation

$$\det(\nabla^2\varphi(x))\rho_\beta(\nabla\varphi(x)) = \rho_\alpha(x). \tag{2.7}$$

With suitable boundary conditions, this characterizes the Brenier potential up to an additive constant among convex solutions. The identity holds pointwise in the classical regime; for nonsmooth convex potentials it is interpreted through the Alexandrov Monge–Ampère measure. The convexity constraint selects the degenerate-elliptic branch and forces $\det(\nabla^2\varphi(x)) \geqslant 0$ wherever the Hessian exists.

Remark 2.33 (Numerical Monge–Ampère solvers). The Monge–Ampère operator should be viewed as a fully nonlinear, degenerate elliptic analogue of the Laplacian. Proposition 2.34 makes this analogy literal at first order, where the equation becomes a weighted Poisson equation. The nonlinear operator $\varphi \mapsto \det(\nabla^2\varphi)$, however, is elliptic only on the convex branch, together with the second boundary condition $\nabla\varphi(\Omega) = \Lambda$. This is the numerical difficulty: a scheme is not reliable merely because it approximates the determinant of the Hessian. It must also select the convex Alexandrov/viscosity solution and discretize the boundary condition consistently.
Several complementary approaches have been developed. Geometric discretizations go back to Oliker–Prussner and reflector-design methods; PDE-based solvers include Newton methods, monotone

wide-stencil finite differences and variational discretizations; semi-discrete and power-diagram methods exploit the convex-cell structure of the transport map. Representative references include [407, 95, 351, 42, 222, 41, 385, 511]. We do not describe these schemes here; the point is that numerical Monge–Ampère transport is primarily a problem of consistent discretization of this nonlinear Laplacian-like operator.

The following proposition records the infinitesimal form of this nonlinear equation. It is useful both conceptually and numerically: close to a smooth reference density, Monge–Ampère transport reduces at first order to a weighted Poisson equation.

Proposition 2.34 (Linearization of the Monge–Ampère equation)**.** *Let $\rho_\varepsilon = \rho_0 + \varepsilon r + o(\varepsilon)$ be a smooth perturbation of a positive reference density ρ_0 on a smooth bounded domain Ω, with $\int_\Omega r\,\mathrm{d}x = 0$. Assume that the expansions hold in a topology strong enough to differentiate the change-of-variables identity. If $T_\varepsilon(x) = x + \varepsilon\nabla u(x) + o(\varepsilon)$ transports $\rho_0\mathrm{d}x$ to $\rho_\varepsilon\mathrm{d}x$, then, to first order,*

$$-\nabla\cdot(\rho_0\nabla u) = r.$$

In particular, when ρ_0 is constant, the linearized equation is $-\Delta u = r/\rho_0$. If the source and target occupy the same fixed domain and no mass crosses its boundary, the accompanying first-order boundary condition is $\rho_0\partial_n u = 0$ on $\partial\Omega$; together with $\int_\Omega u\,\mathrm{d}x = 0$, it fixes the additive constant.

Proof. The change-of-variables equation for T_ε is

$$\rho_0(x) = \rho_\varepsilon(T_\varepsilon(x))\det(\nabla T_\varepsilon(x)).$$

Expanding $\rho_\varepsilon(x+\varepsilon\nabla u) = \rho_0(x) + \varepsilon r(x) + \varepsilon\langle\nabla\rho_0, \nabla u\rangle + o(\varepsilon)$ and $\det(\mathrm{Id} + \varepsilon\nabla^2 u) = 1 + \varepsilon\Delta u + o(\varepsilon)$ gives

$$\rho_0 = \rho_0 + \varepsilon\big(r + \langle\nabla\rho_0, \nabla u\rangle + \rho_0\Delta u\big) + o(\varepsilon) = \rho_0 + \varepsilon\big(r + \nabla\cdot(\rho_0\nabla u)\big) + o(\varepsilon).$$

The first-order term must vanish. □

2.5 Beyond the Quadratic Euclidean Cost

The quadratic Euclidean cost is the model case, but optimal-map theory also covers many non-quadratic costs. The key point is to separate three roles: a convexity-like structure gives potentials, a twist condition prevents splitting, and curvature-type conditions give regularity.

$\mathcal{W}_p$ costs. The quadratic cost is special because it identifies the optimal map with the Euclidean gradient of an ordinary convex potential. For the Wasserstein cost, normalized as $c(x,y) = \|x-y\|^p/p$ with $p > 1$, the same Monge-map picture survives, but the potential is adapted to the cost. More generally, for $c(x,y) = h(x-y)$ with h smooth and strictly convex, absolute continuity of the source again rules out splitting and, under the usual finite-cost assumptions, yields a unique optimal map. This map is characterized by a c-convex potential f, equivalently a source-side $\bar{c}$-concave potential in the convention of Definition 5.14; Section 5.3 develops these potential classes. At differentiability points of f,

$$\nabla f(x) = \nabla_x c(x, T(x)).$$

For $c(x,y) = \|x-y\|^p/p$, this gives the explicit relation

$$T(x) = x - \|\nabla f(x)\|^{q-2}\nabla f(x), \qquad \frac{1}{p} + \frac{1}{q} = 1.$$

Thus the Brenier formula $T = \nabla\varphi$ should be read as the quadratic representative of a broader c-convex theory. The Euclidean theory for strictly convex displacement costs is developed in Gangbo–McCann's work on the geometry of optimal transportation [231]; see also the general treatment in [539].

Squared geodesic distance. On a Riemannian manifold, the natural analogue of the quadratic Euclidean cost is the squared geodesic distance $c(x,y) = d_M(x,y)^2/2$. The optimal map is no longer written with a vector-space subtraction, but with the exponential map:

$$T(x) = \exp_x(-\nabla\varphi(x)),$$

where φ is c-convex. This is the intrinsic version of the formula $T = x - \nabla\varphi$ in normal coordinates. Ignoring cut-locus and measurable-selection issues, the associated McCann interpolation is

$$T_t(x) = \exp_x\Big(t \exp_x^{-1}(T(x))\Big) = \exp_x(-t\nabla\varphi(x)), \qquad \alpha_t = (T_t)_\sharp \alpha, \qquad t \in [0,1].$$

More generally, if π is an optimal coupling and $G_t(x,y) = \exp_x(t \exp_x^{-1}(y))$ selects the point at time t on a minimizing geodesic from x to y, then $\alpha_t = (G_t)_\sharp \pi$. The main additional issues are precisely the cut locus, possible non-uniqueness of minimizing geodesics, and the regularity of the exponential map, which is why the Euclidean statement is usually presented first. The Riemannian polar-factorization theorem of McCann [372] gives the corresponding optimal-map framework, and these ideas feed into displacement convexity and the general manifold theory of optimal transport [371, 539].

Figure 2.8 illustrates this intrinsic interpolation on a positively curved hemisphere.

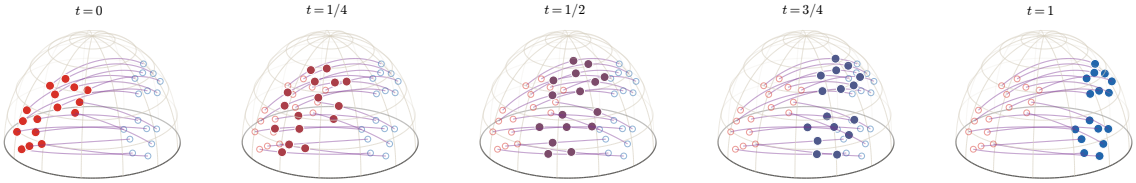


Figure 2.8: Intrinsic McCann interpolation on the upper hemisphere. Here $\mathcal{X} = \{x \in \mathbb{R}^3 : \|x\| = 1,\ x_3 \geqslant 0\}$ and $d_{\mathcal{X}}(x,y) = \arccos(\langle x, y\rangle)$. The thin violet curves are minimizing geodesics, and the colored points show the interpolation at $t = 0, 1/4, 1/2, 3/4, 1$.

Poincaré disk. Negative curvature gives the complementary intrinsic interpolation shown in Figure 2.9.

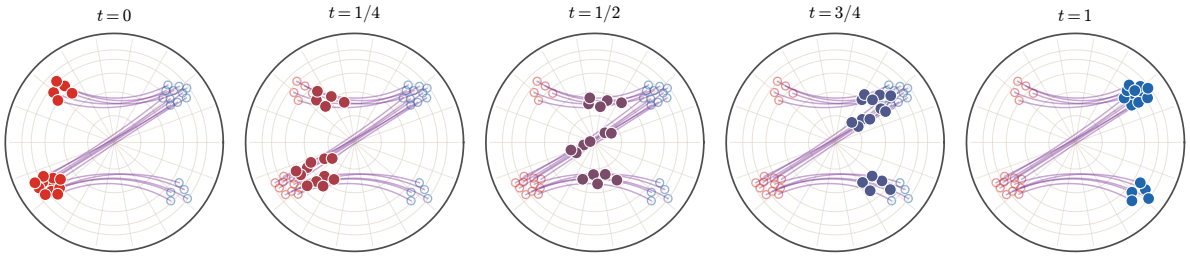


Figure 2.9: Intrinsic McCann interpolation in the Poincaré disk. Here $\mathcal{X} = \{x \in \mathbb{R}^2 : \|x\| < 1\}$ and $d_{\mathcal{X}}(x,y) = \operatorname{arcosh}\Big(1 + 2\|x-y\|^2/((1-\|x\|^2)(1-\|y\|^2))\Big)$. The black circle is the ideal boundary, the faint grid is intrinsic, and the thin violet curves are minimizing geodesics; the colored points show the interpolation at $t = 0, 1/4, 1/2, 3/4, 1$.

Twist condition. The first non-degeneracy condition asks that the first-order information at a source point identifies at most one target point. This is the structural hypothesis that turns an optimal relation into a map.

Definition 2.35 (Twist condition)**.** Let $X, Y \subset \mathbb{R}^d$ and let $c \in C^1(X \times Y)$. The cost c satisfies the twist condition from X to Y if, for every fixed $x \in X$, the map

$$y \in Y \longmapsto \nabla_x c(x,y) \in \mathbb{R}^d$$

is injective. The dual twist condition is the analogous injectivity of $x \mapsto \nabla_y c(x,y)$ for every fixed y.

Proposition 2.36 (Twist prevents splitting)**.** *Assume that $c \in C^1(X \times Y)$ satisfies the twist condition from X to Y. Suppose that optimal pairs for the cost c admit a source potential f with the following first-order certificate: f is differentiable on a full α-measure subset of X, and whenever x belongs to this differentiability set and y can be paired optimally with x, one has $\nabla f(x) = \nabla_x c(x,y)$. Then, for α-almost every source point x, there is at most one*

target point y that can be paired optimally with x. Consequently, any optimal transport relation satisfying these hypotheses is single-valued α-almost everywhere, hence is induced by a Monge map.

Proof. The proof uses the Kantorovich duality formalism developed later in Chapter 5. Let (f, g) be optimal Kantorovich potentials and let π be an optimal plan. Complementary slackness gives $f(x) + g(y) = c(x, y)$ for π-almost every (x, y), while dual feasibility gives $f(z) + g(y) \leqslant c(z, y)$ for all z. Thus, for almost every contact pair (x, y), the function $z \mapsto c(z, y) - g(y)$ touches f from above at x. At a point where f is differentiable, this gives $\nabla f(x) = \nabla_x c(x, y)$. If the cost is twisted, this equation determines at most one y. Hence no optimal plan can split the mass of such an x between several target points. Since differentiability holds on a full α-measure set, the plan is concentrated on the graph of a measurable map there. □

The quadratic cost satisfies twist since $\nabla_x \|x - y\|^2 = 2(x - y)$; the bilinear cost $c(x, y) = -\langle x, y\rangle$ satisfies twist since $\nabla_x c(x, y) = -y$; more generally $c(x, y) = h(x - y)$ is twisted when h is smooth strictly convex and ∇h is injective. On a Riemannian manifold, the squared geodesic cost is twisted locally away from the cut locus. By contrast, a separated cost $a(x) + b(y)$ is never twisted, since $\nabla_x c$ does not see y.

Ma–Trudinger–Wang curvature. Twist gives a map, but it does not by itself make this map continuous or smooth. For a general smooth cost, the relevant structural hypothesis is the Ma–Trudinger–Wang (MTW) condition, introduced for a priori estimates of the generated Jacobian equation associated with optimal transport [360, 523].

Definition 2.37 (Weak MTW condition)**.** Assume that $c \in C^4(X \times Y)$, that $X, Y \subset \mathbb{R}^d$, that c is twisted, and that the mixed Hessian $\nabla^2_{xy} c(x, y)$ is invertible. For (x, p) in the image of the change of variables $(x, y) \mapsto (x, -\nabla_x c(x, y))$, let $Y(x, p)$ be defined by $-\nabla_x c(x, Y(x, p)) = p$, and set $A_{i,j}(x, p) := -\partial^2_{x_i x_j} c(x, Y(x, p))$. The cost satisfies the weak MTW condition if

$$\sum_{i,j,k,l} \partial^2_{p_k p_l} A_{i,j}(x, p)\, \xi_i \xi_j \eta_k \eta_l \geqslant 0 \qquad \text{whenever } \langle \xi, \eta\rangle = 0.$$

A strong MTW condition asks for a positive lower bound, proportional to $\|\xi\|^2 \|\eta\|^2$, on the same orthogonal directions.

The tensor above is often called the cost-sectional curvature. With the sign convention used here, it measures how the negative x-Hessian of the cost bends when the target point is varied through the dual momentum p.

Proposition 2.38 (MTW condition and regularity)**.** *Assume that c is smooth, twisted and non-degenerate, that the source and target domains satisfy the appropriate mutual c-convexity assumptions, and that the source and target densities are positive and bounded above and below. Under the weak MTW condition, the optimal map is locally Hölder continuous; with the corresponding global domain hypotheses, it is continuous up to the boundary. Under the strong MTW condition and higher smoothness of the data and domains, one obtains higher regularity estimates.*

Proof. This is the regularity theory of Ma–Trudinger–Wang, Trudinger–Wang and Loeper, stated in structural form rather than with its full boundary hypotheses. The c-convex potential solves a generated Jacobian equation. Twist and non-degeneracy turn its optimal relation into a single-valued map, while the weak MTW inequality controls the geometry of contact sets and yields interior Hölder estimates. Strong MTW provides a quantitative curvature lower bound; combined with smooth densities, domain geometry and elliptic estimates, it yields higher regularity. Conversely, Loeper proved that weak MTW is necessary for continuity for arbitrary smooth positive data [360, 523, 349, 539]. □

The flat quadratic and bilinear costs have zero MTW curvature, hence satisfy the weak condition. For the squared Riemannian distance, the MTW tensor is a refined curvature condition on the cost: near the diagonal it recovers sectional curvature, negative sectional curvature gives an obstruction, and global regularity also depends on cut-locus and domain-convexity issues. There are nevertheless important curved examples. Round spheres have uniformly positive cost-sectional curvature away from the cut locus and satisfy the strong MTW condition [350]; finite products of round spheres are nonnegatively cross-curved and therefore satisfy weak MTW [305, 216]. Nonnegative cross-curvature,

and hence weak MTW, is also inherited by Riemannian submersions, yielding further examples such as complex projective spaces [305]. Many smooth strictly convex costs still fail MTW, which is why twist is enough for existence of a map but not for regularity.

2.6 One-Dimensional Transport and Quantiles

One-dimensional OT is the model case where order replaces geometry and all constructions become explicit.

Cumulative and quantile functions. In one dimension, optimal transport is completely explicit. The cumulative distribution function orders the mass, and the optimal coupling is obtained by matching equal quantile levels. This special case is both a computational tool and the template for several linearized constructions used later.

Definition 2.39 (Cumulative and quantile functions)**.** For $\alpha \in \mathcal{M}_+^1(\mathbb{R})$, its cumulative distribution function is

$$C_\alpha(x) := \alpha((-\infty, x]). \tag{2.8}$$

Its generalized inverse, or quantile function, is

$$C_\alpha^{-1}(r) := \inf\{x \in \mathbb{R} \,;\, C_\alpha(x) \geqslant r\}, \qquad r \in (0,1). \tag{2.9}$$

Figure 2.10 follows three smooth laws through these representations. Their one, two, and three modes have moderately varying weights and widths. Each density mode produces a change of slope in the cumulative function, while intervals of low density become steep portions of the quantile function. The aligned quarter-mass guides make the inversion between the last two panels explicit.

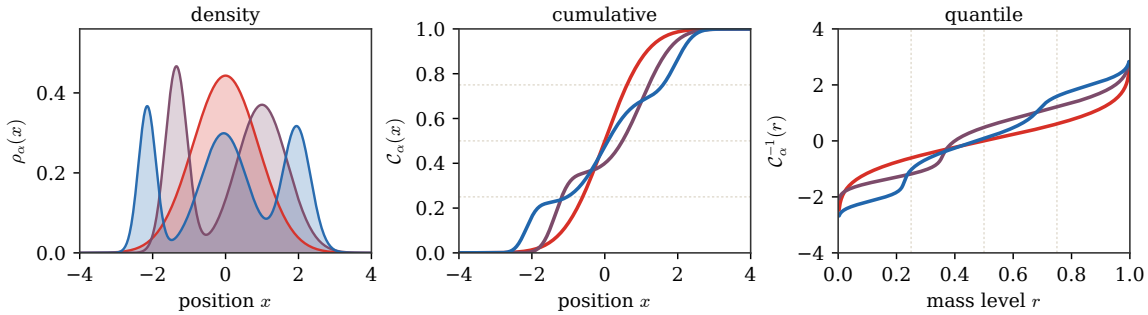


Figure 2.10: Densities, cumulative functions and quantiles for three Gaussian mixtures. Transparent red, violet and blue density areas represent mixtures with one, two and three modes, respectively, with unequal component weights and widths. The cumulative functions integrate the modes into successive increases of mass, while inversion exchanges slowly varying cumulative regions with steep quantile regions. Faint guides mark the same quarter-mass levels in the last two panels.

Proposition 2.40 (Quantile push-forward)**.** *One has* $(C_\alpha^{-1})_\sharp \mathrm{Leb}_{[0,1]} = \alpha$*. If* α *has no atoms, then* $(C_\alpha)_\sharp \alpha = \mathrm{Leb}_{[0,1]}$.

Proof. For simplicity, assume first that α has a strictly positive density, so that C_α is strictly increasing and continuous. Denote $\gamma := (C_\alpha^{-1})_\sharp \mathrm{Leb}_{[0,1]}$. It is enough to prove that $C_\gamma = C_\alpha$. For every x,

$$C_\gamma(x) = \int_0^1 \mathbf{1}_{(-\infty,x]}(C_\alpha^{-1}(z))\mathrm{d}z = \int_0^1 \mathbf{1}_{[0,C_\alpha(x)]}(z)\mathrm{d}z = C_\alpha(x),$$

where we used $C_\alpha^{-1}(z) \leqslant x$ if and only if $z \leqslant C_\alpha(x)$. General measures follow from the same argument with generalized inverses and right-continuity of the cumulative distribution function. If α has no atoms, the probability integral transform gives $(C_\alpha)_\sharp \alpha = \mathrm{Leb}_{[0,1]}$. □

1D Monge solutions. The quantile construction becomes a deterministic Monge map only when the source has no atoms, because the cumulative distribution function can then be used as a genuine change of variable. This gives an explicit one-dimensional Monge solution in the atomless case. If α has atoms,

a deterministic map may fail to realize the same quantile matching because an atom cannot be split; the corresponding relaxed statement is treated in Section 3.3, paragraph Kantorovich solution in 1D.

Theorem 2.41 (One-dimensional Monge solution). *Let $\alpha, \beta \in \mathcal{M}_+^1(\mathbb{R})$, assume that α has no atoms, and let $h : \mathbb{R} \to \mathbb{R}$ be convex. Consider the cost $c(x, y) = h(x - y)$. Write $q_\alpha := C_\alpha^{-1}$ and $q_\beta := C_\beta^{-1}$, and assume that $h(q_\alpha - q_\beta) \in L^1(0, 1)$. Then*

$$T = q_\beta \circ C_\alpha = C_\beta^{-1} \circ C_\alpha \tag{2.10}$$

satisfies $T_\sharp \alpha = \beta$ and minimizes

$$\min_{S_\sharp \alpha = \beta} \int h(x - S(x)) \mathrm{d}\alpha(x).$$

In particular, $h(t) = |t|^p$ gives the usual one-dimensional Monge map for every $1 \leqslant p < +\infty$ whenever the source has no atoms.

Proof. By Proposition 2.40, atomlessness of α gives $(C_\alpha)_\sharp \alpha = \mathrm{Leb}_{[0,1]}$, while $(q_\beta)_\sharp \mathrm{Leb}_{[0,1]} = \beta$. Hence $T_\sharp \alpha = \beta$.

Let S be any admissible map and set $\pi_S = (\mathrm{Id}, S)_\sharp \alpha$. The one-dimensional uncrossing theorem for relaxed couplings, proved later as Theorem 3.39, shows that the quantile coupling $\pi^\star = (q_\alpha, q_\beta)_\sharp \mathrm{Leb}_{[0,1]}$ is optimal among all couplings. Since α has no atoms, $(\mathrm{Id}, T)_\sharp \alpha = \pi^\star$, and therefore

$$\int h(x - T(x)) \mathrm{d}\alpha(x) = \int h(x - y) \mathrm{d}\pi^\star(x, y) \leqslant \int h(x - y) \mathrm{d}\pi_S(x, y) = \int h(x - S(x)) \mathrm{d}\alpha(x).$$

This proves optimality of T among Monge maps. □

Remark 2.42 (Composition is one-dimensional). In dimension one, optimal maps compose. Assume for simplicity that the intermediate laws have no atoms, so that the monotone rearrangements

$$T_{\alpha \to \beta} = C_\beta^{-1} \circ C_\alpha, \qquad T_{\beta \to \gamma} = C_\gamma^{-1} \circ C_\beta$$

are well defined α- and β-almost everywhere. Then

$$T_{\beta \to \gamma} \circ T_{\alpha \to \beta} = T_{\alpha \to \gamma} \qquad \alpha\text{-a.e.}$$

Indeed, each map is nondecreasing and sends quantile level r to the same quantile level of the target law. This semigroup property is special to the ordered line.

The obstruction in higher dimension is already visible for the most elementary Gaussian maps.

Example 2.43 (Linear obstruction to composing Brenier maps). In higher dimension, Brenier maps for the quadratic cost are gradients of convex functions, and such maps do not generally remain gradients after composition. The simplest obstruction is linear. If $T_1(x) = A_1 x$ and $T_2(x) = A_2 x$ with A_1, A_2 symmetric positive definite, then $T_2 \circ T_1$ has matrix $A_2 A_1$. It is a gradient field only when this product is symmetric, equivalently $A_1 A_2 = A_2 A_1$. Gaussian transport gives a concrete instance: between nondegenerate Gaussian laws, the Brenier map is affine with symmetric positive definite linear part. Compositions of Gaussian optimal maps are therefore optimal only in special commuting situations, for instance when all covariance matrices are simultaneously diagonalizable. Otherwise the composition contains a rotational or shearing component and is not the Brenier map between the initial and final Gaussians.

For discrete measures, one cannot directly apply the map formula when the source has atoms, but if the measures are uniform on the same number of Dirac masses, then it is exactly the sorting formula of Proposition 1.2.

Proposition 2.44 (One-dimensional Wasserstein formulas). *Let $1 \leqslant p < +\infty$ and let $\alpha, \beta \in \mathcal{M}_+^1(\mathbb{R})$ have finite p-th moments. Then*

$$\mathcal{W}_p(\alpha, \beta)^p = \int_0^1 \left| C_\alpha^{-1}(r) - C_\beta^{-1}(r) \right|^p \mathrm{d}r = \| C_\alpha^{-1} - C_\beta^{-1} \|_{L^p([0,1])}^p. \tag{2.11}$$

For $p = 1$, this is equivalently

$$\mathcal{W}_1(\alpha, \beta) = \| C_\alpha - C_\beta \|_{L^1(\mathbb{R})} = \int_{\mathbb{R}} \left| C_\alpha(x) - C_\beta(x) \right| \mathrm{d}x = \int_{\mathbb{R}} \left| \int_{-\infty}^x \mathrm{d}(\alpha - \beta) \right| \mathrm{d}x. \tag{2.12}$$

Proof. The first formula follows from Theorem 3.39: the optimal coupling is obtained by taking the same quantile level r for both measures. For $p = 1$, use the layer-cake identity. If q_α and q_β are the two quantile functions, then

$$\int_0^1 |q_\alpha(r) - q_\beta(r)| \mathrm{d}r = \int_{\mathbb{R}} \lambda\big(\{r : q_\alpha(r) \leqslant x < q_\beta(r)\} \cup \{r : q_\beta(r) \leqslant x < q_\alpha(r)\}\big) \mathrm{d}x.$$

The measure of the displayed set is exactly $|C_\alpha(x) - C_\beta(x)|$ for almost every x. □

For $p = 2$, Bobkov and Ledoux [68] gave a complementary representation involving only cumulative distribution functions. It gives the same exact value as the quantile formula, but replaces the inverse-CDF step by an integral over CDF values. This is useful when cumulative functions are easier to estimate or aggregate than quantiles. This cumulative formula has recently been used to design data-parallel estimators of sliced Wasserstein distances [532].

Proposition 2.45 (Bobkov–Ledoux cumulative formula)**.** *Let $\alpha, \beta \in \mathcal{M}_+^1([a,b])$ and write $s_+ := \max(s, 0)$. Then*

$$\mathcal{W}_2^2(\alpha,\beta) = 2\int_a^b \int_x^b \Big[\big(C_\alpha(x) - C_\beta(y)\big)_+ + \big(C_\beta(x) - C_\alpha(y)\big)_+\Big] \mathrm{d}y\, \mathrm{d}x. \tag{2.13}$$

Proof. For any $u, v \in [a,b]$, one has

$$|u - v|^2 = 2\int_a^b \int_x^b \Big[\mathbf{1}_{\{u \leqslant x \leqslant y < v\}} + \mathbf{1}_{\{v \leqslant x \leqslant y < u\}}\Big] \mathrm{d}y\, \mathrm{d}x,$$

because, when $u < v$, the integration domain is the triangle $u \leqslant x \leqslant y < v$, whose area is $(v-u)^2/2$, and the case $v < u$ is symmetric. Apply this identity with $u = q_\alpha(r)$ and $v = q_\beta(r)$, where $q_\alpha = C_\alpha^{-1}$ and $q_\beta = C_\beta^{-1}$, and integrate with respect to $r \in (0,1)$. Fubini's theorem applies since the integrand is bounded. For $x \leqslant y$, the generalized inverse identities give, up to endpoint sets of zero Lebesgue measure,

$$\lambda\left\{r\ ;\ q_\alpha(r) \leqslant x \text{ and } q_\beta(r) > y\right\} = \big(C_\alpha(x) - C_\beta(y)\big)_+,$$

and the same argument with α and β exchanged gives the second positive part. □

> **Remark 2.46 (Quantiles linearize one-dimensional Wasserstein geometry).** Formula (2.11) means that the map $\alpha \mapsto C_\alpha^{-1}$ embeds one-dimensional Wasserstein geometry isometrically into $L^p(0,1)$. Its image is the closed convex cone of almost-everywhere nondecreasing functions with finite p-th moment, not the whole linear space. For $p = 2$, the Wasserstein metric on probability measures over the real line is therefore Hilbertian in the precise sense of being isometric to a convex subset of a Hilbert space. This linearization explains several later simplifications: one-dimensional barycenters are averages of quantiles in Proposition 12.7; sliced Wasserstein applies the same formula after every projection direction in Definition 11.22 and (11.34); the Hilbert embedding of SW_2 in Remark 11.27 collects projected quantile functions; and linear OT coordinates use a reference map to produce the Euclidean embedding (11.51). The statistical formulas in Chapter 10 should also be read in this light: empirical OT on the line is the study of empirical quantile processes, while (2.13) gives an equivalent cumulative representation when inverse CDFs are inconvenient.

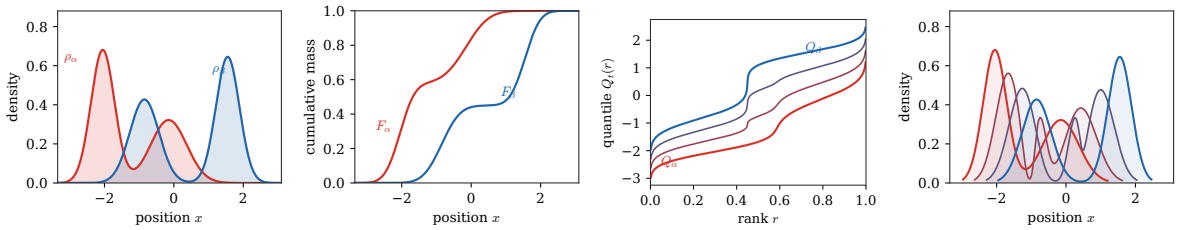
Figure 2.11 gathers the density, cumulative and quantile representations of the same one-dimensional transport problem, together with its displacement interpolation.

The last panel is the one-dimensional specialization of the displacement interpolation of Section 2.4. In quantile coordinates, the interpolating measure is characterized by

$$C_{\alpha_t}^{-1}(r) = (1-t)C_\alpha^{-1}(r) + tC_\beta^{-1}(r), \qquad r \in (0,1).$$

The histogram-equalization figure in Section 1.1 is the same construction applied to pixel intensities.

> **Remark 2.47 ($\mathcal{W}_1$ is a norm).** For $p = 1$, formula (2.12) shows that $\mathcal{W}_1$ is a norm on signed measures with zero total mass once they are identified with their cumulative primitives. This norm property is not specific to the line: in any dimension, and more generally on metric spaces after restricting to zero-mass signed measures for which the Kantorovich–Rubinstein supremum is finite, duality defines the norm studied in Section 6.5. The one-dimensional formula is special only because the norm has



densities cumulative functions quantiles displacement geodesic

Figure 2.11: One-dimensional transport through quantiles. The same two smooth laws are shown as densities, cumulative functions and quantile functions. The last panel displays the displacement interpolation obtained by the linear quantile path $Q_t = (1-t)Q_\alpha + tQ_\beta$, which is the explicit one-dimensional $\mathcal{W}_2$ geodesic.

the explicit primitive representation $\|C_\alpha - C_\beta\|_{L^1(\mathbb{R})}$.

Other classical one-dimensional distances are obtained by replacing the L^1 norm of cumulative functions with L^2 or L^∞ norms; under suitable tightness assumptions, such norms also metrize convergence in law and lead for instance to Cramér–von Mises and Kolmogorov–Smirnov-type distances.

OT on trees. The line is the simplest tree. On a general tree, the order is no longer total, but each edge still defines a cut and hence a cumulative imbalance. This gives an exact formula for the $\mathcal{W}_1$ cost associated with the tree geodesic distance. The formula is classical in fast earth-mover computations on tree metrics [342] and is also the mechanism behind tree-sliced Wasserstein distances [328].

Proposition 2.48 (Cumulative formula on a tree)**.** *Let $\mathsf{T} = (V, E, \ell)$ be a finite weighted tree, where each edge e has length $\ell_e > 0$, and let d_T be its geodesic distance. Choose a root o. For an edge $e = (u, v)$ oriented away from o, let V_e be the set of vertices in the connected component containing v after removing e. If $\alpha, \beta \in \mathcal{P}(V)$, then*

$$\mathcal{W}_{1,d_\mathsf{T}}(\alpha,\beta) = \min_{\pi\in\mathcal{U}(\alpha,\beta)} \sum_{x,y\in V} d_\mathsf{T}(x,y)\pi_{xy} = \sum_{e\in E} \ell_e \,|\alpha(V_e) - \beta(V_e)|\,. \tag{2.14}$$

The value is computable in $O(|V|)$ arithmetic operations once the signed masses $\alpha - \beta$ are stored on the vertices, by one postorder traversal of the rooted tree.

Proof. Let $\sigma = \alpha - \beta$. Removing an edge e splits the tree into V_e and its complement. For any coupling π, the net amount of mass crossing e from V_e to $V \setminus V_e$ minus the amount crossing in the reverse direction is fixed and equals $\sigma(V_e)$. Hence the total amount of transported mass whose path uses e is at least $|\sigma(V_e)|$. Since every unit of mass crossing e pays the length ℓ_e, summing over edges gives

$$\sum_{x,y} d_\mathsf{T}(x,y)\pi_{x,y} = \sum_{e\in E} \ell_e \sum_{x,y:\, e\in[x,y]} \pi_{x,y} \geqslant \sum_{e\in E} \ell_e |\sigma(V_e)|,$$

where $[x, y]$ is the unique path between x and y.

It remains to realize this lower bound. Process the rooted tree from the leaves to the root. At a vertex $v \neq o$, compute the total signed mass in the subtree below v, namely $F_e = \sigma(V_e)$ for the edge e joining v to its parent. If $F_e > 0$, send F_e units of surplus from this subtree to the parent side; if $F_e < 0$, import $-F_e$ units from the parent side into the subtree. After this operation the subtree has zero net imbalance and can be collapsed into its parent. Continuing upward balances all subtrees because $\sigma(V) = 0$. Decomposing these edge fluxes into source-to-target paths yields a coupling whose traffic across each edge is exactly $|F_e|$, and therefore whose cost is the right-hand side of (2.14). The same postorder pass computes all F_e, proving the complexity claim. □

For a chain rooted at one end, the sets V_e are rays, so (2.14) is the discrete version of the one-dimensional identity (2.12). This edgewise cumulative decoupling is specific to the geodesic-distance cost d_T, i.e. to $\mathcal{W}_1$. For powers d_T^p with $p > 1$, the term $\big(\sum_{e\in[x,y]} \ell_e\big)^p$ couples all edges along a path, so the simple sum of absolute subtree imbalances no longer gives the optimal value. The tree structure remains algorithmically useful in a different sense: tree metrics give fast surrogates and embeddings for EMD-type computations on histograms [289, 15, 342], and randomized or data-adapted trees lead

to tree-sliced Wasserstein distances that trade the exact Euclidean ground metric for much faster one-dimensional/tree computations [328].

The same tree can nevertheless be used as a genuine geodesic space. If π is an optimal coupling for the squared tree distance d_{T}^2, the associated displacement interpolation moves each packet of mass $\pi_{i,j}$ at constant speed along the unique path from x_i to x_j. Figure 2.12 shows this tree analogue of McCann interpolation. The intermediate measures are not necessarily supported on the original vertices: mass may sit inside edges while it travels through the branching structure.

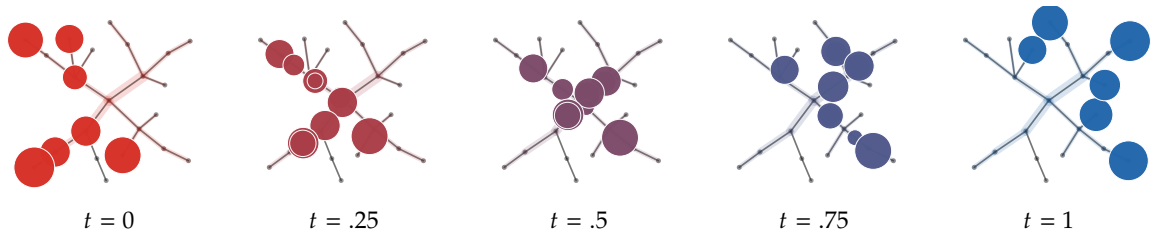


Figure 2.12: McCann interpolation on a finite tree. A quadratic optimal plan is computed between two non-uniform vertex histograms for the squared geodesic distance on the tree. Each transported packet then moves along the unique tree path connecting its source and target vertices. Circle areas encode transported masses, colors interpolate from the source measure α in red to the target measure β in blue, and faint colored corridors mark the tree branches carrying transport.

Triangular rearrangements. There is another canonical way to build transport maps in several dimensions: transport one coordinate at a time by conditional one-dimensional quantiles. This construction goes back to Knothe and Rosenblatt [310, 454]. It is not usually cost-optimal, but it gives a deterministic rearrangement under weak assumptions and clarifies how multivariate transport can be reduced recursively to scalar monotone maps.

Proposition 2.49 (Knothe–Rosenblatt triangular rearrangement)**.** *Let $\alpha, \beta \in \mathcal{M}_+^1(\mathbb{R}^d)$. Assume that the first marginal of α and, recursively, the one-dimensional conditional source laws used below are atomless almost everywhere. Fix measurable versions of all regular conditional distributions. Then there is a triangular map*

$$T(x_1, \dots, x_d) = (T_1(x_1), T_2(x_1, x_2), \dots, T_d(x_1, \dots, x_d))$$

such that $T_\sharp \alpha = \beta$ and, for each k, the function $x_k \mapsto T_k(x_1, \dots, x_k)$ is nondecreasing for $\alpha_{<k}$-almost every $x_{<k} = (x_1, \dots, x_{k-1})$.

Proof. The construction is recursive. For $k = 1$, let T_1 be the monotone rearrangement between the first marginal of α and the first marginal of β. Suppose that $T_1, \dots, T_{k-1}$ have already been constructed. Write $x_{<k} = (x_1, \dots, x_{k-1})$ and $T_{<k} = (T_1, \dots, T_{k-1})$. By induction, $(T_{<k})_\sharp \alpha_{<k} = \beta_{<k}$, where $\alpha_{<k}$ and $\beta_{<k}$ are the first $(k-1)$-coordinate marginals. Let $\alpha^k_{x_{<k}}$ and $\beta^k_{y_{<k}}$ be regular conditional laws of the k-th coordinate given the previous coordinates. Define $T_k(x_{<k}, \cdot)$ as the one-dimensional monotone rearrangement from $\alpha^k_{x_{<k}}$ to $\beta^k_{T_{<k}(x_{<k})}$.

The map $T_k(x_{<k}, \cdot)$ is nondecreasing by the one-dimensional rearrangement theorem. The chain rule for disintegrations then shows that, after step k, the first k coordinates of $T_\sharp \alpha$ have the same law as the first k coordinates of β. At $k = d$ this gives $T_\sharp \alpha = \beta$. Target atoms are handled by generalized quantiles. If a source conditional law has atoms that must be split, the same recursive construction defines a triangular Markov kernel rather than a deterministic map. □

The recursive construction in the proof is an algorithm: it repeatedly applies the one-dimensional quantile rearrangement to conditional distributions.

Figure 2.13 shows the two-dimensional mechanism on image histograms. The first stage transports only the horizontal marginal, so the middle pivot has the same x-marginal as the target but keeps the source vertical conditionals. The second stage then transports each vertical conditional law inside the corresponding column.

Algorithm 2.1: Knothe–Rosenblatt triangular rearrangement

Input: Probability measures α, β on $\mathbb{R}^d$ with regular conditional laws, with every source conditional law used below atomless.
Output: Knothe–Rosenblatt triangular map T.
Compute first-coordinate rearrangement: $T_1 = (F_{\beta_1})^{-1} \circ F_{\alpha_1}$.
For $k = 2, \dots, d$ **do**:
  Set $x_{<k} = (x_1, \dots, x_{k-1})$.
  Compute conditional laws $\alpha^k_{x_{<k}}$ and $\beta^k_{T_{<k}(x_{<k})}$.
  Set $T_k(x_{<k}, x_k) = \left(F_{\beta^k_{T_{<k}(x_{<k})}}\right)^{-1} \circ F_{\alpha^k_{x_{<k}}}(x_k)$.
Return $T(x) = (T_1(x_1), T_2(x_1, x_2), \dots, T_d(x_1, \dots, x_d))$.

source x-move x-move pivot y-move y-move target

Figure 2.13: Triangular rearrangement between the same cat and heart densities as in Figure 2.6. The panels are computed directly on image histograms. The first three transitions move mass horizontally by the monotone rearrangement between the x-marginals; the pivot has the target horizontal marginal. The last three transitions keep each column fixed and move mass vertically by one-dimensional monotone rearrangements between conditional laws.

This construction transports successively along coordinate axes and is therefore often called axis-wise transport. It depends on the chosen ordering of coordinates and is not generally optimal for the quadratic cost. It is nevertheless a useful limiting object: Brenier maps for increasingly anisotropic quadratic costs converge to triangular rearrangements under suitable assumptions [112].

Proposition 2.50 (Anisotropic Brenier maps converge to Knothe–Rosenblatt)**.** *Let α, β be compactly supported probability measures on $\mathbb{R}^d$ with densities bounded above and below on their rectangular supports. In particular, the source and target conditional laws entering the Knothe–Rosenblatt construction are atomless almost everywhere. For $\varepsilon > 0$, set*

$$c_\varepsilon(x, y) := \sum_{k=1}^{d} \varepsilon^{k-1} |x_k - y_k|^2 .$$

Let T_ε be the Monge map from α to β for the cost c_ε, and let T_{KR} be the triangular Knothe–Rosenblatt rearrangement with the coordinate order used above. Then

$$T_\varepsilon \longrightarrow T_{\mathrm{KR}} \qquad \text{in } L^2(\alpha; \mathbb{R}^d) \qquad \text{as } \varepsilon \to 0.$$

Proof. We summarize the scale-separated argument of [112]. Let $\pi_\varepsilon = (\mathrm{Id}, T_\varepsilon)_\sharp \alpha$. For a coupling γ, set $I_k(\gamma) := \int |x_k - y_k|^2 \, \mathrm{d}\gamma(x, y)$ and $F_\varepsilon(\gamma) := \sum_{k=1}^d \varepsilon^{k-1} I_k(\gamma)$. Since the supports are compact, every sequence $\varepsilon \downarrow 0$ has a subsequence for which π_ε converges weakly to a coupling π between α and β. The coordinate costs are bounded and continuous on the common compact support. Dividing F_ε by its leading coefficient and passing to the limit shows that π minimizes I_1. The first source marginal is atomless, so the (x_1, y_1)-marginal of π is the unique monotone coupling; equivalently, $y_1 = T_1(x_1)$ under π.

The next scale identifies the second coordinate. Optimality against the Knothe coupling π_{KR} gives $F_\varepsilon(\pi_\varepsilon) \leqslant F_\varepsilon(\pi_{\mathrm{KR}})$. Since $I_1(\pi_\varepsilon)$ is bounded below by the optimal first-coordinate value $I_1(\pi_{\mathrm{KR}})$, cancellation and division by ε yield

$$\sum_{k=2}^{d} \varepsilon^{k-2} I_k(\pi_\varepsilon) \leqslant \sum_{k=2}^{d} \varepsilon^{k-2} I_k(\pi_{\mathrm{KR}}).$$

Passing to the limit shows that π minimizes I_2 among couplings with the already identified first-coordinate marginal. Disintegration with respect to (x_1, y_1) turns this into the monotone transport between the conditional laws of x_2 and y_2, so $y_2 = T_2(x_1, x_2)$ under π.

The higher-coordinate induction repeats this comparison after disintegrating over the already identified coordinates. Its technical point is to control the residual suboptimality of the earlier coordinates before dividing by the next power of ε; the atomlessness of both source and target conditionals makes the preceding monotone couplings invertible almost everywhere. The scale-separated estimate proved in [112] then gives successively $y_k = T_k(x_1, \dots, x_k)$ for every k. Consequently, every weak limit is the graph coupling $\pi_{\mathrm{KR}} = (\mathrm{Id}, T_{\mathrm{KR}})_\sharp \alpha$, and uniqueness of this limit gives convergence of the whole family.

Finally, let $X \sim \alpha$. The graph couplings are the laws of $(X, T_\varepsilon(X))$, and they converge weakly to the law of $(X, T_{\mathrm{KR}}(X))$. To turn this into convergence of maps, fix $\zeta > 0$. By Lusin's theorem, there is a compact set K with $\alpha(K) > 1 - \zeta$ on which T_{KR} is continuous. On K, the set

$$\{(x, y) \; ; \; x \in K, \; \|y - T_{\mathrm{KR}}(x)\| \geqslant \delta\}$$

is closed and has zero mass under the limiting graph coupling. Portmanteau's theorem gives

$$\limsup_{\varepsilon \to 0} \alpha(\{x \; ; \; x \in K, \; \|T_\varepsilon(x) - T_{\mathrm{KR}}(x)\| \geqslant \delta\}) = 0.$$

Adding the complement of K and letting $\zeta \to 0$ proves convergence in α-probability. Since all maps take values in a common compact set, their squared differences are uniformly integrable, and convergence therefore holds in $L^2(\alpha)$. □

2.7 Gaussian Measures and the Bures Metric

Gaussian measures form the most important finite-dimensional family preserved by quadratic optimal transport. The mean moves linearly, while the covariance follows the Bures–Wasserstein geometry of positive semidefinite matrices.

One-dimensional Gaussians. Let $\alpha = \mathcal{N}(m_\alpha, \sigma_\alpha^2)$ and $\beta = \mathcal{N}(m_\beta, \sigma_\beta^2)$ be two nondegenerate Gaussians on $\mathbb{R}$. Then

$$T(x) = \frac{\sigma_\beta}{\sigma_\alpha}(x - m_\alpha) + m_\beta$$

satisfies $T_\sharp \alpha = \beta$. It is the derivative of the convex function

$$\varphi(x) = \frac{\sigma_\beta}{2\sigma_\alpha}(x - m_\alpha)^2 + m_\beta x,$$

so Brenier's theorem shows that it is the optimal quadratic transport. The associated distance is

$$\mathcal{W}_2(\alpha, \beta)^2 = \int_{\mathbb{R}} \left(\frac{\sigma_\beta}{\sigma_\alpha}(x - m_\alpha) + m_\beta - x \right)^2 \mathrm{d}\alpha(x) = (m_\alpha - m_\beta)^2 + (\sigma_\alpha - \sigma_\beta)^2.$$

The formula extends by continuity to Dirac masses, although the affine Monge map itself only pushes a Dirac source to another Dirac. Thus the OT geometry of one-dimensional Gaussians is the Euclidean geometry of the closed half-plane $(m, \sigma) \in \mathbb{R} \times \mathbb{R}_+$. By contrast, the Fisher–Rao boundary $\sigma = 0$ is infinitely far from every nondegenerate Gaussian, and the KL divergence from a nondegenerate Gaussian to a singular one is infinite.

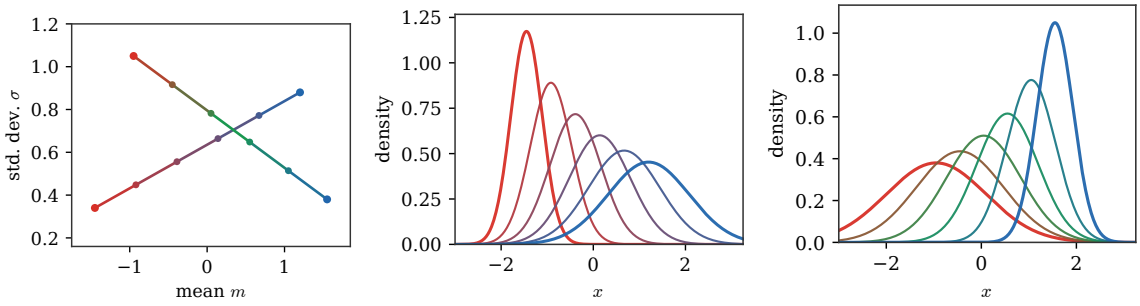
Figure 2.14 displays this Euclidean half-plane geometry and the corresponding displacement interpolation of the Gaussian densities.

Remark 2.51 (Comparison with the Fisher–Rao metric with mean variation). The previous formula shows that the Wasserstein geometry of one-dimensional Gaussians is Euclidean in (m, σ). The Fisher–Rao geometry obtained from the local expansion of the Kullback–Leibler divergence is different as soon as the mean is allowed to vary [157]. For $\sigma, \sigma' > 0$,

$$\mathrm{KL}\left(\mathcal{N}(m, \sigma^2) \mid \mathcal{N}(m', (\sigma')^2)\right) = \log \frac{\sigma'}{\sigma} + \frac{\sigma^2 + (m - m')^2}{2(\sigma')^2} - \frac{1}{2}.$$

Expanding the first argument around (m, σ) gives, for increments (h, s),

$$\mathrm{KL}\left(\mathcal{N}(m + \varepsilon h, (\sigma + \varepsilon s)^2) \mid \mathcal{N}(m, \sigma^2)\right) = \frac{\varepsilon^2}{2}\left(\frac{h^2}{\sigma^2} + 2\frac{s^2}{\sigma^2}\right) + o(\varepsilon^2).$$



(m, σ) half-plane | first geodesic | second geodesic

Figure 2.14: One-dimensional Gaussian $\mathcal{W}_2$ geodesics. In the coordinates (m, σ), the geodesics are Euclidean segments in the upper half-plane. Both paths share red and blue endpoint colors; the first interpolates directly between them, whereas the second passes through green at mid-time. The same palettes identify both segments in the left panel.

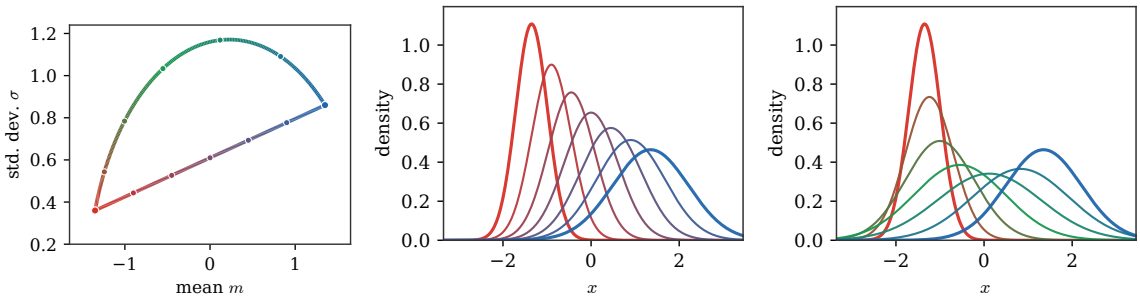
Thus the Fisher–Rao metric on the Gaussian half-plane is

$$g^{\mathrm{FR}}_{(m,\sigma)}((h,s),(h',s')) = \frac{hh' + 2ss'}{\sigma^2}.$$

Setting $z = m/\sqrt{2} + i\sigma$ identifies this metric with twice the usual hyperbolic metric on the upper half-plane. Its geodesics are therefore vertical lines or Euclidean semicircles orthogonal to the boundary $\sigma = 0$ after this horizontal rescaling. Consequently

$$d_{\mathrm{FR}}\big(\mathcal{N}(m,\sigma^2), \mathcal{N}(m',(\sigma')^2)\big) = \sqrt{2}\,\operatorname{arcosh}\left(1 + \frac{(m-m')^2 + 2(\sigma-\sigma')^2}{4\sigma\sigma'}\right).$$

In contrast, $\mathcal{W}_2^2 = (m-m')^2 + (\sigma-\sigma')^2$. Wasserstein geodesics are straight segments in the (m, σ) half-plane and reach $\sigma = 0$ at finite distance, whereas Fisher–Rao geodesics bend away from the boundary and the boundary $\sigma = 0$ is infinitely far away. Figure 2.15 displays the resulting difference in both parameter space and density space.



parameter geodesics | $\mathcal{W}_2$ interpolation | Fisher–Rao interpolation

Figure 2.15: Wasserstein and Fisher–Rao geodesics in the one-dimensional Gaussian family. Both interpolations share red and blue endpoint colors. The Wasserstein curves interpolate directly between them, whereas the Fisher–Rao curves pass through green at mid-time; the same palettes identify both geodesics in the left parameter-space panel. The two density panels use the same endpoint Gaussians and time samples, but the Fisher–Rao path expands the standard deviation along its hyperbolic arc before returning to the target scale.

Multivariate Gaussians. As in the scalar case, a natural approach is to seek the optimal transport map within the class of affine maps. If

$$\alpha = \mathcal{N}(m_\alpha, \Sigma_\alpha), \qquad \beta = \mathcal{N}(m_\beta, \Sigma_\beta), \qquad T(x) = m_\beta + A(x - m_\alpha), \tag{2.15}$$

then T is the gradient of the convex quadratic potential $\varphi(x) = \langle m_\beta,\, x\rangle + \langle A(x - m_\alpha),\, x - m_\alpha\rangle/2$ if and only if A is symmetric positive semidefinite.

Proposition 2.52 (Affine push-forward of Gaussians)**.** *One has $T_\sharp\alpha = \beta$ if and only if*

$$A\Sigma_\alpha A^\top = \Sigma_\beta. \tag{2.16}$$

Proof. An affine function maps a Gaussian to a Gaussian, so the law of $T(X)$ is determined by its mean and covariance. If $X \sim \alpha$ and $Y = T(X)$, then $\mathbb{E}(Y) = m_\beta + A\mathbb{E}(X - m_\alpha) = m_\beta$, and

$$\mathbb{E}((Y - m_\beta)(Y - m_\beta)^\top) = A\mathbb{E}((X - m_\alpha)(X - m_\alpha)^\top)A^\top = A\Sigma_\alpha A^\top.$$

Thus $A\Sigma_\alpha A^\top = \Sigma_\beta$ is necessary and sufficient for Y to have the same mean and covariance as β; since both laws are Gaussian, this is equivalent to equality in distribution. □

The covariance equation is quadratic in A. Under the symmetry constraint imposed by Brenier's theorem, it becomes $A\Sigma_\alpha A = \Sigma_\beta$. The covariance part of the resulting cost is a matrix metric.

Definition 2.53 (Bures metric)**.** For positive semidefinite covariance matrices Σ and Λ, the Bures metric is

$$\mathcal{B}(\Sigma, \Lambda)^2 := \operatorname{tr}\left(\Sigma + \Lambda - 2(\Sigma^{1/2}\Lambda\Sigma^{1/2})^{1/2}\right). \tag{2.17}$$

The next proposition solves the covariance equation and shows that this metric is exactly the covariance contribution to Gaussian $\mathcal{W}_2$.

Proposition 2.54 (Gaussian $\mathcal{W}_2$ formula and Bures covariance term)**.** *Assume that Σ_α and Σ_β are positive definite. The covariance equation and its unique symmetric positive-definite solution are*

$$A\Sigma_\alpha A = \Sigma_\beta, \quad A \succ 0, \qquad \Longleftrightarrow \qquad A = \Sigma_\alpha^{-1/2}\left(\Sigma_\alpha^{1/2}\Sigma_\beta\Sigma_\alpha^{1/2}\right)^{1/2}\Sigma_\alpha^{-1/2}. \tag{2.18}$$

The affine map $T(x) = m_\beta + A(x - m_\alpha)$ is the optimal quadratic-cost transport from $\mathcal{N}(m_\alpha, \Sigma_\alpha)$ to $\mathcal{N}(m_\beta, \Sigma_\beta)$, and

$$\mathcal{W}_2(\alpha, \beta)^2 = \|m_\alpha - m_\beta\|^2 + \mathcal{B}(\Sigma_\alpha, \Sigma_\beta)^2, \tag{2.19}$$

where $\mathcal{B}$ is the Bures metric of Definition 2.53.

Proof. Multiplying $A\Sigma_\alpha A = \Sigma_\beta$ on the left and right by $\Sigma_\alpha^{1/2}$ gives $(\Sigma_\alpha^{1/2}A\Sigma_\alpha^{1/2})^2 = \Sigma_\alpha^{1/2}\Sigma_\beta\Sigma_\alpha^{1/2}$. Since A is symmetric positive, $\Sigma_\alpha^{1/2}A\Sigma_\alpha^{1/2}$ is symmetric positive and is therefore the unique positive square root of the right-hand side. Conversely, the matrix in (2.18) is symmetric positive and satisfies the covariance equation.

By Proposition 2.52, this affine map pushes α to β. It is the gradient of a convex quadratic potential, so Brenier's theorem implies optimality. If $X \sim \alpha$, then

$$\begin{aligned}
\mathbb{E}\|X - T(X)\|^2 &= \|m_\alpha - m_\beta\|^2 + \mathbb{E}\|(\mathrm{Id} - A)(X - m_\alpha)\|^2 \\
&= \|m_\alpha - m_\beta\|^2 + \operatorname{tr}((\mathrm{Id} - A)\Sigma_\alpha(\mathrm{Id} - A)^\top) \\
&= \|m_\alpha - m_\beta\|^2 + \operatorname{tr}(\Sigma_\alpha) + \operatorname{tr}(A\Sigma_\alpha A) - 2\operatorname{tr}(A\Sigma_\alpha) \\
&= \|m_\alpha - m_\beta\|^2 + \operatorname{tr}(\Sigma_\alpha) + \operatorname{tr}(\Sigma_\beta) - 2\operatorname{tr}((\Sigma_\alpha^{1/2}\Sigma_\beta\Sigma_\alpha^{1/2})^{1/2}),
\end{aligned}$$

which is the desired expression. The formula for singular covariance matrices follows by adding $\eta\mathrm{Id}$ and letting $\eta \downarrow 0$. □

Remark 2.55 (Common-generator elliptical laws). The Gaussian formula has a useful, but deliberately limited, extension beyond Gaussian tails. Fix an absolutely continuous, centered, radial probability measure ρ on $\mathbb{R}^d$ with covariance Id. For positive definite matrices $\Sigma_\alpha, \Sigma_\beta$, define two elliptically contoured laws with this same radial generator by

$$\alpha = (z \mapsto m_\alpha + \Sigma_\alpha^{1/2}z)_\sharp\rho, \qquad \beta = (z \mapsto m_\beta + \Sigma_\beta^{1/2}z)_\sharp\rho.$$

For this pair, the same Brenier matrix (2.18) gives the optimal map $T(x) = m_\beta + A(x - m_\alpha)$, and the same formula (2.19) holds. Indeed, if $X = m_\alpha + \Sigma_\alpha^{1/2}Z$ with $Z \sim \rho$, then $A\Sigma_\alpha A = \Sigma_\beta$, hence $A\Sigma_\alpha^{1/2} = \Sigma_\beta^{1/2}Q$ for some orthogonal matrix Q. Since ρ is radial, QZ has the same law as Z, so $T_\sharp\alpha = \beta$.

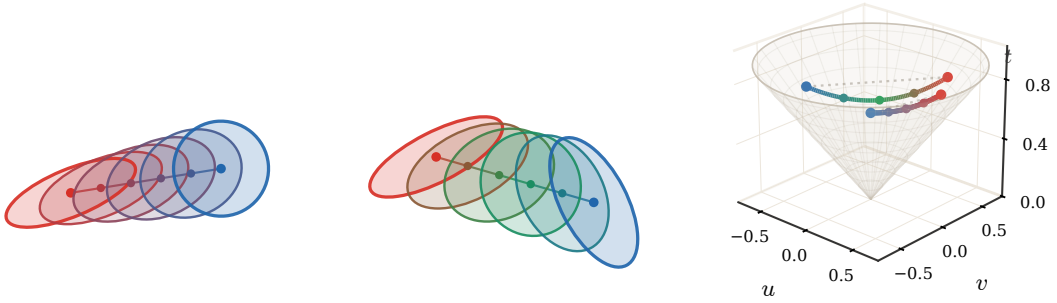
Since A is symmetric positive definite, T is the gradient of a convex quadratic potential and Brenier's theorem gives optimality. Finally, because ρ is centered with covariance Id, the transport cost depends only on the means and covariance matrices and gives the Bures expression. The common-generator assumption is the point: for two elliptical laws with different radial profiles, one must also transport the radial variable, and the covariance/Bures term is no longer the full transport cost.

The covariance term $\mathcal{B}$ is the Bures–Wasserstein metric on positive semidefinite matrices [89, 236, 54]. It separates the Euclidean displacement of the mean from the intrinsic transport geometry of covariance ellipsoids.

Remark 2.56 (The 2×2 covariance cone). For 2×2 covariance matrices, the geometry can be visualized without suppressing any degree of freedom. The following orthonormal coordinates for the Frobenius inner product separate the trace t, diagonal anisotropy u and correlation v:

$$\Sigma = \begin{pmatrix} a & c \\ c & b \end{pmatrix} \quad \longleftrightarrow \quad (t,u,v) = \frac{1}{\sqrt{2}}(a+b, a-b, 2c), \qquad \Sigma = \frac{1}{\sqrt{2}} \begin{pmatrix} t+u & v \\ v & t-u \end{pmatrix}, \tag{2.20}$$
$$2\det(\Sigma) = t^2 - u^2 - v^2, \qquad \mathbb{S}^2_+ \quad \longleftrightarrow \quad \mathcal{K}_3 := \left\{(t,u,v) \in \mathbb{R}^3 : t \geqslant \sqrt{u^2+v^2}\right\}.$$

Thus $\mathbb{S}^2_+$ is the three-dimensional Lorentz, or ice-cream, cone: its apex represents the zero matrix, its nonzero boundary represents rank-one covariances, and its interior represents positive definite covariances. This identification describes the set of admissible matrices, not its Bures geometry. In particular, the Bures distance is not the ambient Euclidean distance of $\mathbb{R}^3$.
Figure 2.16 makes this distinction visible. The first two panels show two Bures–Wasserstein interpolations through covariance ellipses, and the right panel places the same paths in $\mathcal{K}_3$. Their colored curves remain inside the cone but generally bend away from the faint gray Euclidean chords, because they are geodesics for the Bures metric rather than for the ambient Frobenius metric.



anisotropic to isotropic rotated anisotropies covariance cone

Figure 2.16: Two-dimensional Gaussian $\mathcal{W}_2$ geodesics. Means move linearly, while covariance ellipses follow the Bures–Wasserstein interpolation. The left path uses a direct red-to-blue palette; the middle path shares the same endpoint colors but passes through green. The right panel displays these same two palettes along the corresponding covariance paths in the 2×2 positive-semidefinite cone, with u and v horizontal and t vertical; faint gray chords show the ambient Euclidean segments for comparison.

Alternate formulation of the Bures metric. Beyond its Gaussian origin, the Bures metric can be recovered from two simpler ambient geometries. It is induced either by $\mathcal{W}_2$ after retaining only second moments, or by the Frobenius norm after retaining only the covariance product $AA^\top$.

The first interpretation is distribution-free: it identifies all probability measures with the same second-moment matrix.

Proposition 2.57 (Bures metric as a second-moment quotient). *For $\alpha \in \mathcal{P}_2(\mathbb{R}^d)$, set its raw second-moment matrix to be*

$$\Phi_2(\alpha) := \int_{\mathbb{R}^d} uu^\top \mathrm{d}\alpha(u) \in \mathbb{S}^d_+.$$

For $\Sigma, \Lambda \in \mathbb{S}^d_+$,

$$\min_{\Phi_2(\alpha)=\Sigma,\ \Phi_2(\beta)=\Lambda} \mathcal{W}_2(\alpha,\beta)^2 = \mathcal{B}(\Sigma,\Lambda)^2.$$

In other words, the Bures metric is the Wasserstein quotient distance induced on positive semidefinite matrices by identifying all probability measures with the same raw second-moment matrix.

Proof. Assume first that Σ and Λ are positive definite. Fix two admissible laws α, β and an arbitrary coupling π between them, and write $K := \int_{\mathbb{R}^d \times \mathbb{R}^d} uv^\top \mathrm{d}\pi(u,v)$. Then $\int \|u-v\|^2 \mathrm{d}\pi(u,v) = \mathrm{tr}(\Sigma) + \mathrm{tr}(\Lambda) - 2\,\mathrm{tr}(K)$. The block second-moment matrix

$$\int \begin{pmatrix} u \\ v \end{pmatrix} \begin{pmatrix} u \\ v \end{pmatrix}^\top \mathrm{d}\pi(u,v) = \begin{pmatrix} \Sigma & K \\ K^\top & \Lambda \end{pmatrix}$$

is positive semidefinite. The Schur complement gives $R := \Sigma^{-1/2} K \Lambda^{-1/2}$ with $RR^\top \preceq \mathrm{Id}$, so $K = \Sigma^{1/2} R \Lambda^{1/2}$ with $\|R\|_{\mathrm{op}} \leqslant 1$. By duality between the nuclear and operator norms,

$$\mathrm{tr}(K) \leqslant \sup_{\|R\|_{\mathrm{op}} \leqslant 1} \mathrm{tr}(\Sigma^{1/2} R \Lambda^{1/2}) = \|\Lambda^{1/2} \Sigma^{1/2}\|_* = \mathrm{tr}\left((\Sigma^{1/2} \Lambda \Sigma^{1/2})^{1/2} \right).$$

For singular matrices, fix $\eta > 0$. Adding $\eta \mathrm{Id}$ to the two diagonal blocks preserves block positivity, so the same Schur-complement argument gives

$$\mathrm{tr}(K) \leqslant \mathrm{tr}\left(((\Sigma + \eta \mathrm{Id})^{1/2} (\Lambda + \eta \mathrm{Id}) (\Sigma + \eta \mathrm{Id})^{1/2})^{1/2} \right).$$

Letting $\eta \downarrow 0$ and using continuity of the matrix square root gives the same trace bound with Σ, Λ. Hence every admissible coupling has cost at least $\mathcal{B}(\Sigma, \Lambda)^2$.

Conversely, choose the centered Gaussian laws $\mathcal{N}(0, \Sigma)$ and $\mathcal{N}(0, \Lambda)$. They satisfy the two second-moment constraints, and Proposition 2.54 gives $\mathcal{W}_2(\mathcal{N}(0,\Sigma), \mathcal{N}(0,\Lambda))^2 = \mathcal{B}(\Sigma,\Lambda)^2$, again by continuity in the singular case. This admissible pair attains the lower bound, proving the claim. □

The second interpretation is purely matrix-based: it identifies all matrix factors with the same covariance product.

Proposition 2.58 (Bures metric as a Procrustes metric)**.** *Let $\Sigma, \Lambda \in \mathbb{S}^d_+$, let $d' \geqslant d$, and fix any $A_0, B_0 \in \mathbb{R}^{d \times d'}$ such that $A_0 A_0^\top = \Sigma$ and $B_0 B_0^\top = \Lambda$. Then*

$$\mathcal{B}(\Sigma,\Lambda)^2 = \min_{\substack{A,B \in \mathbb{R}^{d \times d'} \\ AA^\top = \Sigma,\ BB^\top = \Lambda}} \|A - B\|_{\mathrm{F}}^2 = \min_{R \in \mathrm{O}(d')} \|A_0 - B_0 R\|_{\mathrm{F}}^2. \tag{2.21}$$

For $d' = d$, one may in particular take $A_0 = \Sigma^{1/2}$ and $B_0 = \Lambda^{1/2}$; the second minimum is then over $\mathrm{O}(d)$.

Proof. Two matrices in $\mathbb{R}^{d \times d'}$ have the same row-Gram matrix if and only if they differ by right multiplication by an orthogonal matrix. Indeed, equality of the Gram matrices makes the linear map sending each row of one matrix to the corresponding row of the other well defined and isometric on their row spans; this isometry extends to an element of $\mathrm{O}(d')$. Therefore every admissible pair has the form $A = A_0 U$ and $B = B_0 V$ for some $U, V \in \mathrm{O}(d')$. Right invariance of the Frobenius norm gives

$$\min_{U,V \in \mathrm{O}(d')} \|A_0 U - B_0 V\|_{\mathrm{F}}^2 = \min_{R \in \mathrm{O}(d')} \|A_0 - B_0 R\|_{\mathrm{F}}^2,$$

where one may take $R = VU^\top$. Thus the two factor minima in (2.21) coincide and are independent of the chosen representatives A_0, B_0.

Expanding the last objective and applying the orthogonal Procrustes formula, which follows directly from a singular-value decomposition, yields

$$\min_{R \in \mathrm{O}(d')} \|A_0 - B_0 R\|_{\mathrm{F}}^2 = \mathrm{tr}(\Sigma) + \mathrm{tr}(\Lambda) - 2\|A_0^\top B_0\|_*.$$

Moreover, $(A_0^\top B_0)^\top (A_0^\top B_0) = B_0^\top \Sigma B_0$, whose nonzero eigenvalues are those of $\Sigma^{1/2} B_0 B_0^\top \Sigma^{1/2} = \Sigma^{1/2} \Lambda \Sigma^{1/2}$. Consequently, $\|A_0^\top B_0\|_* = \mathrm{tr}((\Sigma^{1/2} \Lambda \Sigma^{1/2})^{1/2})$. Definition 2.53 now identifies the resulting value with $\mathcal{B}(\Sigma,\Lambda)^2$. □

Thus the map $A \mapsto AA^\top$ identifies $\mathbb{S}^d_+$ with the orbit space $\mathbb{R}^{d\times d'}/\mathrm{O}(d')$, and $\mathcal{B}$ is the quotient distance induced by the Frobenius metric. This is the matrix counterpart of the quotient transport geometries developed in Section 11.3.

These two representations make the metric and convexity properties of $\mathcal{B}$ transparent.

Proposition 2.59 (Metric and convexity properties of the Bures term)**.** *The function $\mathcal{B}$ is a distance on positive semidefinite covariance matrices. Moreover, $\mathcal{B}^2$ is jointly convex: for $t \in [0,1]$,*

$$\mathcal{B}^2((1-t)\Sigma_0 + t\Sigma_1, (1-t)\Lambda_0 + t\Lambda_1) \leqslant (1-t)\mathcal{B}^2(\Sigma_0,\Lambda_0) + t\mathcal{B}^2(\Sigma_1,\Lambda_1).$$

Proof. The free-factor formula in Proposition 2.58 is symmetric and nonnegative. Since its minima are attained, it vanishes if and only if two factors coincide, which is equivalent to $\Sigma = \Lambda$. For the triangle inequality, use the square-factor specialization and let Q_1, Q_2 be optimal orthogonal alignments for (Σ,Λ) and (Λ,Γ), respectively. Since Q_2Q_1 is admissible for (Σ,Γ),

$$\mathcal{B}(\Sigma,\Gamma) \leqslant \|\Sigma^{1/2} - \Gamma^{1/2}Q_2Q_1\|_{\mathrm{F}} \leqslant \|\Sigma^{1/2} - \Lambda^{1/2}Q_1\|_{\mathrm{F}} + \|\Lambda^{1/2} - \Gamma^{1/2}Q_2\|_{\mathrm{F}} = \mathcal{B}(\Sigma,\Lambda) + \mathcal{B}(\Lambda,\Gamma).$$

For convexity, use the free-factor formula in Proposition 2.58. For $i \in \{0,1\}$, choose optimal square factors (U_i, V_i) for (Σ_i,Λ_i) and define the block factors

$$U_t = [\sqrt{1-t}\,U_0, \sqrt{t}\,U_1], \qquad V_t = [\sqrt{1-t}\,V_0, \sqrt{t}\,V_1].$$

Set $\Sigma_t = (1-t)\Sigma_0 + t\Sigma_1$ and $\Lambda_t = (1-t)\Lambda_0 + t\Lambda_1$. Then $U_tU_t^\top = \Sigma_t$ and $V_tV_t^\top = \Lambda_t$. Applying the free-factor formula with $2d$ columns gives

$$\mathcal{B}^2(\Sigma_t,\Lambda_t) \leqslant \|U_t - V_t\|_{\mathrm{F}}^2 = (1-t)\|U_0 - V_0\|_{\mathrm{F}}^2 + t\|U_1 - V_1\|_{\mathrm{F}}^2 = (1-t)\mathcal{B}^2(\Sigma_0,\Lambda_0) + t\mathcal{B}^2(\Sigma_1,\Lambda_1),$$

which proves joint convexity. □

> **Remark 2.60 (Diagonal covariances and Hellinger geometry).** If $\Sigma_\alpha = \mathrm{diag}(r_i)_i$ and $\Sigma_\beta = \mathrm{diag}(s_i)_i$ are diagonal, the Bures metric reduces to the Euclidean distance between square roots, $\mathcal{B}(\Sigma_\alpha,\Sigma_\beta) = \|\sqrt{r} - \sqrt{s}\|_2$. This is the finite-dimensional Hellinger geometry on the nonnegative covariance coordinates: variances are compared after the amplitude change of variables $r_i \mapsto \sqrt{r_i}$.

Comparison with the Fisher–Rao metric for zero-mean Gaussians. The Bures metric is the covariance geometry induced by quadratic optimal transport. A different, information-geometric metric is obtained by expanding the Kullback–Leibler divergence between centered Gaussians. For $\Sigma,\Sigma' \in \mathbb{S}^d_{++}$, write

$$\mathrm{KL}(\Sigma \mid \Sigma') := \mathrm{KL}\left(\mathcal{N}(0,\Sigma) \mid \mathcal{N}(0,\Sigma')\right) = \frac{1}{2}\left(\mathrm{tr}((\Sigma')^{-1}\Sigma) - d - \log\det((\Sigma')^{-1}\Sigma)\right).$$

If $D = D^\top$ and $\Sigma + \varepsilon D$ remains positive definite, then

$$\mathrm{KL}(\Sigma + \varepsilon D \mid \Sigma) = \frac{\varepsilon^2}{2}\langle Q_\Sigma(D), D\rangle_F + o(\varepsilon^2), \qquad Q_\Sigma(D) = \frac{1}{2}\Sigma^{-1}D\Sigma^{-1},$$

where $\langle A,B\rangle_F = \mathrm{tr}(A^\top B)$ is the Frobenius pairing. Thus the local quadratic form is

$$g_\Sigma(D,E) = \frac{1}{2}\,\mathrm{tr}(\Sigma^{-1}D\Sigma^{-1}E).$$

With this normalization, the Fisher–Rao, or affine-invariant, geodesic distance on $\mathbb{S}^d_{++}$ is

$$d_{\mathrm{FR}}(\Sigma,\Sigma')^2 = \frac{1}{2}\|\log(\Sigma^{-1/2}\Sigma'\Sigma^{-1/2})\|_F^2,$$

where log denotes the principal matrix logarithm, and the corresponding geodesic is

$$\Sigma_t^{\mathrm{FR}} = \Sigma^{1/2}\left(\Sigma^{-1/2}\Sigma'\Sigma^{-1/2}\right)^t\Sigma^{1/2}.$$

The contrast with Bures is especially visible near the boundary of the covariance cone. For example, $\mathcal{B}^2(\Sigma,0) = \mathrm{tr}(\Sigma)$, and more generally the Bures formula extends continuously to $\mathbb{S}^d_+$; rank-deficient covariance matrices are therefore at finite distance and the closed cone is part of the metric completion.

Fisher–Rao is different. If an eigenvalue of $\Sigma^{-1/2}\Sigma'\Sigma^{-1/2}$ tends to zero, then the corresponding logarithm diverges in d_{FR}. Hence the boundary, made of degenerate covariance matrices and representing infinitely anisotropic Gaussian limits on fixed-trace sections, is at infinite distance.

In the coordinates (2.20), write $z = (t, u, v)$ and $z' = (t', u', v')$, and set

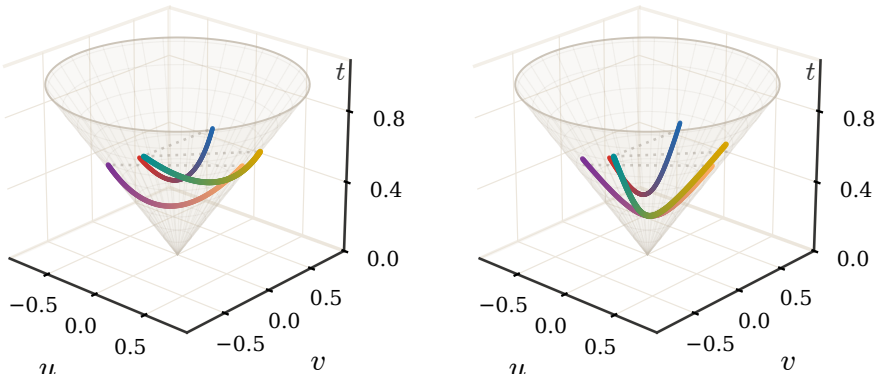
$$\Delta_z = t^2 - u^2 - v^2, \qquad \langle z, z' \rangle_{\mathrm{E}} = tt' + uu' + vv', \qquad \langle z, z' \rangle_{\mathrm{L}} = tt' - uu' - vv'.$$

Then the two distances admit the side-by-side expressions

$$\mathcal{B}(\Sigma, \Sigma')^2 = \sqrt{2}(t + t') - 2\sqrt{\langle z, z' \rangle_{\mathrm{E}} + \sqrt{\Delta_z \Delta_{z'}}}, \tag{2.22}$$

$$d_{\mathrm{FR}}(\Sigma, \Sigma')^2 = \frac{1}{2}\left((\log \lambda_+)^2 + (\log \lambda_-)^2\right), \qquad \lambda_\pm = \frac{\langle z, z' \rangle_{\mathrm{L}} \pm \sqrt{\langle z, z' \rangle_{\mathrm{L}}^2 - \Delta_z \Delta_{z'}}}{\Delta_z}, \tag{2.23}$$

where the Fisher–Rao formula requires $\Delta_z, \Delta_{z'} > 0$ and $\lambda_\pm$ are the eigenvalues of $\Sigma^{-1}\Sigma'$. Figure 2.17 shows this contrast in the 2×2 ice-cream cone: the Bures path reaches a rank-one covariance, while the Fisher–Rao path is drawn only to a small positive-definite regularization of the same rank-one limit.



Bures–Wasserstein Fisher–Rao

Figure 2.17: Bures–Wasserstein and Fisher–Rao covariance geodesics in the 2×2 positive-semidefinite cone. Both panels use the coordinates $u = (a - b)/\sqrt{2}$, $v = \sqrt{2}c$, $t = (a + b)/\sqrt{2}$ for $\Sigma = \left(\begin{smallmatrix} a & c \\ c & b \end{smallmatrix}\right)$. The three trajectories use distinct red-to-blue, orange-to-violet, and teal-to-gold palettes, repeated identically in both panels to identify corresponding endpoints; faint gray chords show the ambient Euclidean segments. The Bures geodesics can reach rank-deficient covariances on the boundary, whereas the Fisher–Rao geodesics are drawn toward small positive-definite regularizations. The omitted boundary endpoints correspond to rank-one limiting covariances, which are not at finite Fisher–Rao distance.

CHAPTER 3

Kantorovich Relaxation

Kantorovich's relaxation is the decisive move that turns transport into convex optimization. This chapter replaces deterministic maps by couplings, first for discrete histograms and then for general measures; it develops the resulting linear-programming structure, existence theory and characterization of optimizers by cyclical monotonicity. The metric geometry induced when the cost is a power of a ground distance is the subject of the next chapter. Historically, this viewpoint grew from Kantorovich's economic planning work [294] and is now the standard foundation of OT [538, 539, 446].

3.1 Discrete Relaxation

The discrete relaxation is the cleanest place to see mass splitting. It replaces permutations by a transportation polytope and reveals the linear-programming structure that algorithms exploit.

Mass splitting and couplings. Monge's discrete matching problem is problematic because it cannot be applied when $n \neq m$. A faithful formulation must keep track of masses $(\mathrm{a}_i, \mathrm{b}_j)$. The continuous Monge problem (2.5), based on push-forward maps, has the same obstruction in another form: there may be no map T such that $T_\sharp \alpha = \beta$. This happens, for instance, when a single Dirac mass should be sent to several Dirac masses.

This lack of mass splitting also makes the Monge formulation asymmetric in α and β: one can map two Dirac masses to a single one, but not the other way around. Even when a feasible map exists, the resulting optimization problem is non-convex and therefore difficult to solve numerically.

Kantorovich's key idea [294] is to relax the deterministic nature of transportation. Instead of requiring each source point x_i to be sent to exactly one target, the mass at x_i may be dispatched across several locations. This moves from deterministic transport maps to probabilistic, or fuzzy, transport plans. The relaxation is encoded, in place of a permutation σ or a map T, by a coupling matrix $\mathrm{P} \in \mathbb{R}_+^{n \times m}$, where $\mathrm{P}_{i,j}$ describes the amount of mass flowing from x_i to y_j in the formalism of discrete measures

$$\alpha = \sum_i \mathrm{a}_i \delta_{x_i}, \qquad \beta = \sum_j \mathrm{b}_j \delta_{y_j}.$$

Definition 3.1 (Discrete couplings and mass conservation)**.** Admissible couplings are only constrained to satisfy conservation of mass:

$$\mathrm{U}(\mathrm{a}, \mathrm{b}) \coloneqq \left\{ \mathrm{P} \in \mathbb{R}_+^{n \times m} \;;\; \mathrm{P}\mathbb{1}_m = \mathrm{a} \quad \text{and} \quad \mathrm{P}^\top \mathbb{1}_n = \mathrm{b} \right\}. \tag{3.1}$$

Equivalently,

$$\mathrm{P}\mathbb{1}_m = \Big(\sum_j \mathrm{P}_{i,j}\Big)_i \in \mathbb{R}^n, \qquad \mathrm{P}^\top \mathbb{1}_n = \Big(\sum_i \mathrm{P}_{i,j}\Big)_j \in \mathbb{R}^m.$$

Remark 3.2 (Small transportation polytopes). The definition is already informative in the smallest dimensions. If $n = m = 1$, mass conservation fixes the only entry, so the feasible set is a singleton. The same happens for $(n, m) = (2, 1)$, and by symmetry for $(n, m) = (1, 2)$: the unique coupling is forced by its only column, or by its only row.

The first nontrivial case is $(n, m) = (2, 2)$. Let $\mathrm{a} = (p, 1-p)$ and $\mathrm{b} = (q, 1-q)$ with $p, q \in [0, 1]$. Once $s \coloneqq \mathrm{P}_{1,1}$ is chosen, the marginal constraints force all other entries, hence every coupling has the form

$$\mathrm{P}(s) = \begin{pmatrix} s & p - s \\ q - s & 1 - p - q + s \end{pmatrix}.$$

The nonnegativity constraints are exactly $s \in [\max(0, p + q - 1), \min(p, q)]$, so $\mathrm{U}(\mathrm{a}, \mathrm{b})$ is a segment, possibly reduced to a point at the boundary. In general, when all marginal entries are positive, the transportation polytope has affine dimension $(n-1)(m-1)$ before the nonnegativity inequalities cut

out its faces.

The first useful consequence of this relaxation is feasibility: there is always at least one admissible plan, obtained by making source and target independent.

Definition 3.3 (Discrete product coupling)**.** Given weights $a \in \Sigma_n$ and $b \in \Sigma_m$, the discrete product, or trivial, coupling is

$$(a \otimes b)_{i,j} := a_i b_j.$$

It belongs to $U(a, b)$ and corresponds to choosing the source and target labels independently.

This feasible set is the bounded intersection of an affine space with the nonnegative orthant, hence a convex polytope. In one dimension, an ordered coupling can be read as a matrix: rows index source bins, columns index target bins, and the marginal constraints appear as prescribed row and column sums.

Proposition 3.4 (Discrete product optimality is degenerate)**.** *Assume that the zero-mass rows and columns have been removed, so that* $a_i > 0$ *and* $b_j > 0$*, and let* C *be a finite cost matrix. The product plan* $a \otimes b$ *minimizes* $P \mapsto \langle C, P\rangle$ *over* $U(a, b)$ *if and only if every coupling* $P \in U(a, b)$ *minimizes it.*

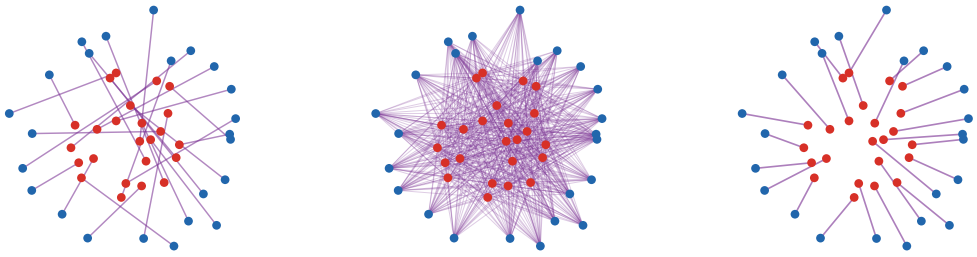
Proof. The reverse implication is immediate. Assume conversely that $a \otimes b$ is optimal and let $Q \in U(a, b)$ be arbitrary. Since all entries of $a \otimes b$ are positive, there exists $t > 0$ small enough that $R := (1+t)(a \otimes b) - tQ$ has nonnegative entries. Its row and column sums are $R\mathbb{1}_m = (1+t)a - ta = a$ and $R^\top \mathbb{1}_n = (1+t)b - tb = b$, so $R \in U(a, b)$. Moreover, $a \otimes b = \frac{1}{1+t} R + \frac{t}{1+t} Q$. By optimality of $a \otimes b$, both $\langle C, R\rangle$ and $\langle C, Q\rangle$ are at least $\langle C, a \otimes b\rangle$. Taking the scalar product of the convex-combination identity with C gives

$$\langle C, a \otimes b\rangle = \frac{1}{1+t}\langle C, R\rangle + \frac{t}{1+t}\langle C, Q\rangle.$$

A convex average of two numbers not smaller than $\langle C, a \otimes b\rangle$ can equal $\langle C, a \otimes b\rangle$ only if both numbers are equal to it. Hence $\langle C, Q\rangle = \langle C, a \otimes b\rangle$, and Q is optimal. Since Q was arbitrary, all couplings are optimal. □

Thus the product plan is mainly a feasibility witness. Except in the degenerate situation where the linear cost is constant on the whole transportation polytope, it is not expected to solve optimal transport. It is also maximally diffuse: when all masses are positive, it has nm positive entries, whereas Proposition 3.8 shows below that sparse optimal plans with at most $n + m - 1$ positive entries always exist.

Figure 3.1 contrasts deterministic, product and optimal couplings through weighted transport segments.



deterministic graph product plan $a \otimes b$ optimal plan

Figure 3.1: Discrete couplings represented as straight transport segments on the canonical point clouds used in the matching section. The deterministic graph is a feasible Monge-type plan, the product plan spreads every source mass over all targets, and the optimal Kantorovich plan minimizes the quadratic transport cost. Line width and opacity encode the transported mass.

Figure 3.2 gives the complementary matrix representation and attaches the prescribed marginals directly to each plan.

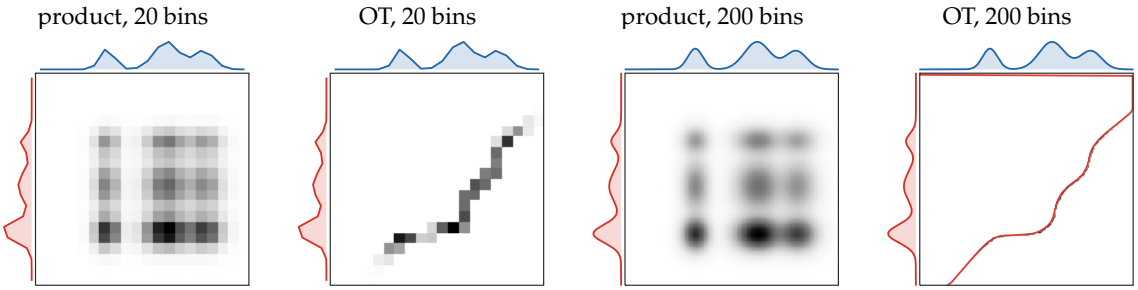


Figure 3.2: Coupling matrices with their prescribed marginals. The central grayscale image displays $P_{i,j}$; the red curve on the left is the source marginal a, and the blue curve on top is the target marginal b. The independent product plan is diffuse, whereas the one-dimensional optimal plan concentrates near the monotone quantile correspondence. Only the dense 200-bin optimal panel overlays the barycentric projection $i \mapsto \sum_j P_{i,j} j / a_i$, because that curve is meaningful visually only at sufficient resolution.

Whereas the Monge formulation is intrinsically asymmetric, Kantorovich's relaxed formulation is symmetric at the level of feasible sets: a coupling P belongs to $U(a, b)$ if and only if $P^\top$ belongs to $U(b, a)$.

Linear-programming structure. Kantorovich, aiming for economic planning, made a strong simplifying assumption: the cost of transportation should be linear in the amount of transported mass.

Definition 3.5 (Discrete Kantorovich problem)**.** For $a \in \Sigma_n$, $b \in \Sigma_m$, and $C \in \mathbb{R}^{n \times m}$, define

$$L_C(a, b) := \min_{P \in U(a,b)} \langle C, P \rangle = \min_{P \in U(a,b)} \sum_{i,j} C_{i,j} P_{i,j}. \tag{3.2}$$

A minimizer P is an optimal coupling, or optimal transport plan.

Here $C_{i,j}$ is the cost of moving one unit of mass from x_i to y_j. This is a linear program, and its solutions need not be unique.

The discrete transport value has opposite curvature in its two types of data: mixing marginals gives convexity, whereas minimizing a linear objective over couplings gives concavity in the cost matrix.

Proposition 3.6 (Convexity in the marginals and concavity in the cost)**.** *For every fixed cost matrix* $C \in \mathbb{R}^{n \times m}$*, the map*

$$(a, b) \longmapsto L_C(a, b)$$

is jointly convex on $\Sigma_n \times \Sigma_m$*. For every fixed* $(a, b) \in \Sigma_n \times \Sigma_m$*, the map*

$$C \longmapsto L_C(a, b)$$

is concave on $\mathbb{R}^{n \times m}$*.*

Proof. For $r \in \{0, 1\}$, let P_r be optimal between a_r and b_r. For $t \in [0, 1]$, the matrix $P_t = (1 - t)P_0 + tP_1$ belongs to $U((1 - t)a_0 + ta_1, (1 - t)b_0 + tb_1)$. Therefore

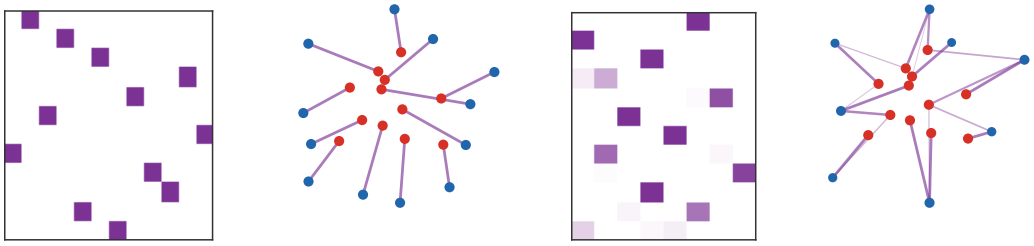
$$L_C\big((1 - t)a_0 + ta_1, (1 - t)b_0 + tb_1\big) \leqslant \langle C, P_t \rangle = (1 - t)L_C(a_0, b_0) + tL_C(a_1, b_1).$$

For fixed marginals, every $P \in U(a, b)$ satisfies

$$\langle (1 - t)C_0 + tC_1, P \rangle \geqslant (1 - t)L_{C_0}(a, b) + tL_{C_1}(a, b).$$

Minimizing the left-hand side over P proves concavity in C. □

Figure 3.3 contrasts the permutation plans inherited from equal-weight assignments with the genuinely splitting couplings permitted by nonuniform marginals.



permutation matrix P_σ assignment segments splitting matrix P splitting segments

Figure 3.3: From permutation matrices to splitting couplings. When the two empirical measures have the same number of atoms and uniform weights, an optimal plan can be a permutation matrix. Once target masses are nonuniform, the admissible set $U(a, b)$ also contains plans in which one source sends mass to several targets and several sources merge into the same target.

Sparsity here is not peculiar to transport. For nonnegative variables, the relevant quantity is the rank of the constraint operator—not the raw number of listed constraints, which may contain redundancies. The following standard linear-programming principle makes this precise; see, for instance, [52].

Proposition 3.7 (Rank-controlled sparse minimizers)**.** *Let $\mathcal{A} : \mathbb{R}^{n\times m} \to \mathbb{R}^q$ be linear, let $\mathbf{d} \in \mathbb{R}^q$ and $C \in \mathbb{R}^{n\times m}$, and consider*

$$\inf_{P\geqslant 0,\ \mathcal{A}(P)\leqslant \mathbf{d}} \langle C, P\rangle, \tag{3.3}$$

where the inequality is componentwise. If the feasible set is nonempty and the objective is bounded below on it, then (3.3) admits a minimizer $P^\star$ satisfying

$$\#\left\{(i,j)\ ;\ P^\star_{i,j} > 0\right\} \leqslant \operatorname{rank}(\mathcal{A}). \tag{3.4}$$

Here $\operatorname{rank}(\mathcal{A}) = \dim(\operatorname{Im}\mathcal{A})$. The same conclusion holds when $\mathcal{A}(P) \leqslant \mathbf{d}$ is replaced by $\mathcal{A}(P) = \mathbf{d}$.

Proof. The feasible region is a nonempty polyhedron, and the objective is bounded below on it. The attainment theorem for linear programs therefore provides a minimizer. Among all minimizers, choose $P^\star$ with the smallest number of positive entries, and denote its support by S.

Suppose that $\#S > \operatorname{rank}(\mathcal{A})$. The restriction of $\mathcal{A}$ to the $\#S$-dimensional space of matrices supported on S then has a nontrivial kernel. Hence there is a nonzero matrix H, supported on S, such that $\mathcal{A}(H) = 0$. Since $P^\star$ is strictly positive on S, both $P^\star + tH$ and $P^\star - tH$ remain nonnegative for all sufficiently small $t > 0$. Moreover, $\mathcal{A}(P^\star \pm tH) = \mathcal{A}(P^\star)$, so both perturbations satisfy exactly the same linear inequalities, or equalities, as $P^\star$. Optimality forces $\langle C, H\rangle = 0$; otherwise one of the two signs would strictly decrease the objective.

Choose $\sigma \in \{-1, 1\}$ such that σH has a negative entry, and set $t_\star := \min_{(i,j):\,\sigma H_{i,j}<0} P^\star_{i,j}/(-\sigma H_{i,j}) > 0$. Then $P^\star + t_\star \sigma H$ is nonnegative, feasible, and has the same cost as $P^\star$. At least one of its positive entries has become zero, while no new entry has appeared outside S. This contradicts the minimality of S and proves (3.4). □

Proposition 3.8 (Sparse optimal plans)**.** *Assume $a_i > 0$, $b_j > 0$ and $\sum_i a_i = \sum_j b_j = 1$. The linear program (3.2) admits an optimal coupling with at most $n + m - 1$ nonzero entries.*

Proof. The transportation polytope is nonempty and compact, so the minimum is finite. Apply the equality case of Proposition 3.7 to the marginal operator $\mathcal{A}_{\mathrm{marg}}(P) := (P\mathbb{1}_m, P^\top \mathbb{1}_n) \in \mathbb{R}^{n+m}$. Its rank is $n + m - 1$. Indeed, every matrix satisfies $\sum_i (P\mathbb{1}_m)_i = \sum_j (P^\top \mathbb{1}_n)_j$, so there is one linear dependence among the $n + m$ marginal coordinates. It is the only one: if $u \in \mathbb{R}^n$ and $v \in \mathbb{R}^m$ satisfy

$$\langle u, P\mathbb{1}_m\rangle + \langle v, P^\top \mathbb{1}_n\rangle = 0 \qquad \text{for every } P,$$

then testing this identity on each elementary matrix gives $u_i + v_j = 0$ for every (i, j). Thus all u_i equal one scalar and all v_j equal its opposite, so the annihilator of $\operatorname{Im}(\mathcal{A}_{\mathrm{marg}})$ is one-dimensional. It follows that $\operatorname{rank}(\mathcal{A}_{\mathrm{marg}}) = n + m - 1$, and Proposition 3.7 supplies the claimed optimal coupling. □

For transportation polytopes, this rank argument also has a graph interpretation. A cycle in the bipartite support produces an alternating nonzero perturbation with zero row and column sums, exactly the kernel direction used in the general proof. Consequently, the positive support of an extreme coupling is a forest.

Proposition 3.9 (North-west corner feasible plan)**.** *Let* $a \in \mathbb{R}^n_+$ *and* $b \in \mathbb{R}^m_+$ *have the same positive total mass. The following greedy sweep constructs a coupling* $P \in U(a, b)$ *with at most* $n + m - 1$ *positive entries and an acyclic positive support. Starting from* $(i, j) = (1, 1)$ *with residual masses* $r_i = a_i$ *and* $s_j = b_j$*, skip zero residuals, set* $P_{i,j} = \min(r_i, s_j)$*, subtract this value from both residuals, and advance every index whose residual has become zero. Repeat until all residual masses are exhausted.*

Proof. All assignments are nonnegative. At each step, the mass placed in entry (i, j) is subtracted from exactly one current row residual and one current column residual, so no row or column can receive more mass than prescribed. Conversely, an index is advanced only when its residual has been fully filled. When the algorithm stops, the total assigned mass is $\sum_i a_i = \sum_j b_j$, hence all row and column sums are exactly a and b.

Each positive assignment exhausts at least one current row or one current column. Before the final assignment, at most $n - 1$ row advances and $m - 1$ column advances can occur without terminating the construction. Hence the number of positive entries is at most $(n - 1) + (m - 1) + 1 = n + m - 1$. For acyclicity, view the positive support as a bipartite graph. Once a row or column index is advanced, it never appears again, so each new positive edge either starts a new component or attaches at least one new vertex to the component currently being swept. No edge is ever added between two old vertices of the same component, so no cycle can be created. □

Algorithm 3.1: North-west corner coupling

Input: Source weights $a \in \Sigma_n$ and target weights $b \in \Sigma_m$.
Output: Sparse feasible coupling $P \in U(a, b)$.
Initialize: Set $P = 0$, $r = a$, $s = b$, and $(i, j) = (1, 1)$.
While $i \leqslant n$ and $j \leqslant m$ **do**:
  $\eta = \min(r_i, s_j)$, $P_{i,j} \leftarrow \eta$.
  Update residuals: $r_i \leftarrow r_i - \eta$, $s_j \leftarrow s_j - \eta$.
  If $r_i = 0$ **then**:
    Set $i \leftarrow i + 1$.
  If $s_j = 0$ **then**:
    Set $j \leftarrow j + 1$.
Return P.

The north-west corner rule, summarized in Algorithm 3.1, does not use the cost matrix and is therefore not meant to solve (3.2). Its role is algorithmic: an acyclic support corresponds to linearly independent marginal constraints. When the support has fewer than $n + m - 1$ positive entries, transportation simplex implementations complete it with zero-mass basic variables to obtain a degenerate basic feasible solution. This gives a cheap initialization for the pivoting methods discussed in Section 3.2.

One-dimensional cases. In one dimension, the transportation polytope has a canonical monotone optimizer. This is the weighted version of the sorting rule from Section 1.1.

Proposition 3.10 (One-dimensional weighted sweep)**.** *Let* $x_1 \leqslant \cdots \leqslant x_n$ *and* $y_1 \leqslant \cdots \leqslant y_m$ *be points on the line, and let* $c(x, y) = h(x - y)$ *with* h *convex. The north-west corner plan between the sorted weighted atoms is optimal for* (3.2)*. Consequently, for unsorted one-dimensional inputs, an optimal plan is obtained in* $O(n \log n + m \log m)$ *time by sorting and then sweeping the masses once from left to right.*

Proof. The north-west plan is monotone: if $i < i'$ and $j > j'$, it cannot put positive mass on both (i, j) and (i', j'), because the sweep exhausts rows and columns in increasing order. To prove optimality, start from an arbitrary feasible plan P. If $P_{1,1} < \min(a_1, b_1)$, then row 1 sends positive mass to some column $j > 1$, while column 1 receives positive mass from some row $i > 1$. Moving $\eta = \min(P_{1,j}, P_{i,1})$ from $(1, j)$ and $(i, 1)$ to $(1, 1)$ and (i, j) preserves both marginals. Convexity of h gives the Monge inequality

$$h(x_1 - y_j) + h(x_i - y_1) \geqslant h(x_1 - y_1) + h(x_i - y_j),$$

so this exchange does not increase the cost. Each exchange either removes the chosen entry in row 1 or removes the chosen entry in column 1. Repeating finitely many times forces $P_{1,1} = \min(a_1, b_1)$, exactly the first north-west assignment. It exhausts row 1, column 1, or both. Removing every exhausted index reduces the problem to a smaller ordered transportation problem. Induction on $n + m$ produces the complete north-west plan without increasing cost. Sorting costs $O(n \log n + m \log m)$ and the sweep uses at most $n + m - 1$ positive assignments. □

Permutation matrices as couplings. We now restrict attention to the special case $n = m$ and uniform weights $a = b = \mathbb{1}_n/n$. In this case a matching can be encoded as a matrix with exactly one active entry per row and per column.

Definition 3.11 (Permutation matrices)**.** For a permutation $\sigma \in \mathrm{Perm}(n)$, its permutation matrix P_σ is

$$(P_\sigma)_{i,j} = \begin{cases} 1 & \text{if} \quad j = \sigma(i), \\ 0 & \text{otherwise.} \end{cases}$$

The set of all permutation matrices is

$$\mathcal{P}_n^{\mathrm{perm}} := \{P_\sigma \; ; \; \sigma \in \mathrm{Perm}(n)\}\,.$$

The corresponding probability coupling is P_σ/n. If the matching cost matrix is C, then

$$\langle C, P_\sigma/n \rangle = \frac{1}{n} \sum_{i=1}^{n} C_{i,\sigma(i)}.$$

Thus the assignment problem is the minimization of a linear function over the discrete, non-convex set of permutation matrices.

The convex relaxation replaces this finite set by all bistochastic matrices.

Definition 3.12 (Birkhoff polytope)**.** The Birkhoff polytope is the convex set of bistochastic matrices

$$\mathcal{B}_n := \left\{P \in \mathbb{R}_+^{n\times n} \; ; \; P\mathbb{1}_n = \mathbb{1}_n \quad \text{and} \quad P^\top \mathbb{1}_n = \mathbb{1}_n\right\}.$$

Then $U(\mathbb{1}_n/n, \mathbb{1}_n/n) = \mathcal{B}_n/n$, and permutation couplings are included in this convex relaxation. More precisely,

$$\mathcal{P}_n^{\mathrm{perm}} = \mathcal{B}_n \cap \{0,1\}^{n\times n} \subset \mathcal{B}_n,$$

so before using the structure of $\mathcal{B}_n$ one first obtains only the relaxed inequality

$$\min_{P\in\mathcal{B}_n} \langle C, P\rangle \leqslant \min_{P\in\mathcal{P}_n^{\mathrm{perm}}} \langle C, P\rangle.$$

The next elementary facts lead to the Birkhoff–von Neumann theorem [61, 542], which explains why this relaxation is tight for uniform matching.

Definition 3.13 (Extreme points)**.** For a convex set C in a finite-dimensional vector space,

$$\mathrm{Extr}(C) := \{x \in C \; ; \; x = (y+z)/2, \; y, z \in C \Rightarrow y = z = x\}\,.$$

Proposition 3.14 (Existence of extreme points)**.** *If C is a non-empty compact convex subset of a finite-dimensional vector space, then* $\mathrm{Extr}(C)$ *is non-empty.*

Proof. Among all non-empty faces of C, choose one of minimal affine dimension. If this face contained two distinct points, maximizing a linear functional that is not constant on the face would produce a non-empty proper exposed subface, contradicting minimality. Hence the minimal face is a singleton, and its point is extreme. □

Example 3.15 (An unbounded convex set without extreme points). Compactness cannot be dropped from Proposition 3.14. For instance, the affine line $C = \mathbb{R} \times \{0\} \subset \mathbb{R}^2$ is closed, convex and unbounded, but it has no extreme point: every $(x, 0)$ is the midpoint of the distinct points $(x-1, 0)$ and $(x+1, 0)$.

Proposition 3.16 (Linear programs have extreme minimizers)**.** *Let C be a non-empty compact convex set. For every linear form ℓ,*

$$\operatorname{Extr}(C) \cap \operatorname*{argmin}_{x \in C} \ell(x) \neq \emptyset.$$

Proof. The set $S = \operatorname{argmin}_{x \in C} \ell(x)$ is non-empty, compact and convex. By Proposition 3.14, it has an extreme point x. If $x = (y + z)/2$ with $y, z \in C$, then by linearity and optimality of x, both y and z also minimize ℓ on C, hence $y, z \in S$. Since x is extreme in S, $y = z = x$. Thus x is extreme in C. □

Theorem 3.17 (Birkhoff–von Neumann)**.** *The extreme points of $\mathcal{B}_n$ are exactly the permutation matrices.*

Figure 3.4 shows the non-extreme mechanism used in the proof below. The displayed matrix is bistochastic but not a permutation matrix: the three unit entries already behave like isolated matching edges, while the scattered fractional support contains a longer alternating cycle.

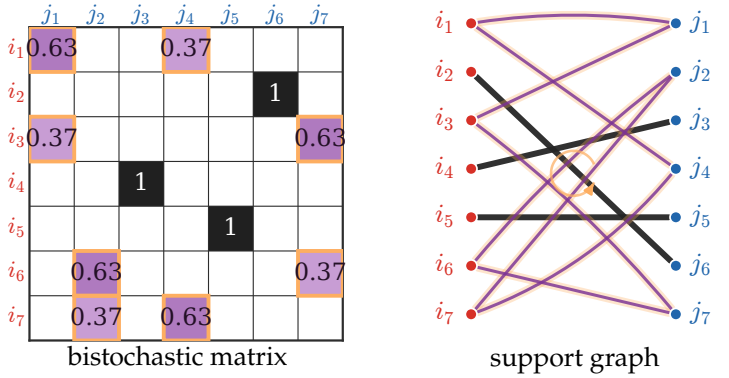


Figure 3.4: Cycle certificate in the Birkhoff–von Neumann proof. The left panel is a 7×7 bistochastic matrix which is not a permutation matrix. It has only three isolated entries equal to one, while the non-binary entries are scattered across four rows and columns. The right panel shows the bipartite positive-support graph, with the column nodes sorted as $j_1, \ldots, j_7$ from top to bottom to match the matrix order: red nodes are rows, blue nodes are columns, thin purple edges correspond to $0 < \mathrm{P}_{i,j} < 1$, and bold black edges correspond to isolated entries $\mathrm{P}_{i,j} = 1$. The orange halo marks the longer alternating fractional cycle along which one can add and subtract mass while preserving all row and column sums.

Proof. We first prove that permutation matrices are extreme. Let $\mathrm{P}_\sigma \in \mathcal{P}_n^{\mathrm{perm}}$ and assume that $\mathrm{P}_\sigma = (\mathrm{Q} + \mathrm{R})/2$ with $\mathrm{Q}, \mathrm{R} \in \mathcal{B}_n$. Every bistochastic matrix has entries in $[0, 1]$. Since the only extreme points of $[0, 1]$ are 0 and 1, each entry of P_σ fixes the corresponding entries of Q and R: if $(\mathrm{P}_\sigma)_{i,j} = 0$, then $\mathrm{Q}_{i,j} = \mathrm{R}_{i,j} = 0$, while if $(\mathrm{P}_\sigma)_{i,j} = 1$, then $\mathrm{Q}_{i,j} = \mathrm{R}_{i,j} = 1$. Hence $\mathrm{Q} = \mathrm{R} = \mathrm{P}_\sigma$, so P_σ is extreme.

We now prove the converse by contrapositive. Pick $\mathrm{P} \in \mathcal{B}_n \setminus \mathcal{P}_n^{\mathrm{perm}}$. Since an integral bistochastic matrix is necessarily a permutation matrix, P has at least one fractional entry. We shall split $\mathrm{P} = \frac{\mathrm{Q}+\mathrm{R}}{2}$ with $\mathrm{Q}, \mathrm{R} \in \mathcal{B}_n$ and $\mathrm{Q} \neq \mathrm{R}$, proving that P is not extreme.

Associate with P the bipartite graph whose left vertices are the rows, whose right vertices are the columns, and whose edges are the fractional entries $0 < \mathrm{P}_{i,j} < 1$. An entry equal to 1 uses the whole mass of its row and column, so it is isolated in the positive support and does not appear in this fractional graph. If a left vertex i is incident to a fractional edge (i, j_1), then it must be incident to at least one other fractional edge. Indeed, the row sum is one; after the contribution $\mathrm{P}_{i,j_1} \in (0, 1)$, a positive amount $1 - \mathrm{P}_{i,j_1}$ remains in the same row, and it cannot be carried by an entry equal to 1. The same argument applies to right vertices, using the column sums. Thus every non-isolated vertex of the fractional graph has degree at least two.

Starting from any fractional edge, one may therefore walk through adjacent fractional edges without immediately backtracking and without getting stuck. Since the graph is finite, some vertex is eventually visited twice; the portion of the walk between the two visits contains a cycle. Choose a shortest such cycle and write it in alternating form $(i_1, j_1, i_2, j_2, \ldots, i_p, j_p)$, with $i_{p+1} = i_1$, where both (i_s, j_s) and (i_{s+1}, j_s) are fractional for every s. The minimality of the cycle implies that the vertices i_s are all distinct and that the vertices j_s are all distinct. In particular,

$$0 < \mathrm{P}_{i_s,j_s} < 1 \quad \text{and} \quad 0 < \mathrm{P}_{i_{s+1},j_s} < 1.$$

Define

$$\varepsilon := \min_{1 \leqslant s \leqslant p} \left\{ \mathrm{P}_{i_s,j_s},\ \mathrm{P}_{i_{s+1},j_s},\ 1 - \mathrm{P}_{i_s,j_s},\ 1 - \mathrm{P}_{i_{s+1},j_s} \right\}.$$

All these numbers are positive, so $\varepsilon > 0$. Split the cycle edges into the two alternating families

$$A := \{(i_s, j_s)\}_{s=1}^{p}, \qquad B := \{(i_{s+1}, j_s)\}_{s=1}^{p}.$$

We now perform the standard alternating-cycle perturbation:

$$\mathrm{Q}_{i,j} := \begin{cases} \mathrm{P}_{i,j}, & (i,j) \notin A \cup B, \\ \mathrm{P}_{i,j} + \varepsilon/2, & (i,j) \in A, \\ \mathrm{P}_{i,j} - \varepsilon/2, & (i,j) \in B, \end{cases} \qquad \mathrm{R}_{i,j} := \begin{cases} \mathrm{P}_{i,j}, & (i,j) \notin A \cup B, \\ \mathrm{P}_{i,j} - \varepsilon/2, & (i,j) \in A, \\ \mathrm{P}_{i,j} + \varepsilon/2, & (i,j) \in B. \end{cases}$$

By the definition of ε, all modified entries stay in $[0,1]$, so Q and R are nonnegative. Each row vertex i_s of the cycle is incident to exactly one edge of A and one edge of B; the $+\varepsilon/2$ and $-\varepsilon/2$ perturbations therefore cancel in that row. The same cancellation holds in each column vertex j_s, and all other rows and columns are unchanged. Consequently, $\mathrm{Q}\mathbb{1}_n = \mathrm{R}\mathbb{1}_n = \mathbb{1}_n$ and $\mathrm{Q}^\top \mathbb{1}_n = \mathrm{R}^\top \mathbb{1}_n = \mathbb{1}_n$, so $\mathrm{Q}, \mathrm{R} \in \mathcal{B}_n$. Finally, $\mathrm{Q} \neq \mathrm{R}$ because $\varepsilon > 0$ and the cycle is non-empty, while by construction $\mathrm{P} = (\mathrm{Q} + \mathrm{R})/2$. Thus P is not extreme. Consequently every extreme point of $\mathcal{B}_n$ is integral, and every integral bistochastic matrix is a permutation matrix. □

Corollary 3.18 (Kantorovich for matching)**.** *If $m = n$ and $\mathrm{a} = \mathrm{b} = \mathbb{1}_n/n$, then the discrete Kantorovich problem (3.2) admits an optimal solution of the form P_σ/n. The associated permutation σ solves the assignment problem of Section 1.1.*

Proof. The feasible set is $\mathcal{B}_n/n$. By Proposition 3.16, the linear objective has an optimal extreme point. Since scaling preserves extreme points and Theorem 3.17 identifies the extreme points of $\mathcal{B}_n$, this optimizer is P_σ/n for some permutation σ. Its cost is exactly $n^{-1} \sum_i \mathrm{C}_{i,\sigma(i)}$, so σ is an optimal assignment. □

Equivalently, for uniform empirical measures, one can always choose a permutation matrix among the minimizers of the relaxed Kantorovich problem: the relaxation is tight for assignment problems.

> **Remark 3.19 (General discrete case).** For general input measures, one does not have equivalence between Monge and Kantorovich problems, since the Monge constraint can be empty. In finite dimension, however, the support of an optimal coupling still enjoys strong sparsity: one can choose an optimal basic feasible plan whose bipartite support is cycle-free, hence with at most $n + m - 1$ nonzero entries. Figure 3.3 illustrates the difference between the tight uniform matching case and the genuinely splitting nonuniform case.

Constructive Birkhoff–von Neumann decomposition. The same combinatorial idea gives both a finite convex decomposition of every bistochastic matrix and an effective procedure for extracting its permutation matrices. The existence statement follows immediately from the extreme-point theorem above and Carathéodory's theorem; the algorithmic statement makes the matching argument constructive.

Corollary 3.20 (Birkhoff–von Neumann decomposition)**.** *Every $\mathrm{P} \in \mathcal{B}_n$ admits a representation*

$$\mathrm{P} = \sum_{r=1}^{L} \lambda_r \mathrm{P}_{\sigma_r}, \qquad \lambda_r > 0, \qquad \sum_{r=1}^{L} \lambda_r = 1, \qquad L \leqslant (n-1)^2 + 1,$$

where each P_{σ_r} is a permutation matrix.

Proof. Let $\mathcal{K}$ be the convex hull of the permutation matrices. Clearly $\mathcal{K} \subset \mathcal{B}_n$. Conversely, if some $\mathrm{P} \in \mathcal{B}_n$ did not belong to the compact convex set $\mathcal{K}$, strict separation would provide a linear form ℓ such that $\ell(\mathrm{P}) < \min_{\mathrm{Q} \in \mathcal{K}} \ell(\mathrm{Q})$. Proposition 3.16 and Theorem 3.17 give

$$\min_{\mathrm{Q} \in \mathcal{B}_n} \ell(\mathrm{Q}) = \min_{\sigma \in \mathrm{Perm}(n)} \ell(\mathrm{P}_\sigma) = \min_{\mathrm{Q} \in \mathcal{K}} \ell(\mathrm{Q}),$$

contradicting $\mathrm{P} \in \mathcal{B}_n$. Hence $\mathcal{B}_n = \mathcal{K}$.

The $2n$ row- and column-sum equations have rank $2n - 1$, and the strictly positive matrix $\mathbb{1}_n \mathbb{1}_n{}^\top/n$ lies in their affine space. Therefore $\mathcal{B}_n$ has affine dimension $n^2 - (2n-1) = (n-1)^2$. Carathéodory's theorem in this affine hull now reduces any convex decomposition to at most $(n-1)^2 + 1$ permutation matrices; zero coefficients can be discarded. □

The decomposition can be computed without first enumerating the extreme points. Hall's theorem states that a finite bipartite graph admits a matching covering all row vertices if and only if every row subset I has at least $|I|$ neighbors [265]. To apply it, let $\mathrm{R} \geqslant 0$ have every row and column sum equal to the same number $s > 0$, and let G_{R} be its support graph. If $N(I)$ denotes the neighbor set of I, then

$$s|I| = \sum_{i \in I} \sum_{j \in N(I)} \mathrm{R}_{i,j} \leqslant \sum_{j \in N(I)} \sum_{i} \mathrm{R}_{i,j} = s|N(I)|.$$

Thus $|N(I)| \geqslant |I|$, so G_{R} contains a perfect matching. The following algorithm extracts that matching by successive augmenting paths and then removes the largest possible multiple of its permutation matrix.

Algorithm 3.2: Birkhoff–von Neumann decomposition

Input: $\mathrm{P} \in \mathcal{B}_n$.
Initialize: $\mathrm{R} \leftarrow \mathrm{P}$, $s \leftarrow 1$, and an empty list $\mathcal{D}$.
While $s > 0$:
  Form the support graph G_{R} with edge (i, j) whenever $\mathrm{R}_{i,j} > 0$.
  Initialize matching: $M \leftarrow \varnothing$.
  For $q = 1, \dots, n$ **do**:
    Choose an unmatched row i and find by breadth-first search an M-augmenting path A from i to an unmatched column, alternating between edges outside and inside M.
    Set $M \leftarrow M \triangle A$.
  Set $\sigma(i) = j$ for every $(i, j) \in M$ and $\lambda \leftarrow \min_i \mathrm{R}_{i,\sigma(i)}$.
  Append (λ, σ) to $\mathcal{D}$, set $\mathrm{R} \leftarrow \mathrm{R} - \lambda \mathrm{P}_\sigma$, and set $s \leftarrow s - \lambda$.

Return $\mathcal{D}$, so that $\mathrm{P} = \sum_{(\lambda,\sigma) \in \mathcal{D}} \lambda \mathrm{P}_\sigma$.

Hall's condition proves that each augmenting-path search succeeds. Indeed, if a search rooted at an unmatched row were to fail, its alternating tree would reach a row set I and its full neighbor set $N(I)$. Every reached column would be matched to a distinct reached row other than the root, giving $|N(I)| = |I| - 1$ and contradicting Hall's condition. This augmenting-path argument is the standard computational proof of Hall's theorem; bistochasticity supplies its neighborhood hypothesis here.

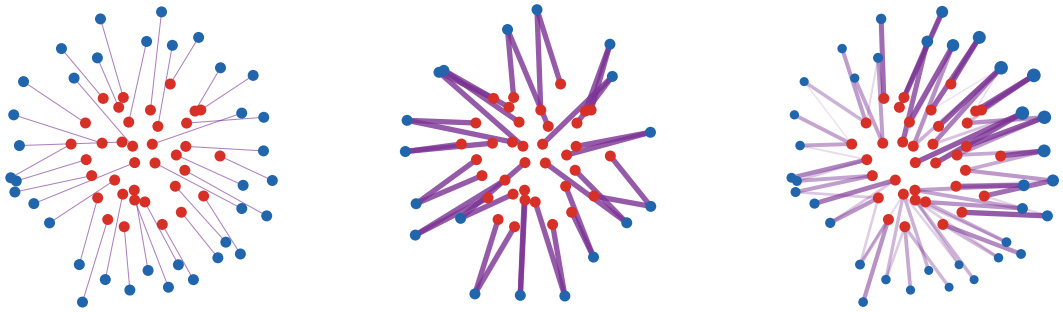
The residual update preserves nonnegativity and changes every row and column sum from s to $s - \lambda$. To verify both termination and the sharper term bound, normalize the current residual as $\widehat{\mathrm{P}} = \mathrm{R}/s \in \mathcal{B}_n$ and let F be its minimal face. The chosen permutation matrix belongs to F because its support is contained in that of $\widehat{\mathrm{P}}$. If $\lambda < s$, then

$$\widehat{\mathrm{P}}' := \frac{\mathrm{R} - \lambda \mathrm{P}_\sigma}{s - \lambda}$$

is the next normalized residual. Maximality of $\lambda = \min_i \mathrm{R}_{i,\sigma(i)}$ makes at least one formerly positive coordinate vanish, so $\widehat{\mathrm{P}}'$ lies in a proper face of F. Each nonterminal iteration therefore lowers the face dimension by at least one; if $\lambda = s$, the residual vanishes. Since $\dim(\mathcal{B}_n) = (n-1)^2$, Algorithm 3.2 terminates after at most $(n-1)^2 + 1$ extracted permutations. Summing its residual updates gives the asserted decomposition and $\sum_{(\lambda,\sigma) \in \mathcal{D}} \lambda = 1$.

If the current support graph has E edges, the elementary matching subroutine performs at most n breadth-first searches and costs $O(nE)$. Thus a dense implementation costs $O(n^3)$ per extracted permutation and, using the term bound above, at most $O(n^5)$ arithmetic and graph operations overall, with $O(n^2)$ memory. This is a conservative worst-case estimate rather than a bit-complexity bound; faster bipartite-matching routines improve it, and sparse residual supports reduce E substantially [316, 84].

Rational weights. The strict assignment model is tied to equal cardinalities and equal weights, whereas the coupling set $\mathrm{U}(\mathrm{a}, \mathrm{b})$ of Definition 3.1 accommodates arbitrary discrete masses and different support sizes. Figure 3.5 contrasts these regimes. For rational weights, the relaxed problem can in fact be reduced back to a larger uniform assignment problem.



$n = m$ $m < n$ $n = m$ but $b_j \neq 1/n$

Figure 3.5: From assignments to transport plans, using the same disk-to-annulus geometry as Figure 1.5. In the balanced equal-weight case, each source atom is matched to one target atom. With a target cloud that has half as many atoms, or with strongly nonuniform target weights, the coupling matrix can merge or split mass; segment thickness and opacity encode its nonzero entries, and blue marker areas encode the prescribed target masses.

Proposition 3.21 (Rational weights as duplicated uniform matching)**.** *Let*

$$\alpha = \sum_{i=1}^{n} \frac{k_i}{N} \delta_{x_i}, \qquad \beta = \sum_{j=1}^{m} \frac{\ell_j}{N} \delta_{y_j}, \qquad \sum_i k_i = \sum_j \ell_j = N,$$

with positive integers k_i, ℓ_j. Replace each x_i by k_i identical copies and each y_j by ℓ_j identical copies, producing two uniform N-point clouds. Then the duplicated assignment problem and the discrete Kantorovich problem between α and β have the same optimal value. Moreover, there exists an optimal coupling of the form

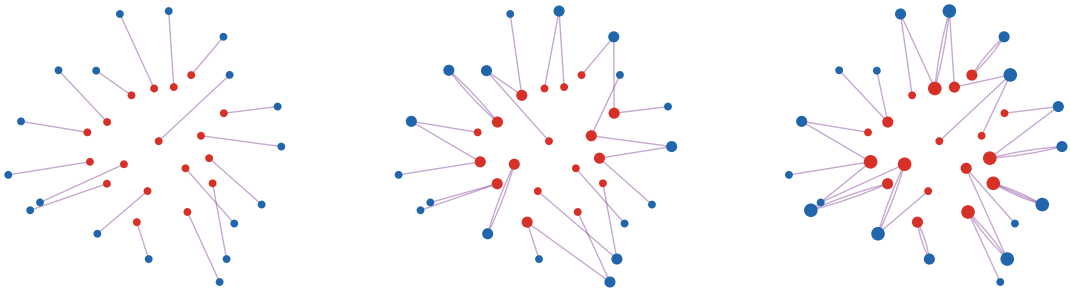
$$\mathrm{P}_{i,j} = \frac{n_{i,j}}{N}, \qquad n_{i,j} \in \mathbb{N}, \qquad \sum_j n_{i,j} = k_i, \quad \sum_i n_{i,j} = \ell_j,$$

and couplings whose scaled entries $N\mathrm{P}_{i,j}$ are integral are exactly the collapsed assignments between duplicated clouds. Fractional optimal couplings may nevertheless coexist when the optimum is degenerate.

Proof. Any assignment between the duplicated source and target clouds defines integers $n_{i,j}$ counting how many copied particles of type x_i are matched to copied particles of type y_j. These counts satisfy $\sum_j n_{i,j} = k_i$ and $\sum_i n_{i,j} = \ell_j$, and the associated coupling $\mathrm{P}_{i,j} = n_{i,j}/N$ has marginals k_i/N and ℓ_j/N. The assignment cost is $\frac{1}{N} \sum_{i,j} n_{i,j} c(x_i, y_j) = \sum_{i,j} \mathrm{P}_{i,j} c(x_i, y_j)$. Conversely, any nonnegative integer count matrix with those row and column sums can be realized by allocating the k_i copies of each x_i among the target copies according to $(n_{i,j})_j$. It remains to show that restricting to integral counts does not increase the optimum. Scale a feasible coupling by N and write $\mathrm{Q} = N\mathrm{P}$. The constraints on Q have integer right-hand sides $(k_1, \ldots, k_n, \ell_1, \ldots, \ell_m)$. After multiplying the target rows by -1, their coefficient matrix is the oriented node-edge incidence matrix of a bipartite graph and is therefore totally unimodular. Every vertex of this transportation polytope is integral. Since a linear objective attains its minimum at a vertex, an integral optimal Q exists, proving equality of the two optimal values. The argument establishes existence, not integrality of every optimizer; convex combinations of distinct integral optima give fractional optima. □

This network-flow integrality mechanism complements the Birkhoff–von Neumann theorem above (Theorem 3.17): in both cases, a linear transport relaxation has an optimizer represented by integer edge flows after scaling. The equal-unit-margin case specializes these flows to permutation matrices, whereas a degenerate optimal face may also contain fractional couplings.

Figure 3.6 makes this reduction explicit by replacing each rational mass with identical unit copies and then regrouping the resulting uniform assignment.



$k_i = \ell_j = 1$ $k_i, \ell_j \in \{1, 2\}$ $k_i, \ell_j \in \{1, 2, 3\}$

Figure 3.6: Rational weights as duplicated uniform matchings, using the same disk-to-annulus geometry as Figure 3.5 with fewer displayed atoms. The red and blue locations are kept fixed, while disk areas encode the integer multiplicities k_i and ℓ_j. Solving the assignment problem after duplicating particles produces several collapsed segments attached to high-multiplicity atoms; this is the integer count matrix of Proposition 3.21.

Applications of discrete transport. Discrete Kantorovich transport applies whenever weighted data points or histograms must be compared through a geometry-aware correspondence. The following examples illustrate information transfer between datasets, comparison of visual distributions, and inference of temporal relations between cell populations.

Example 3.22 (Application to domain adaptation). In unsupervised domain adaptation, labeled source samples $(x_i^s, y_i^s)_i$ and unlabeled target samples $(x_j^t)_j$ define empirical laws $\alpha_s = \sum_i a_i \delta_{x_i^s}$ and $\alpha_t = \sum_j b_j \delta_{x_j^t}$. Writing $\mathrm{a} = (a_i)_i$ and $\mathrm{b} = (b_j)_j$, a Kantorovich coupling $\mathrm{P} \in \mathrm{U}(\mathrm{a}, \mathrm{b})$ gives soft correspondences between the two clouds. For $b_j > 0$, disintegrating P with respect to its target marginal and averaging the source labels gives the backward barycentric rule $\widetilde{y}_j = b_j^{-1} \sum_i \mathrm{P}_{i,j} y_i^s$, for either real-valued labels or one-hot class vectors. This is the label-space analogue of the barycentric projection introduced later in Definition 12.31. Alternatively, the transport can be optimized jointly with a classifier by adding label-prediction terms to the feature cost. This is the mechanism behind OT domain adaptation and JDOT: the plan is not only a distance certificate, but an explicit cross-domain alignment [160, 159]. Learning or adapting the ground cost is the inverse viewpoint developed later in Section 12.5.

Example 3.23 (Application to visual distributions). Images, color histograms, texture descriptors and shape samples can all be represented as discrete measures. The ground cost then encodes color differences, image-plane displacements or surface distances. This viewpoint underlies the Earth Mover's Distance for image retrieval [457], regularized transport for imaging [552], convolutional transport on geometric domains [495], and Wasserstein barycenters for texture mixing or Radon-domain image processing [76, 75]. The practical gain is that the discrepancy respects geometry: moving a small amount of mass to a nearby color or pixel is cheaper than moving it far away.

Example 3.24 (Application to single-cell population dynamics). Single-cell sequencing observes different cells at each collection time. If $x_i^{[k]} \in \mathcal{X}$ denotes the state of cell i at time t_k, write

$$\alpha_{t_k} = \sum_{i=1}^{n_k} a_i^{[k]} \delta_{x_i^{[k]}} \quad \text{and} \quad \mathrm{a}^{[k]} = (a_i^{[k]})_{i=1}^{n_k} \in \Sigma_{n_k}.$$

Thus a cell is one atom of an empirical population measure, rather than itself a measure over genes as in Example 4.10. A balanced transport between two consecutive snapshots is represented by a matrix $\mathrm{P}^{[k]} \in \mathbb{R}_+^{n_k \times n_{k+1}}$ satisfying

$$\mathrm{P}^{[k]} \mathbb{1}_{n_{k+1}} = \mathrm{a}^{[k]} \quad \text{and} \quad (\mathrm{P}^{[k]})^\top \mathbb{1}_{n_k} = \mathrm{a}^{[k+1]}.$$

Its entry $\mathrm{P}_{i,j}^{[k]}$ is a soft ancestor–descendant mass, not evidence that the same cell was observed twice.

For every row with positive sum, the normalization

$$K^{[k]}_{i,j} := \frac{P^{[k]}_{i,j}}{\sum_{j'=1}^{n_{k+1}} P^{[k]}_{i,j'}}$$

defines a conditional transition probability; multiplying successive matrices $K^{[k]}$ yields population-level lineage probabilities across several collection times.
Waddington-OT modifies this balanced baseline by incorporating growth estimates and relaxing the marginal constraints through unbalanced transport, as discussed in Example 11.4 [476]. Related continuous and entropic constructions include trajectory inference and generative transport models [520, 326, 308]. Figure 3.7 displays three observed population snapshots in the common force-layout embedding used by Waddington-OT; this visualization need not coincide with the ground geometry used to compute $P^{[k]}$.

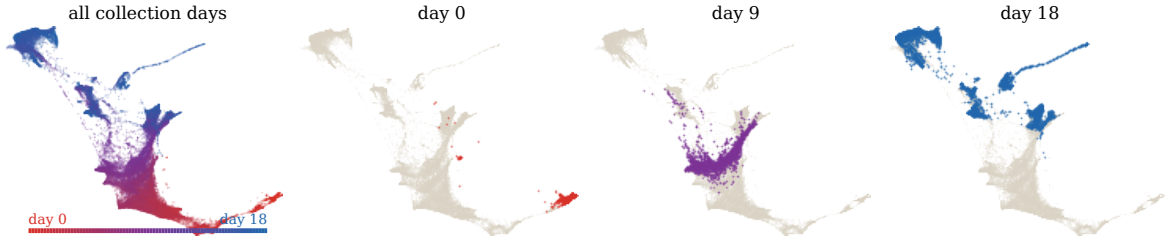


Figure 3.7: Three snapshots of the Waddington-OT single-cell reprogramming time course. The first panel aggregates a representative sample from all collection times and colors cells from red to blue according to time. The remaining panels highlight the populations observed at days 0, 9, and 18, while cells from other times remain in gray. Every panel uses the same official force-layout embedding and viewport; no transport coupling or interpolated trajectory is displayed.

3.2 Linear-Programming Algorithms

The discrete Kantorovich problem is a linear program with much more structure than a generic dense LP. Its variables are the arcs of a complete bipartite network, its equality constraints are flow-conservation constraints, and its extreme points are sparse tree-like couplings. This is why classical transportation algorithms remain important even though the formulation is finite-dimensional and convex.

Transportation simplex and network simplex. The transportation simplex goes back to Dantzig's formulation of the transportation problem [176]. It works on basic feasible couplings, whose positive support is completed into a spanning tree of the bipartite supply-demand graph. Starting from a basis, for instance one produced by the north-west corner rule, reduced costs identify whether an unused arc can decrease the objective. Adding such an arc creates a unique cycle in the tree; one then pushes as much mass as possible around this cycle and removes the exhausted arc. This is exactly the simplex method specialized to the transportation polytope, with pivots that can be implemented using graph operations instead of a generic basis inverse.

The network simplex is the corresponding pivoting method for general minimum-cost-flow problems [51]. It keeps node potentials, reduced costs and a spanning-tree basis, and it has become one of the most effective exact solvers for medium-scale discrete OT. Like the ordinary simplex method, its worst-case number of pivots can be exponential for adversarial instances, but the per-pivot operations exploit sparsity and are very efficient in practice. For theoretical polynomial guarantees, one can instead use strongly polynomial minimum-cost-flow algorithms, such as Orlin's algorithm [409]. In a dense balanced transportation problem with n sources and n targets, the graph has $O(n)$ vertices and $O(n^2)$ arcs, so these general bounds are polynomial but still much heavier than the nearly matrix-vector structure that Sinkhorn will exploit.

Interior-point methods. Generic interior-point methods approach the same LP through a smooth central path. Assume here that all entries of a and b are positive; zero-mass rows and columns must first

be removed. The logarithmic-barrier problem on the resulting transport polytope is

$$\mathrm{P}_\varepsilon := \operatorname*{argmin}_{\substack{\mathrm{P}\mathbb{1}_m=\mathrm{a},\ \mathrm{P}^\top\mathbb{1}_n=\mathrm{b}\\ \mathrm{P}_{i,j}>0}} \langle \mathrm{C}, \mathrm{P}\rangle - \varepsilon \sum_{i,j} \log \mathrm{P}_{i,j}, \tag{3.5}$$

where $\varepsilon > 0$ is decreased along the algorithm. The barrier is singular at the boundary, so each iterate stays strictly inside the transportation polytope; as $\varepsilon \downarrow 0$, the central path approaches the set of LP minimizers. Each Newton step solves a linear system involving the marginal constraints and the current diagonal Hessian $\mathrm{diag}(\varepsilon/\mathrm{P}_{i,j}^2)$, which gives robust polynomial complexity but can be expensive for dense couplings [399].

Figure 3.8 isolates this mechanism on a two-dimensional polytope: decreasing the barrier parameter moves the minimizer along the central path toward the optimal face.

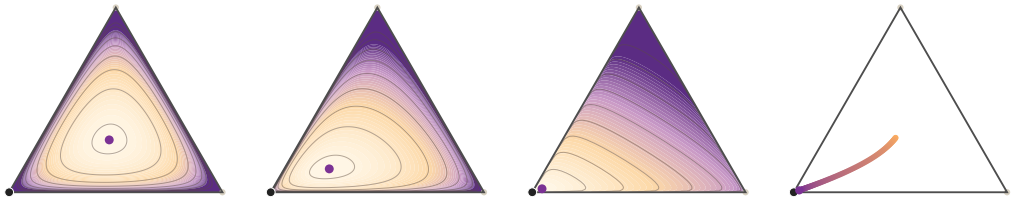


Figure 3.8: Logarithmic-barrier central path for a two-dimensional equilateral triangular slice of a linear program. The first three panels display level sets of $\ell^\top z - \varepsilon \sum_i \log(b_i - \langle a_i, z\rangle)$ for decreasing barrier strengths. The last panel removes the objective background and shows only the central path inside the feasible triangle. Large ε selects a central interior point; decreasing ε moves the minimizer toward the optimal vertex while never touching the boundary. This should be contrasted with entropic OT, where the entropy temperature ε is usually fixed and defines the regularized problem itself.

The comparison with Sinkhorn is therefore subtle. Both methods keep iterates positive, but they use positivity in different ways. Interior-point algorithms solve the original LP by decreasing the barrier parameter ε and following a central path. Sinkhorn fixes an entropic temperature ε and solves a different, KL-regularized OT problem by alternating diagonal scalings. The parameter ε may be decreased in continuation strategies, but for a fixed run it is part of the objective rather than only an algorithmic barrier.

3.3 Relaxation for Arbitrary Measures

This section lifts the finite-dimensional coupling matrix to a joint probability measure. The payoff is that existence, duality and metric properties can be stated for arbitrary laws, including discrete, singular and continuous distributions.

Continuous couplings. The first step is to formalize what it means for a joint law to have prescribed marginals.

Definition 3.25 (Marginals of a joint measure)**.** Let $\pi \in \mathcal{M}_+^1(\mathcal{X}\times\mathcal{Y})$ and let $P_\mathcal{X}(x,y) = x$, $P_\mathcal{Y}(x,y) = y$ be the coordinate projections. The marginals of π are

$$\pi_1 := (P_\mathcal{X})_\sharp \pi \in \mathcal{M}_+^1(\mathcal{X}), \qquad \pi_2 := (P_\mathcal{Y})_\sharp \pi \in \mathcal{M}_+^1(\mathcal{Y}).$$

Equivalently, for all bounded continuous test functions f on $\mathcal{X}$ and g on $\mathcal{Y}$,

$$\int_{\mathcal{X}\times\mathcal{Y}} f(x)\mathrm{d}\pi(x,y) = \int_\mathcal{X} f\,\mathrm{d}\pi_1, \qquad \int_{\mathcal{X}\times\mathcal{Y}} g(y)\mathrm{d}\pi(x,y) = \int_\mathcal{Y} g\,\mathrm{d}\pi_2.$$

A useful mnemonic for the marginal constraint $\pi_1 = \alpha$ and $\pi_2 = \beta$ is the formal row-and-column notation

$$\int_\mathcal{Y} \mathrm{d}\pi(x,y) = \mathrm{d}\alpha(x), \qquad \int_\mathcal{X} \mathrm{d}\pi(x,y) = \mathrm{d}\beta(y),$$

which is made rigorous by Definition 3.25, or equivalently by the identities $\pi(A \times \mathcal{Y}) = \alpha(A)$ and $\pi(\mathcal{X} \times B) = \beta(B)$ for measurable sets $A \subset \mathcal{X}$ and $B \subset \mathcal{Y}$.

Definition 3.26 (Couplings)**.** Given $\alpha \in \mathcal{M}_+^1(\mathcal{X})$ and $\beta \in \mathcal{M}_+^1(\mathcal{Y})$, the set of couplings between α and β is

$$\mathcal{U}(\alpha, \beta) \coloneqq \left\{\pi \in \mathcal{M}_+^1(\mathcal{X} \times \mathcal{Y}) \; ; \; \pi_1 = \alpha \quad \text{and} \quad \pi_2 = \beta\right\}. \tag{3.6}$$

This is the continuous analogue of the transportation polytope (3.1).

Remark 3.27 (Probabilistic interpretation of couplings). If $X \sim \alpha$ and $Y \sim \beta$, then $\pi \in \mathcal{U}(\alpha, \beta)$ means that π is the law of a pair (X, Y) whose coordinates have laws α and β. The coupling encodes the dependence between X and Y. The tensor product $\alpha \otimes \beta$ corresponds to independence, whereas a graph coupling $(\mathrm{Id}, T)_\sharp \alpha$ corresponds to the deterministic relation $Y = T(X)$.
In the discrete case, when $\alpha = \sum_i a_i \delta_{x_i}$ and $\beta = \sum_j b_j \delta_{y_j}$, the constraint $\pi_1 = \alpha$ and $\pi_2 = \beta$ forces every coupling to have the form $\pi = \sum_{i,j} P_{i,j} \delta_{(x_i, y_j)}$ with $P \in U(a, b)$. The discrete formulation is therefore a special case of the continuous one, not merely an approximation.

Unlike the Monge constraint, the coupling constraint is never empty. The continuous feasibility witness is the tensor product coupling, the measure-theoretic version of the discrete product plan above.

Definition 3.28 (Tensor product and trivial coupling)**.** Given $\alpha \in \mathcal{M}_+^1(\mathcal{X})$ and $\beta \in \mathcal{M}_+^1(\mathcal{Y})$, the tensor product coupling $\alpha \otimes \beta$ is the probability measure on $\mathcal{X} \times \mathcal{Y}$ defined by

$$\int_{\mathcal{X}\times\mathcal{Y}} h(x, y) \mathrm{d}(\alpha \otimes \beta)(x, y) = \int_{\mathcal{X}} \left(\int_{\mathcal{Y}} h(x, y) \mathrm{d}\beta(y) \right) \mathrm{d}\alpha(x)$$

for every bounded measurable h. It is also called the trivial coupling because it makes the two coordinates independent.

Indeed, for every $f \in C_b(\mathcal{X})$,

$$\int_{\mathcal{X}\times\mathcal{Y}} f(x) \mathrm{d}(\alpha \otimes \beta)(x, y) = \left(\int_{\mathcal{X}} f(x) \mathrm{d}\alpha(x) \right) \left(\int_{\mathcal{Y}} \mathrm{d}\beta(y) \right) = \int f \mathrm{d}\alpha,$$

and similarly for the second marginal, so $\alpha \otimes \beta \in \mathcal{U}(\alpha, \beta)$.

The next result echoes Proposition 3.4 in the continuous setting. In both cases, the independent coupling is optimal precisely when the objective is flat over the whole admissible set; for continuous costs, this flatness is equivalently the additive form $c(x, y) = u(x) + v(y)$ on the product support.

Proposition 3.29 (Product optimality is degenerate)**.** *Assume that $\mathcal{X}$ and $\mathcal{Y}$ are compact metric spaces and that $c \in C(\mathcal{X} \times \mathcal{Y})$. The following statements are equivalent:*

$$\alpha \otimes \beta \in \operatorname*{argmin}_{\pi \in \mathcal{U}(\alpha,\beta)} \int c \, \mathrm{d}\pi, \qquad \textit{every coupling in } \mathcal{U}(\alpha, \beta) \textit{ is optimal.}$$

They are also equivalent to the additive decomposition of the cost on the product support,

$$c(x, y) = u(x) + v(y), \qquad u \in C(\mathrm{supp}(\alpha)), \quad v \in C(\mathrm{supp}(\beta)).$$

Proof. If every coupling is optimal, then $\alpha \otimes \beta$ is optimal. Conversely, assume that $\alpha \otimes \beta$ is optimal. We first show that, for every $x_0, x_1 \in \mathrm{supp}(\alpha)$ and $y_0, y_1 \in \mathrm{supp}(\beta)$,

$$c(x_0, y_0) + c(x_1, y_1) = c(x_0, y_1) + c(x_1, y_0).$$

Indeed, if this equality failed, after exchanging y_0 and y_1 if necessary one would have a strict inequality

$$c(x_0, y_0) + c(x_1, y_1) > c(x_0, y_1) + c(x_1, y_0).$$

Failure of the equality forces $x_0 \neq x_1$ and $y_0 \neq y_1$. By continuity, the strict inequality persists with a uniform margin on small neighborhoods U_0, U_1 of x_0, x_1 and V_0, V_1 of y_0, y_1, chosen disjoint within

each pair. Since the four points lie in the supports, these neighborhoods have positive marginal mass. Denote by α_i and β_i the normalized restrictions of α to U_i and of β to V_i, and choose

$$0 < \lambda \leqslant \min\{\alpha(U_0)\beta(V_0), \alpha(U_1)\beta(V_1)\}.$$

The exchanged measure

$$\tilde{\pi} = \alpha \otimes \beta - \lambda\, \alpha_0 \otimes \beta_0 - \lambda\, \alpha_1 \otimes \beta_1 + \lambda\, \alpha_0 \otimes \beta_1 + \lambda\, \alpha_1 \otimes \beta_0$$

is nonnegative and has the same two marginals as $\alpha \otimes \beta$. The uniform strict inequality on the neighborhoods implies that $\int c\, \mathrm{d}\tilde{\pi} < \int c\, \mathrm{d}(\alpha \otimes \beta)$, contradicting optimality.

Fixing any $x_\star \in \mathrm{supp}(\alpha)$ and $y_\star \in \mathrm{supp}(\beta)$, the equality of cross differences gives, for all $(x, y) \in \mathrm{supp}(\alpha) \times \mathrm{supp}(\beta)$,

$$c(x, y) = c(x, y_\star) + c(x_\star, y) - c(x_\star, y_\star).$$

Thus $c = u + v$ on the product support, with $u(x) = c(x, y_\star) - c(x_\star, y_\star)$ and $v(y) = c(x_\star, y)$. These functions are continuous on the corresponding supports. Every coupling is concentrated on this product support, so for any $\pi \in \mathcal{U}(\alpha, \beta)$, $\int c\, \mathrm{d}\pi = \int u\, \mathrm{d}\alpha + \int v\, \mathrm{d}\beta$, which depends only on the marginals. Hence all couplings are optimal. □

The tensor product is therefore a trivial feasible coupling, not a typical optimizer. Product optimality means that the cost cannot distinguish between dependences once the marginals are fixed. The continuity assumption is important: if $\alpha = \beta$ is the uniform law on $[0, 1]$ and $c(x, y) = \mathbb{1}_{\{x=y\}}$, then $\alpha \otimes \beta$ has zero cost and is optimal, whereas the identity coupling has cost one. Thus, for arbitrary merely measurable costs, changing the cost on an $\alpha \otimes \beta$-negligible set may affect singular couplings without changing the product cost.

If there exists a map $T : \mathcal{X} \to \mathcal{Y}$ such that $T_\sharp \alpha = \beta$, then the Monge map induces the graph coupling $\pi = (\mathrm{Id}, T)_\sharp \alpha \in \mathcal{U}(\alpha, \beta)$, characterized by

$$\int_{\mathcal{X}\times\mathcal{Y}} h(x, y)\mathrm{d}\pi(x, y) = \int_{\mathcal{X}} h(x, T(x))\mathrm{d}\alpha(x).$$

Applying this identity to $h(x, y) = f(x)$ or $h(x, y) = g(y)$ gives respectively $\pi_1 = \alpha$ and $\pi_2 = \beta$. Thus graph couplings are precisely the Kantorovich representation of deterministic Monge maps. A last important class consists of semi-discrete problems, where α has a density and β is discrete. Every coupling is then supported on the union of the slices $\mathcal{X} \times \{y_j\}$. When an optimal coupling is induced by a map, these slices are selected by a partition of $\mathcal{X}$ into transport cells, as developed in Chapter 6.

> **Remark 3.30 (Book-shifting as a flat Kantorovich face).** The Monge book-shifting example in Example 2.21 has a transparent coupling interpretation. Let α be uniform on $[0, 2]$ and β uniform on $[1, 3]$. For every $\pi \in \mathcal{U}(\alpha, \beta)$,
>
> $$\int |y - x|\mathrm{d}\pi(x, y) \geqslant \int (y - x)\mathrm{d}\pi(x, y) = \int y\mathrm{d}\beta(y) - \int x\mathrm{d}\alpha(x) = 1.$$
>
> Equality holds exactly for couplings concentrated on the half-plane $\{(x, y) : y \geqslant x\}$, where $|y-x| = y-x$. Hence the optimal set is a whole face of the coupling polytope, not a single graph. The translation and book-shifting maps give two graph couplings inside this face, but many non-deterministic couplings are optimal as well.

Continuous Kantorovich problem. The discrete Kantorovich relaxation of Section 3.1, formalized in Definition 3.5, extends from nonnegative coupling matrices to coupling measures between arbitrary probability measures.

Definition 3.31 (Continuous Kantorovich problem)**.** For $\alpha \in \mathcal{M}_+^1(\mathcal{X})$, $\beta \in \mathcal{M}_+^1(\mathcal{Y})$, and a nonnegative Borel cost $c : \mathcal{X} \times \mathcal{Y} \to [0, +\infty]$, define

$$\mathcal{L}_c(\alpha, \beta) := \inf_{\pi \in \mathcal{U}(\alpha,\beta)} \int_{\mathcal{X}\times\mathcal{Y}} c(x, y)\mathrm{d}\pi(x, y). \tag{3.7}$$

A minimizer is an optimal coupling, or optimal transport plan. The value is understood in $[0, +\infty]$.

This is an infinite-dimensional linear program over a space of measures.

The following result generalizes Proposition 3.6 from coupling matrices to coupling measures: the Kantorovich value is convex in the marginals but concave in the ground cost.

Proposition 3.32 (Convexity in the marginals and concavity in the cost)**.** *Fix a nonnegative Borel cost c. The extended-valued map*

$$(\alpha, \beta) \longmapsto \mathcal{L}_c(\alpha, \beta)$$

is jointly convex on pairs of probability measures. Conversely, for fixed marginals (α, β), the map $c \mapsto \mathcal{L}_c(\alpha, \beta)$ is concave on the cone of nonnegative Borel costs: for any such c_0, c_1 and $t \in [0,1]$,

$$\mathcal{L}_{(1-t)c_0+tc_1}(\alpha, \beta) \geqslant (1-t)\mathcal{L}_{c_0}(\alpha, \beta) + t\mathcal{L}_{c_1}(\alpha, \beta).$$

The inequality is understood in $[0, +\infty]$; in particular, the map is concave on every convex family of costs on which it is finite.

Proof. For joint convexity, let (α_0, β_0) and (α_1, β_1) be two pairs of probability measures. The claim is immediate at $t \in \{0,1\}$ or if one of the two values on the right-hand side is infinite. Otherwise, for $\eta > 0$, choose $\pi_i \in \mathcal{U}(\alpha_i, \beta_i)$ such that $\int c \mathrm{d}\pi_i \leqslant \mathcal{L}_c(\alpha_i, \beta_i) + \eta$ for $i \in \{0,1\}$. Then $(1-t)\pi_0 + t\pi_1$ couples $(1-t)\alpha_0 + t\alpha_1$ and $(1-t)\beta_0 + t\beta_1$, and hence

$$\mathcal{L}_c\big((1-t)\alpha_0 + t\alpha_1, (1-t)\beta_0 + t\beta_1\big) \leqslant (1-t)\mathcal{L}_c(\alpha_0, \beta_0) + t\mathcal{L}_c(\alpha_1, \beta_1) + \eta.$$

Letting $\eta \to 0$ proves joint convexity.

For concavity, the endpoint cases are immediate, so assume $t \in (0,1)$ and fix $\pi \in \mathcal{U}(\alpha, \beta)$. Linearity of integration for nonnegative functions gives

$$\int \big((1-t)c_0 + tc_1\big)\mathrm{d}\pi \geqslant (1-t)\mathcal{L}_{c_0}(\alpha, \beta) + t\mathcal{L}_{c_1}(\alpha, \beta).$$

Taking the infimum over π proves the stated inequality. □

> **Remark 3.33 (Probabilistic interpretation of Kantorovich's problem).** The continuous Kantorovich problem of Definition 3.31 can equivalently be written as
>
> $$\mathcal{L}_c(\alpha, \beta) = \inf_{X \sim \alpha,\, Y \sim \beta} \mathbb{E}(c(X, Y)).$$
>
> The minimization is not over the marginal laws, which are fixed, but over all possible dependences between the two random variables. OT therefore chooses the cheapest joint law among all couplings.

Existence is one of the decisive gains of Kantorovich relaxation. The Monge problem of Definition 2.14 may have no admissible map, and even when maps exist its infimum need not be attained because graph couplings are not closed under weak limits. By contrast, $\mathcal{U}(\alpha, \beta)$ always contains the product coupling $\alpha \otimes \beta$; tightness and weak closedness then make the direct method available.

Proposition 3.34 (Existence for lower-semicontinuous costs)**.** *Assume that $\mathcal{X}$ and $\mathcal{Y}$ are Polish spaces and that $c : \mathcal{X} \times \mathcal{Y} \to [0, +\infty]$ is lower semicontinuous. Then the Kantorovich problem (3.7) admits at least one minimizer. In particular, this applies to every continuous nonnegative cost on compact metric spaces.*

Proof. The constraint set is non-empty because it contains the product coupling $\alpha \otimes \beta$. It is uniformly tight: given $\varepsilon > 0$, choose compact sets $K_{\mathcal{X}} \subset \mathcal{X}$ and $K_{\mathcal{Y}} \subset \mathcal{Y}$ such that $\alpha(K_{\mathcal{X}}) \geqslant 1 - \varepsilon/2$ and $\beta(K_{\mathcal{Y}}) \geqslant 1 - \varepsilon/2$. Every $\pi \in \mathcal{U}(\alpha, \beta)$ then satisfies $\pi(K_{\mathcal{X}} \times K_{\mathcal{Y}}) \geqslant 1 - \varepsilon$. By Prokhorov's theorem [69, 539], the set of couplings is relatively compact for weak convergence; the weak topology is reviewed in Section 4.2. It is also weakly closed, because the coordinate projections are continuous and therefore preserve the marginal identities in the limit. Hence $\mathcal{U}(\alpha, \beta)$ is weakly compact. Finally, Portmanteau's theorem makes $\pi \mapsto \int c \mathrm{d}\pi$ weakly lower semicontinuous. The direct method gives a minimizer. □

Axiomatic characterization of the relaxation. Kantorovich relaxation is not only a convenient enlargement of Monge's problem: convexity and weak stability single it out as the maximal extension of the pointwise ground cost from pairs of points to pairs of probability measures.

Theorem 3.35 (Kantorovich cost as the maximal convex relaxation)**.** *Assume that $\mathcal{X}$ and $\mathcal{Y}$ are compact metric spaces and that $c : \mathcal{X} \times \mathcal{Y} \to [0, +\infty)$ is continuous. Then $\mathcal{L}_c$ is the pointwise largest proper functional*

$$\mathsf{F} : \mathcal{M}_+^1(\mathcal{X}) \times \mathcal{M}_+^1(\mathcal{Y}) \longrightarrow (-\infty, +\infty]$$

that is jointly convex, weakly lower semicontinuous, and satisfies

$$\mathsf{F}(\delta_x, \delta_y) \leqslant c(x, y) \qquad \text{for every } (x, y) \in \mathcal{X} \times \mathcal{Y}.$$

More precisely, every such F satisfies $\mathsf{F}(\alpha, \beta) \leqslant \mathcal{L}_c(\alpha, \beta)$ for all (α, β), while $\mathcal{L}_c$ itself has these properties and obeys $\mathcal{L}_c(\delta_x, \delta_y) = c(x, y)$.

Proof. Joint convexity follows from Proposition 3.32. For weak lower semicontinuity, consider converging marginals and optimal plans along a subsequence realizing the lower limit. Compactness gives a further weakly convergent subsequence; its limit has the limiting marginals, and lower semicontinuity of the integral gives the claim, exactly as in Proposition 3.34. The only coupling of (δ_x, δ_y) is $\delta_{(x,y)}$, which proves the identity on Dirac pairs.

Fix $\pi \in \mathcal{U}(\alpha, \beta)$. Its marginal constraints are exactly the barycentric identity $(\alpha, \beta) = \int (\delta_x, \delta_y) \mathrm{d}\pi$. Jensen's inequality for lower-semicontinuous convex functionals therefore gives

$$\mathsf{F}(\alpha, \beta) \leqslant \int \mathsf{F}(\delta_x, \delta_y) \mathrm{d}\pi(x, y) \leqslant \int c(x, y) \mathrm{d}\pi(x, y).$$

Taking the infimum over $\pi \in \mathcal{U}(\alpha, \beta)$ proves maximality. □

This is the probability-measure specialization of the convex-envelope theorem of Savaré and Sodini [472, Theorem 4.4], which treats finite measures on completely regular spaces. The word *largest* is essential: joint convexity, weak lower semicontinuity and equality on Dirac pairs only imply that another extension lies below $\mathcal{L}_c$; they do not force it to coincide with $\mathcal{L}_c$.

Monge–Kantorovich equivalence. The proof of Brenier's theorem 2.23 relies on Kantorovich relaxation and duality. It proves that, under its hypotheses, the relaxation is tight: it has the same cost as the Monge problem and its optimal coupling is induced by a map.

Corollary 3.36 (Monge–Kantorovich equivalence under Brenier)**.** *Let $\alpha, \beta \in \mathcal{P}_2(\mathbb{R}^d)$, assume that α is absolutely continuous with respect to Lebesgue measure, and let $c(x, y) = \|x - y\|^2$. If T is the Brenier map solving Monge's problem, then $\pi = (\mathrm{Id}, T)_\sharp \alpha$ is the unique optimal coupling solving the Kantorovich problem. In particular, Monge and Kantorovich costs are the same.*

Proof. The proof of Brenier's theorem shows that the support of any optimal Kantorovich plan is contained in the subdifferential $\partial\varphi$ of a convex function φ. When α has a density, φ is differentiable α-almost everywhere, so $\partial\varphi(x) = \{\nabla\varphi(x)\}$ for α-almost every x. Thus every optimal coupling is concentrated on the graph of $T = \nabla\varphi$ and must equal $(\mathrm{Id}, T)_\sharp \alpha$. The graph coupling is feasible and optimal, and the two formulations have the same value. □

The density assumption is exactly what prevents the relaxed plan from using several destinations at a nonsmooth point.

> **Remark 3.37 (Nonsmooth potentials and splitting).** If α does not have a density, then φ may be non-smooth on a set charged by α, and non-smooth points can lead to mass splitting. For instance, moving δ_0 to $(\delta_{-1} + \delta_{+1})/2$ can be represented by a plan concentrated on the set-valued subdifferential of $\varphi(x) = |x|$, but not by a deterministic map. This is the continuous counterpart of the gap between the uniform matching case of Corollary 3.18 and the general splitting case.

> **Remark 3.38 (Probabilistic form of tightness).** If (X, Y) has the optimal Kantorovich law under the assumptions of Corollary 3.36, then $Y = T(X)$ almost surely with $X \sim \alpha$ and $T(X) \sim \beta$. This is analogous to the Birkhoff–von Neumann result in the fully discrete uniform case: in both settings, the convex relaxation admits an optimizer satisfying the original deterministic constraint. The hypotheses are quite different, however: Birkhoff–von Neumann is finite-dimensional and need

not give uniqueness, whereas Brenier's theorem uses absolute continuity of the source and gives uniqueness of the optimal map almost everywhere.

Kantorovich solution in 1D. The atomless assumption in the Monge statement of Section 2.6 is a limitation of maps, not of one-dimensional optimality. Once couplings are allowed, atoms can be split by assigning subintervals of quantile levels to different target points. The common quantile parameter therefore defines an optimal relaxed coupling for arbitrary probability measures.

Theorem 3.39 (One-dimensional Kantorovich solution)**.** *Let $\alpha, \beta \in \mathcal{M}_+^1(\mathbb{R})$ and let $h : \mathbb{R} \to [0, +\infty)$ be convex. Consider the cost $c(x, y) = h(x - y)$. Write $q_\alpha := C_\alpha^{-1}$ and $q_\beta := C_\beta^{-1}$, and assume that $h(q_\alpha - q_\beta) \in L^1(0, 1)$. Then the quantile coupling*

$$\pi^\star = (q_\alpha, q_\beta)_\sharp \mathrm{Leb}_{[0,1]} = (C_\alpha^{-1}, C_\beta^{-1})_\sharp \mathrm{Leb}_{[0,1]}$$

solves the Kantorovich problem (3.7) for $c(x, y) = h(x - y)$. In particular, $h(t) = |t|^p$ gives the usual one-dimensional optimal coupling for every $p \geqslant 1$.

Proof. The push-forward statement $\pi^\star \in \mathcal{U}(\alpha, \beta)$ follows from Proposition 2.40. It remains to prove optimality.

The key point is the one-dimensional uncrossing inequality. If $x < x'$ and $y > y'$, set $a = x - y$, $\delta = x' - x > 0$ and $\eta = y - y' > 0$. Convexity of h implies that increments are monotone, hence

$$h(a + \eta) - h(a) \leqslant h(a + \delta + \eta) - h(a + \delta),$$

which is exactly $h(x - y) + h(x' - y') \geqslant h(x - y') + h(x' - y)$. Thus removing a crossing never increases the cost. For a finite transport matrix on two ordered grids, if $i < i'$ and $j > j'$ carry crossed masses $\mathrm{P}_{i,j}$ and $\mathrm{P}_{i',j'}$, one may move $\theta = \min(\mathrm{P}_{i,j}, \mathrm{P}_{i',j'})$ units of mass from the crossed entries (i, j), (i', j') to the uncrossed entries (i, j'), (i', j). The row and column sums are unchanged, and the preceding inequality shows that the cost does not increase. Repeating this elementary uncrossing step yields an ordered plan; on an ordered uniform quantile grid, this is the diagonal plan.

For general measures, apply the same argument in quantile coordinates. Let $\pi \in \mathcal{U}(\alpha, \beta)$. Let U_α and U_β be uniform random variables on $(0, 1)$, and denote by $K_\alpha(x, \mathrm{d}r)$ and $K_\beta(y, \mathrm{d}s)$ regular conditional laws of U_α given $q_\alpha(U_\alpha) = x$ and of U_β given $q_\beta(U_\beta) = y$. Given $(X, Y) \sim \pi$, draw conditionally $R \sim K_\alpha(X, \cdot)$ and $S \sim K_\beta(Y, \cdot)$. The law γ of (R, S) has uniform marginals and satisfies $\pi = (q_\alpha, q_\beta)_\sharp \gamma$.

We now justify the approximation. Let $\kappa_M(r) = \max(-M, \min(r, M))$ and set $q_{\alpha,M} = \kappa_M \circ q_\alpha$ and $q_{\beta,M} = \kappa_M \circ q_\beta$. These functions are bounded and nondecreasing. Approximate them almost everywhere by nondecreasing step functions $q_{\alpha,M}^N$ and $q_{\beta,M}^N$ that are constant on the uniform intervals $I_k = ((k - 1)/N, k/N]$. The matrix $G_{k,\ell}^N = \gamma(I_k \times I_\ell)$ is a coupling of the two uniform histograms. Proposition 3.10, applied to the ordered step values, therefore gives

$$\int_0^1 h(q_{\alpha,M}^N(r) - q_{\beta,M}^N(r))\mathrm{d}r \leqslant \int_{(0,1)^2} h(q_{\alpha,M}^N(r) - q_{\beta,M}^N(s))\mathrm{d}\gamma(r, s).$$

For fixed M, the arguments of h remain in the compact interval $[-2M, 2M]$. Dominated convergence as $N \to \infty$ yields the same inequality for $q_{\alpha,M}$ and $q_{\beta,M}$.

Finally, κ_M is nondecreasing and 1-Lipschitz. Hence, for every $x, y \in \mathbb{R}$, there exists $t_M(x, y) \in [0, 1]$ such that $\kappa_M(x) - \kappa_M(y) = t_M(x, y)(x - y)$. Convexity and nonnegativity imply

$$h(t_M(x, y)(x - y)) \leqslant (1 - t_M(x, y))h(0) + t_M(x, y)h(x - y) \leqslant h(0) + h(x - y).$$

The assumed integrability controls the diagonal term. If the cost of π is finite, the same bound controls the right-hand side; if it is infinite, the desired comparison is immediate. Dominated convergence as $M \to \infty$ therefore gives

$$\int_0^1 h(q_\alpha(r) - q_\beta(r))\mathrm{d}r \leqslant \int h(x - y)\mathrm{d}\pi(x, y)$$

for every $\pi \in \mathcal{U}(\alpha, \beta)$. □

This result is strictly more flexible than the Monge formula. If α has an atom, a map can only send that whole atom to one target point, whereas the quantile interval associated with the atom can be coupled with a nontrivial portion of β. The one-dimensional Kantorovich solution therefore handles mass splitting without changing the monotone geometry.

3.4 Cone of Convex Functions and Monge Gap

For quadratic transport, Brenier's theorem 2.23 identifies optimal maps with gradients, or more generally subgradients, of convex functions. Convex functions form a cone under addition and nonnegative scaling, but imposing this global structure directly on a learned map can be difficult. The Monge gap turns transport optimality into a variational certificate: for suitable costs it is itself convex, on the line it becomes an isotonicity penalty, and in learning problems it can be added to an ordinary regression objective.

Definition and optimality certificate. Matching an output law does not identify an optimal map: many maps can have the same push-forward while incurring different transport costs. The Monge gap of Uscidda and Cuturi [526] separates distributional fitting from transport optimality by comparing the graph coupling generated by a map with the best coupling having exactly the same marginals.

Definition 3.40 (Monge gap)**.** Let $\alpha \in \mathcal{M}_+^1(\mathcal{X})$, let $T : \mathcal{X} \to \mathcal{Y}$ be measurable, and assume that the following costs are finite. The c-Monge gap of T relative to α is

$$\mathcal{M}_\alpha^c(T) := \int c(x, T(x))\mathrm{d}\alpha(x) - \mathcal{L}_c(\alpha, T_\sharp\alpha). \tag{3.8}$$

The second marginal is deliberately $T_\sharp\alpha$: the gap tests optimality of T, not agreement with an externally prescribed target.

Proposition 3.41 (Elementary properties of the Monge gap)**.** *Under the standing Polish-space assumptions, suppose that the relevant integrals are finite. Then $\mathcal{M}_\alpha^c(T) \geqslant 0$, with equality if and only if $(\mathrm{Id}, T)_\sharp\alpha$ is an optimal coupling between α and $T_\sharp\alpha$. Moreover,*

$$\mathcal{M}_\alpha^c(T) = \sup_{\eta \in \mathcal{U}(\alpha,\alpha)} \int \big[c(x, T(x)) - c(x, T(z))\big]\mathrm{d}\eta(x, z). \tag{3.9}$$

Finally, the gap is unchanged when $c(x, y)$ is replaced by $c(x, y) + r(x) + s(y)$, provided the modified cost remains admissible and all marginal terms are integrable.

Proof. The graph plan $(\mathrm{Id}, T)_\sharp\alpha$ is feasible for $\mathcal{L}_c(\alpha, T_\sharp\alpha)$, which proves nonnegativity and the equality criterion. Every $\eta \in \mathcal{U}(\alpha, \alpha)$ induces the coupling $(x, z) \mapsto (x, T(z))$. Conversely, disintegrating α along T lifts every coupling in $\mathcal{U}(\alpha, T_\sharp\alpha)$ to such a self-coupling. Subtracting the resulting infimum from the graph cost gives (3.9). Adding $r(x) + s(y)$ shifts both terms in (3.8) by $\int r\mathrm{d}\alpha + \int s\mathrm{d}(T_\sharp\alpha)$. □

Convex gaps and convex potentials. Formula (3.9) need not be convex in T: both occurrences of T lie inside a generic nonlinear cost. Convexity becomes automatic when cost differences are affine in the image variable.

Proposition 3.42 (Convex Monge gaps for affine-difference costs)**.** *Let $\mathcal{Y} \subset \mathbb{R}^m$ be convex and let the nonnegative cost c have the form*

$$c(x, y) = a(x) + b(y) - \langle A(x),\, y\rangle, \tag{3.10}$$

where $A, T \in L^2(\alpha; \mathbb{R}^m)$ and all displayed terms are finite. Then

$$\mathcal{M}_\alpha^c(T) = \sup_{\eta \in \mathcal{U}(\alpha,\alpha)} \int \langle A(x),\, T(z) - T(x)\rangle\mathrm{d}\eta(x, z). \tag{3.11}$$

Consequently, $T \mapsto \mathcal{M}_\alpha^c(T)$ is convex, and it is positively homogeneous when the admissible maps take values in $\mathbb{R}^m$. Moreover, $\mathcal{M}_\alpha^c(T) = 0$ if and only if there is a proper lower-semicontinuous convex function φ such that

$$T(x) \in \partial\varphi(A(x)) \qquad \text{for } \alpha\text{-almost every } x. \tag{3.12}$$

Proof. The marginal terms a and b cancel in (3.9), giving (3.11); this is a supremum of linear functionals of T. Next, disintegrating α along A lifts every coupling between $A_\sharp\alpha$ and $T_\sharp\alpha$. Hence zero gap is equivalent to optimality of $(A, T)_\sharp\alpha$ for the bilinear cost $(u, y) \mapsto -\langle u,\, y\rangle$. Up to marginal terms, this is the quadratic cost $\frac{1}{2}\|u - y\|^2$. By Rockafellar's characterization [452], recalled in Section 3.5, its optimal plans are exactly those concentrated on the subdifferential of a proper lower-semicontinuous convex function. □

The class (3.10) contains $c(x,y) = \frac{1}{2}\|x-y\|^2$, for which $A(x) = x$. It also contains the Bregman divergence of Definition 9.1 with x as base point,

$$B_\Phi(y|x) = \Phi(y) - \Phi(x) - \langle \nabla\Phi(x),\, y - x\rangle,$$

where Φ is differentiable and convex: here $a(x) = \langle \nabla\Phi(x),\, x\rangle - \Phi(x)$, $b(y) = \Phi(y)$ and $A(x) = \nabla\Phi(x)$. The argument order is essential: on a full-dimensional domain, the reversed divergence $B_\Phi(x|y)$ is generally not of the form (3.10) unless Φ is quadratic. Thus the zero-gap condition generalizes Brenier's characterization: in the quadratic case it reduces to $T(x) \in \partial\varphi(x)$ α-almost everywhere, and to $T = \nabla\varphi$ when α has a density.

This cost class is also exhaustive under mild smoothness. If $\mathcal{Y}$ is open and $c(x,\cdot) \in C^2(\mathcal{Y})$, then the gap is convex for every source α for which it is finite—and it is enough to test finitely supported sources—if and only if c has the form (3.10). Indeed, for $\alpha = \frac{1}{2}(\delta_{x_0} + \delta_{x_1})$, $u = T(x_0)$, $v = T(x_1)$ and $g(y) = c(x_0,y) - c(x_1,y)$, one has $\mathcal{M}^c_\alpha(T) = \frac{1}{2}(g(u) - g(v))_+$. Convexity in (u,v) forces $\nabla^2 g = 0$, so every such difference is affine.

One-dimensional order and isotonicity. On the line, the order structure gives a more explicit formula, even for convex displacement costs whose Monge gap is not itself convex.

Definition 3.43 (Increasing rearrangement relative to a source)**.** Let $\alpha \in \mathcal{M}^1_+(\mathbb{R})$ be atomless and let $T : \mathbb{R} \to \mathbb{R}$ be measurable. The increasing rearrangement of T relative to α is

$$T^{\uparrow,\alpha} := C^{-1}_{T_\sharp\alpha} \circ C_\alpha. \tag{3.13}$$

It is the unique α-almost-everywhere nondecreasing map satisfying $(T^{\uparrow,\alpha})_\sharp\alpha = T_\sharp\alpha$.

This rearrangement is not intrinsic to T alone: it depends on α through both the source rank C_α and the output law $T_\sharp\alpha$. If α has atoms, a deterministic increasing rearrangement need not exist because an atom may have to split; the quantile coupling of Theorem 3.39 is then the correct replacement.

Proposition 3.44 (One-dimensional Monge-gap formulas)**.** *Let $\alpha \in \mathcal{M}^1_+(\mathbb{R})$ be atomless, let $c(x,y) = h(x-y)$ with h convex, and assume the relevant terms are integrable. Then*

$$\mathcal{M}^c_\alpha(T) = \int \big[h(x - T(x)) - h(x - T^{\uparrow,\alpha}(x))\big]\,\mathrm{d}\alpha(x). \tag{3.14}$$

If h is strictly convex, this gap vanishes if and only if T has a nondecreasing representative α-almost everywhere.

For the equal-weight atomic measure $\alpha_n = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$ on the uniform grid $x_i = x_1 + (i-1)\Delta$, let $t_i = T(x_i)$ and let $t_{(1)} \leqslant \cdots \leqslant t_{(n)}$ be their order statistics. For $c(x,y) = |x-y|^2$,

$$\mathcal{M}^{|\cdot|^2}_{\alpha_n}(T) = \frac{1}{n}\sum_{i=1}^n \big[(x_i - t_i)^2 - (x_i - t_{(i)})^2\big] = \frac{2\Delta}{n} \sum_{1\leqslant i<j\leqslant n} (t_i - t_j)_+. \tag{3.15}$$

Proof. The one-dimensional rearrangement result of Theorem 2.41, together with its coupling extension in Theorem 3.39, shows that $T^{\uparrow,\alpha}$ is optimal between α and $T_\sharp\alpha$. This gives (3.14). If h is strictly convex, atomlessness makes this optimal map unique α-almost everywhere, so zero gap is equivalent to $T = T^{\uparrow,\alpha}$ α-almost everywhere. For the empirical formula, sorting $(t_i)_i$ gives the increasing optimal assignment and hence the first equality. An adjacent exchange of an inversion $u > v$ decreases the average squared cost by $2\Delta(u-v)/n$; bubble sort exchanges every inverted pair once, which gives the second equality. □

For quadratic cost, formula (3.15) is a convex pairwise hinge penalty whose zero set is the isotonic cone $\{t_1 \leqslant \cdots \leqslant t_n\}$. It is not the usual least-squares isotonic projection: rearrangement preserves the empirical output law, whereas isotonic regression moves and pools values. In higher dimension, Proposition 3.42 replaces scalar order by cyclic monotonicity and convex subgradients. Monge-gap regularization therefore extends the role of isotonic constraints beyond a total order, while its defining comparison keeps $T_\sharp\alpha$ fixed.

Monge-gap-regularized kernel regression. The hinge representation above turns transport optimality into a tractable soft monotonicity prior. Let $x_i = x_1 + (i-1)\Delta$ and let $y_i \in \mathbb{R}$ be noisy observations. For a positive-semidefinite kernel k with reproducing-kernel Hilbert space $\mathcal{H}_k$, consider

$$\min_{T \in \mathcal{H}_k} \left\{ \frac{1}{n} \sum_{i=1}^{n} |T(x_i) - y_i|^2 + \lambda \|T\|_{\mathcal{H}_k}^2 + \gamma \mathcal{M}_{\alpha_n}^{|\cdot|^2}(T) \right\}, \qquad \lambda > 0, \quad \gamma \geqslant 0, \tag{3.16}$$

where $\alpha_n = n^{-1} \sum_i \delta_{x_i}$. The usual orthogonal-projection argument shows that a minimizer has the representer form $T_{\mathrm{q}}(\cdot) = \sum_{j=1}^{n} \mathrm{q}_j k(x_j, \cdot)$. Writing $\mathrm{K}_{i,j} = k(x_i, x_j)$ and $\mathrm{y} = (y_i)_i$, formula (3.15) reduces (3.16) to

$$\min_{\mathrm{q} \in \mathbb{R}^n} \left\{ \frac{1}{n} \|\mathrm{Kq} - \mathrm{y}\|^2 + \lambda \langle \mathrm{q}, \mathrm{Kq} \rangle + \frac{2\gamma\Delta}{n} \sum_{1 \leqslant i < j \leqslant n} \big[(\mathrm{Kq})_i - (\mathrm{Kq})_j \big]_+ \right\}. \tag{3.17}$$

This is a finite-dimensional convex program: the first two terms are convex quadratics because $\mathrm{K} \succeq 0$, and the last term is a sum of hinges of affine functionals. Thus every point along the regularization path can be computed globally, without alternating between regression and transport.

Figure 3.9 shows this path for a Gaussian kernel. The left experiment reproduces a smooth oscillatory example with $n = 55$ observations. The right experiment uses $n = 95$ noisier observations from a signal containing several spatial scales. Increasing γ suppresses order inversions rather than merely smoothing the fit; the black dashed curve is the solution of the hard constraint $T(x_1) \leqslant \cdots \leqslant T(x_n)$.

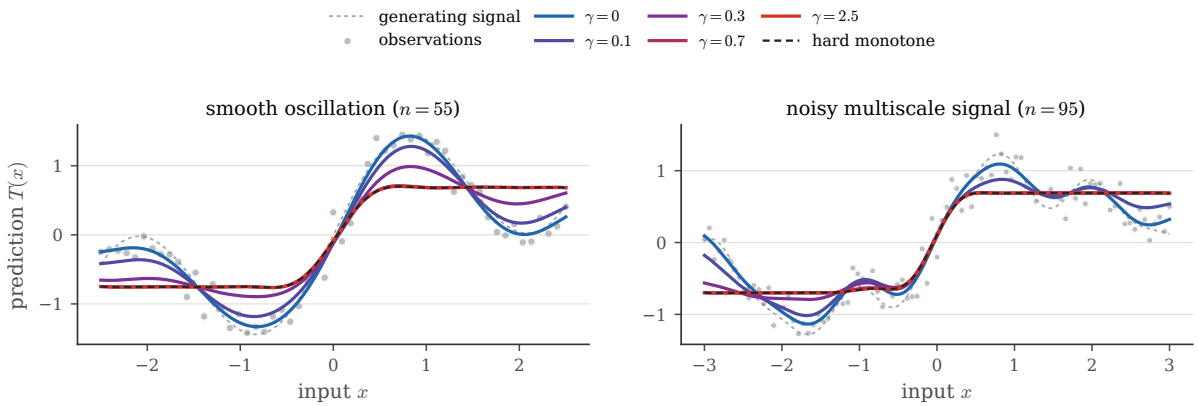


Figure 3.9: A quadratic Monge-gap penalty traces a convex kernel-regression path toward monotonicity. Gray dots are the observations and the light dashed curve is the generating signal. Colored curves solve (3.16) for increasing γ, from ordinary kernel ridge regression in blue to an almost monotone fit in red; the black dashed curve solves the corresponding hard monotonicity-constrained problem. The second experiment uses more points, stronger noise and a more oscillatory signal.

3.5 c-Cyclical Monotonicity

Cyclical monotonicity is the local geometric fingerprint of optimality. It converts a global minimization problem into finite exchange inequalities and is the bridge from Kantorovich plans to convex potentials.

Optimal transport plans behave well when one looks at any finite sub-collection of points in their support: the restriction is still an optimal matching for those points alone. For finitely supported marginals this is immediate, and it leads to the notion of c-cyclical monotonicity.

Support and c-cyclical monotonicity. The support of a coupling is the topological support introduced in Definition 2.4, now applied to a Radon measure on $\mathcal{X} \times \mathcal{Y}$. Thus $(x, y) \in \mathrm{supp}(\pi)$ exactly when every open neighborhood of (x, y) has positive π-mass.

Definition 3.45 (c-cyclical monotonicity)**.** A set $\Gamma \subset \mathcal{X} \times \mathcal{Y}$ is c-cyclically monotone if, for every $k \geqslant 2$,

every finite family $(x_i, y_i)_{i=1}^k \subset \Gamma$ and every permutation σ of $\{1, \dots, k\}$,

$$\sum_{i=1}^k c(x_i, y_i) \leqslant \sum_{i=1}^k c(x_i, y_{\sigma(i)}).$$

Any permutation is a product of cycles, so it suffices to verify the inequality for cyclic permutations,

$$\sum_{i=1}^k c(x_i, y_i) \leqslant \sum_{i=1}^k c(x_i, y_{i+1}), \qquad y_{k+1} = y_1.$$

Optimal matching to optimal transport. Let the marginals be uniform on n points, $\alpha = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$ and $\beta = \frac{1}{n}\sum_{i=1}^n \delta_{y_i}$. By Corollary 3.18, there exists an optimal plan induced by a permutation. Its support $\Gamma = \{(x_i, y_{\sigma(i)})\}_i$ is c-cyclically monotone: otherwise exchanging the finitely many targets along a violating cycle would lower the matching cost. The following theorem, in the spirit of Rockafellar's cyclic-monotonicity theorem [452], says that the same finite-exchange condition holds for any optimal coupling, not only for finite uniform matchings.

Theorem 3.46 (Optimal plans are c-cyclically monotone)**.** *Assume that $c : \mathcal{X} \times \mathcal{Y} \to [0, +\infty)$ is continuous and that the Kantorovich value is finite. For any optimal plan π solving (3.7), $\operatorname{supp}(\pi)$ is c-cyclically monotone.*

Proof. We prove the contrapositive. Suppose that $\operatorname{supp}(\pi)$ is not c-cyclically monotone. Then there exist points $(x_i, y_i)_{i=1}^k$ in the support and a permutation σ such that $\sum_i c(x_i, y_i) > \sum_i c(x_i, y_{\sigma(i)})$. By continuity of c, after shrinking neighborhoods $U_i \ni x_i$ and $V_i \ni y_i$, the same strict inequality holds uniformly for every choice of points in these neighborhoods:

$$\sum_i c(u_i, v_i) > \sum_i c(u_i, \tilde{v}_{\sigma(i)}) \qquad (u_i \in U_i,\ v_i \in V_i,\ \tilde{v}_{\sigma(i)} \in V_{\sigma(i)}).$$

Choose the sets so that $m_i := \pi(U_i \times V_i) > 0$. Taking $0 < \lambda \leqslant \left(\sum_{i=1}^k \frac{1}{m_i}\right)^{-1}$ ensures that the scaled restrictions $\pi_i = \lambda \frac{\pi|_{U_i \times V_i}}{m_i}$ have common mass λ and satisfy $\sum_i \pi_i \leqslant \pi$, even if some rectangles overlap. Let $\alpha_i = (P_{\mathcal{X}})_\sharp \pi_i$ and $\beta_i = (P_{\mathcal{Y}})_\sharp \pi_i$, and define $\tilde{\pi} = \pi - \sum_i \pi_i + \sum_i (\alpha_i \otimes \beta_{\sigma(i)})/\lambda$. The removed and reinserted first marginals are both $\sum_i \alpha_i$, and the removed and reinserted second marginals are both $\sum_i \beta_i$ because σ is a permutation. Hence $\tilde{\pi} \in \mathcal{U}(\alpha, \beta)$. Integrating the uniform strict inequality against the product probability $\otimes_i (\pi_i/\lambda)$ shows that the reinserted crossed terms have strictly smaller cost than the removed diagonal terms. This contradicts the optimality of π. □

Monotonicity. Assume the optimal plan is induced by a measurable map $T : \mathcal{X} \to \mathcal{Y}$, meaning that $\pi = (\mathrm{Id}, T)_\sharp \alpha$. There is a set $G \subset \mathcal{X}$ of full α-measure such that $(x, T(x)) \in \operatorname{supp}(\pi)$ for every $x \in G$. For any $x_1, \dots, x_k \in G$, cyclical monotonicity reads

$$\sum_{i=1}^k c(x_i, T(x_i)) \leqslant \sum_{i=1}^k c(x_i, T(x_{i+1})), \qquad x_{k+1} = x_1.$$

For $c(x, y) = \frac{1}{2}\|x - y\|^2$, taking $k = 2$ gives $\langle T(x) - T(y),\ x - y\rangle \geqslant 0$ for every $x, y \in G$, so the optimal representative of T is monotone on G. Brenier's theorem adds that, when α is absolutely continuous, $T = \nabla\varphi$ almost everywhere for a convex potential φ. The converse fails in dimension $d \geqslant 2$: as shown by the rotation example in the Monge section, a small rotation is monotone yet not a gradient.

One dimension. In one space dimension, with cost $c(x, y) = |x - y|^p$, the two-point inequality becomes

$$|x - T(x)|^p + |y - T(y)|^p \leqslant |x - T(y)|^p + |y - T(x)|^p,$$

For $p > 1$, strict convexity implies that a reversed pair $x < y$ and $T(x) > T(y)$ violates this inequality. Hence every optimal map is nondecreasing on the full-measure transport set, recovering the classical monotone rearrangement. For $p = 1$, uncrossing is non-strict: the monotone rearrangement remains optimal, but other optimal maps and non-deterministic plans can occur, as in the book-shifting example of Remark 3.30.

The coupling formalism now provides a well-posed convex relaxation together with structural certificates for its optimizers. The next chapter specializes the cost to powers of a metric and turns these optimal values into a geometry on probability measures.

CHAPTER 4

Wasserstein Space

The preceding chapter introduced couplings as a convex relaxation of transport maps and established the Kantorovich problem as a well-posed optimization principle. When the ground cost is a power of a metric, its optimal value also equips probability measures with a metric geometry. This chapter develops the resulting Wasserstein space, from gluing and displacement geodesics to topology, stable maps on measures and distributional robustness.

4.1 Wasserstein Distances

This section proves that OT costs are genuine distances when the ground cost comes from a metric. It also constructs displacement geodesics and explains how Kantorovich transport relaxes the directed Monge formulation.

OT defines a distance. An important feature of OT is that it defines a distance between histograms and probability measures as soon as the cost matrix satisfies certain suitable properties. Indeed, OT can be understood as a canonical way to lift a ground distance between points to a distance between histograms or measures. The proof of this result relies on a "gluing lemma", which we first prove in the discrete case.

Lemma 4.1 (Discrete gluing lemma)**.** *Given* $(a, b, c) \in \Sigma_n \times \Sigma_p \times \Sigma_m$*, let* $P \in U(a, b)$ *and* $Q \in U(b, c)$*. Then there exists a nonnegative three-way array* $S \in \mathbb{R}_+^{n\times p\times m}$ *such that its two-way marginals satisfy*

$$\sum_k S_{i,j,k} = P_{i,j} \quad \text{and} \quad \sum_i S_{i,j,k} = Q_{j,k}.$$

Consequently, the marginal between the first and third variables, $R_{i,k} := \sum_j S_{i,j,k}$*, belongs to* $U(a, c)$*. For the canonical construction below, this glued coupling is the twisted matrix product* $R = P\,\mathrm{diag}(1/b)Q$*, with* $R_{i,k} = \sum_{j:b_j>0} P_{i,j}Q_{j,k}/b_j$*. In the matrix notation,* $1/b_j$ *is understood as* 0 *when* $b_j = 0$*.*

Proof.

One verifies that

$$S_{i,j,k} = \begin{cases} \frac{P_{i,j}Q_{j,k}}{b_j} & \text{if} \quad b_j \neq 0 \\ 0 \text{ otherwise} \end{cases} \tag{4.1}$$

has the required marginals. Indeed, if $b_j \neq 0$,

$$\sum_k S_{i,j,k} = \sum_k \frac{P_{i,j}Q_{j,k}}{b_j} = \frac{P_{i,j}}{b_j}(Q\mathbb{1}_m)_j = \frac{P_{i,j}}{b_j}b_j.$$

If $b_j = 0$, then necessarily $P_{i,j} = 0$ and $\sum_k S_{i,j,k} = 0 = P_{i,j}$. The same computation gives the other prescribed marginal:

$$\sum_i S_{i,j,k} = \begin{cases} \frac{Q_{j,k}}{b_j}\sum_i P_{i,j} = Q_{j,k} & \text{if} \quad b_j > 0, \\ 0 = Q_{j,k} & \text{if} \quad b_j = 0. \end{cases}$$

Summing over j then gives the displayed formula for R. Its row and column sums are $\sum_k R_{i,k} = \sum_j P_{i,j} = a_i$ and $\sum_i R_{i,k} = \sum_j Q_{j,k} = c_k$, so $R \in U(a, c)$. □

Figure 4.1 displays the two input couplings, their glued marginal and the direct optimal coupling in a common matrix format.

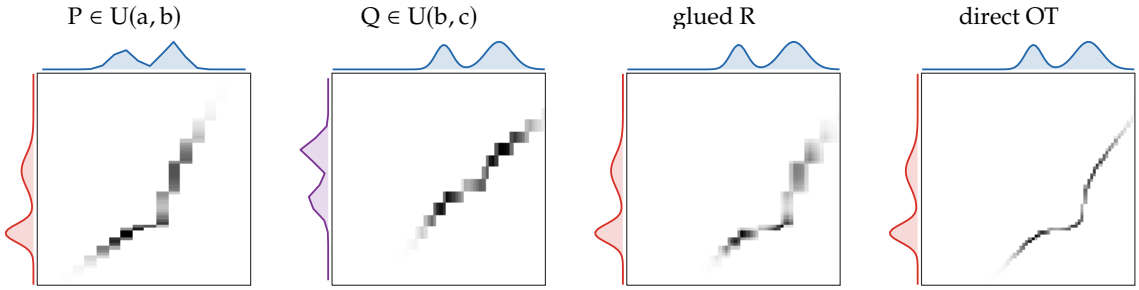


Figure 4.1: Discrete gluing lemma in matrix form. The first two panels are optimal one-dimensional couplings through an intermediate marginal b. The third panel shows the induced marginal $R = P\,\mathrm{diag}(1/b)Q$ between a and c; it is feasible and is the coupling used in the triangle-inequality proof. Because the intermediate marginal is represented on a coarser grid, the glued coupling is more mediated than the direct optimal coupling shown on the right. The thin box frames only the coupling matrix in each panel, while the attached marginal strips remain outside it.

When the cost matrix is the pth power of a distance matrix, the discrete Kantorovich value becomes a metric on histograms.

Definition 4.2 (Discrete Wasserstein distance)**.** Let $D \in \mathbb{R}_+^{n\times n}$ be a distance matrix on $[\![n]\!]$ and let $p \geqslant 1$. The discrete p-Wasserstein distance between histograms $a, b \in \Sigma_n$ is

$$W_p(a,b) := L_{D^p}(a,b)^{1/p}. \tag{4.2}$$

It depends on the chosen ground distance D.

Proposition 4.3 (Metric property of discrete Wasserstein distance)**.** *For every distance matrix* D *on* $[\![n]\!]$*, Definition 4.2 defines a distance on* Σ_n*:* W_p *is symmetric, positive,* $W_p(a,b) = 0$ *if and only if* $a = b$*, and it satisfies the triangle inequality*

$$\forall\, a, b, c \in \Sigma_n, \quad W_p(a,c) \leqslant W_p(a,b) + W_p(b,c).$$

Proof. Transposition maps U(a, b) bijectively onto U(b, a) and preserves the cost because D is symmetric. This proves symmetry. If $a = b$, the diagonal coupling $\mathrm{diag}(a)$ has zero cost. Conversely, if an admissible coupling has zero cost, every off-diagonal entry must vanish because $D_{i,j} > 0$ for $i \neq j$. The coupling is therefore diagonal, and its two marginals coincide. This proves definiteness.

For the triangle inequality, let $P \in U(a,b)$ and $Q \in U(b,c)$ be optimal. Lemma 4.1 provides $S \in \mathbb{R}_+^{n\times n\times n}$ with these two marginals. Its first-third marginal $R_{i,k} = \sum_j S_{i,j,k}$ belongs to U(a, c). The ground-metric triangle inequality and Minkowski's inequality give

$$\begin{aligned} W_p(a,c) &\leqslant \left(\sum_{i,j,k} D_{i,k}^p S_{i,j,k}\right)^{1/p} \\ &\leqslant \left(\sum_{i,j,k} (D_{i,j} + D_{j,k})^p S_{i,j,k}\right)^{1/p} \leqslant \left(\sum_{i,j,k} D_{i,j}^p S_{i,j,k}\right)^{1/p} + \left(\sum_{i,j,k} D_{j,k}^p S_{i,j,k}\right)^{1/p} \\ &= W_p(a,b) + W_p(b,c). \end{aligned}$$

□

Continuous gluing. Proposition 4.3 generalizes from histogram to arbitrary measures that need not be discrete. For this, one needs the following general gluing lemma.

Lemma 4.4 (Gluing lemma)**.** *Let* $(\alpha,\beta,\gamma) \in \mathcal{M}_+^1(\mathcal{X}) \times \mathcal{M}_+^1(\mathcal{Y}) \times \mathcal{M}_+^1(\mathcal{Z})$*, where* $\mathcal{X}$*,* $\mathcal{Y}$ *and* $\mathcal{Z}$ *are Polish spaces in the sense of Definition 2.3. Given* $\pi \in \mathcal{U}(\alpha,\beta)$ *and* $\xi \in \mathcal{U}(\beta,\gamma)$*, there exists a probability measure* $\sigma \in \mathcal{M}_+^1(\mathcal{X}\times\mathcal{Y}\times\mathcal{Z})$ *such that*

$$(P_{\mathcal{X},\mathcal{Y}})_\sharp \sigma = \pi \quad \text{and} \quad (P_{\mathcal{Y},\mathcal{Z}})_\sharp \sigma = \xi$$

where $P_{\mathcal{X},\mathcal{Y}}(x,y,z) = (x,y)$ *and* $P_{\mathcal{Y},\mathcal{Z}}(x,y,z) = (y,z)$ *are the coordinate projections.*

Proof. Disintegrate the two couplings with respect to their common marginal β as $\pi(\mathrm{d}x,\mathrm{d}y) = \pi_y(\mathrm{d}x)\beta(\mathrm{d}y)$ and $\xi(\mathrm{d}y,\mathrm{d}z) = \beta(\mathrm{d}y)\xi_y(\mathrm{d}z)$. The Polish hypotheses guarantee the existence of the probability kernels $y \mapsto \pi_y$ and $y \mapsto \xi_y$. Define their conditional product by

$$\int g \,\mathrm{d}\sigma := \int_{\mathcal{Y}} \int_{\mathcal{X}} \int_{\mathcal{Z}} g(x,y,z)\, \xi_y(\mathrm{d}z)\pi_y(\mathrm{d}x)\beta(\mathrm{d}y)$$

for every bounded Borel function g. Taking $g(x,y,z) = h(x,y)$ recovers π, while taking $g(x,y,z) = h(y,z)$ recovers ξ, proving the two marginal identities.

In the discrete setting, $\pi_{y_j} = \sum_i (\mathrm{P}_{i,j}/\mathrm{b}_j)\delta_{x_i}$ and $\xi_{y_j} = \sum_k (\mathrm{Q}_{j,k}/\mathrm{b}_j)\delta_{z_k}$ whenever $\mathrm{b}_j > 0$. Hence $\sigma = \sum_{i,j,k} \mathrm{S}_{i,j,k}\delta_{(x_i,y_j,z_k)}$, where $\mathrm{S}_{i,j,k} = \mathrm{P}_{i,j}\mathrm{Q}_{j,k}/\mathrm{b}_j$, with the value zero when $\mathrm{b}_j = 0$, exactly as in (4.1). □

Finite moments provide the natural domain for metric transport costs.

Definition 4.5 (p-Wasserstein space)**.** Let $(\mathcal{X}, d)$ be Polish, let $p \geqslant 1$, and fix $x_0 \in \mathcal{X}$. The p-Wasserstein space is

$$\mathcal{P}_p(\mathcal{X}) := \left\{ \alpha \in \mathcal{M}_+^1(\mathcal{X}) \,;\, \int d(x,x_0)^p \,\mathrm{d}\alpha(x) < +\infty \right\},$$

which is independent of the choice of x_0.

If $\alpha, \beta \in \mathcal{P}_p(\mathcal{X})$, then the product coupling has finite p-cost by the triangle inequality, so the Kantorovich value is finite. Proposition 3.34 supplies an optimal coupling because d^p is continuous and nonnegative [539, 469].

Using this gluing lemma, we can now construct the Wasserstein distance in the general setting of arbitrary distributions on a Polish space.

Definition 4.6 (Wasserstein distance)**.** Let $(\mathcal{X}, d)$ be a Polish metric space and $p \geqslant 1$. On the space $\mathcal{P}_p(\mathcal{X})$ of Definition 4.5, the p-Wasserstein distance is

$$\mathcal{W}_p(\alpha,\beta) := \mathcal{L}_{d^p}(\alpha,\beta)^{1/p} = \left(\inf_{\pi \in \mathcal{U}(\alpha,\beta)} \int_{\mathcal{X}\times\mathcal{X}} d(x,y)^p \,\mathrm{d}\pi(x,y) \right)^{1/p}. \tag{4.3}$$

It depends on the ground distance d.

The gluing lemma upgrades this optimal-transport value from a discrepancy to a genuine metric.

Proposition 4.7 (Metric property of the Wasserstein distance)**.** *Definition 4.6 defines a distance on $\mathcal{P}_p(\mathcal{X})$: $\mathcal{W}_p$ is symmetric and nonnegative, $\mathcal{W}_p(\alpha,\beta) = 0$ if and only if $\alpha = \beta$, and*

$$\forall\, (\alpha,\beta,\gamma) \in \mathcal{P}_p(\mathcal{X})^3, \qquad \mathcal{W}_p(\alpha,\gamma) \leqslant \mathcal{W}_p(\alpha,\beta) + \mathcal{W}_p(\beta,\gamma).$$

Proof. The value is finite on $\mathcal{P}_p(\mathcal{X})$ because the product coupling has finite p-cost. If $\tau(x,y) = (y,x)$ is the flip map, then $\tau_\sharp \pi \in \mathcal{U}(\beta,\alpha)$ whenever $\pi \in \mathcal{U}(\alpha,\beta)$, and the transport cost is unchanged. This proves symmetry. The diagonal coupling $(\mathrm{Id},\mathrm{Id})_\sharp\alpha$ proves $\mathcal{W}_p(\alpha,\alpha) = 0$.

Conversely, Proposition 3.34 supplies an optimal coupling π. If $\mathcal{W}_p(\alpha,\beta) = 0$, then $d(x,y) = 0$ for π-almost every (x,y), so $x = y$ almost surely under π. For every bounded Borel function f,

$$\int f \,\mathrm{d}\alpha = \int f(x)\,\mathrm{d}\pi(x,y) = \int f(y)\,\mathrm{d}\pi(x,y) = \int f \,\mathrm{d}\beta,$$

and therefore $\alpha = \beta$.

For the triangle inequality, consider optimal couplings $\pi \in \mathcal{U}(\alpha,\beta)$ and $\xi \in \mathcal{U}(\beta,\gamma)$, and glue them into σ by Lemma 4.4. Let $P_{1,3}(x,y,z) = (x,z)$ and define $\eta := (P_{1,3})_\sharp\sigma$. Then $\eta \in \mathcal{U}(\alpha,\gamma)$. If the two original couplings are induced by maps T_1 and T_2, this construction is induced by $T_2 \circ T_1$; gluing therefore

extends composition of deterministic maps. The triangle inequality and Minkowski's inequality give

$$\begin{aligned}\mathcal{W}_p(\alpha,\gamma) &\leqslant \left(\int d(x,z)^p \,\mathrm{d}\sigma(x,y,z)\right)^{1/p}\\ &\leqslant \left(\int (d(x,y)+d(y,z))^p \,\mathrm{d}\sigma(x,y,z)\right)^{1/p}\\ &\leqslant \left(\int d(x,y)^p \,\mathrm{d}\sigma(x,y,z)\right)^{1/p} + \left(\int d(y,z)^p \,\mathrm{d}\sigma(x,y,z)\right)^{1/p}\\ &= \left(\int d(x,y)^p \,\mathrm{d}\pi\right)^{1/p} + \left(\int d(y,z)^p \,\mathrm{d}\xi\right)^{1/p} = \mathcal{W}_p(\alpha,\beta)+\mathcal{W}_p(\beta,\gamma).\end{aligned}$$

□

Remark 4.8 (Gluing is the metric engine). The proof of Proposition 4.7 is the prototype for many metric constructions in the book. The only genuinely transport-specific step is the gluing lemma, Lemma 4.4: two couplings sharing a marginal can be lifted to a joint law, and the ordinary triangle inequality is then applied pointwise before integrating. The same pattern already appeared in the discrete proof of Proposition 4.3, where the glued tensor is written explicitly in (4.1). It reappears whenever a new distance is built from transport: quotient Wasserstein uses the same metric inheritance after optimizing over a group action in Proposition 11.35; conditional Wasserstein glues fiberwise couplings in Proposition 11.47 and along geodesics in Proposition 11.48; the conic construction of unbalanced OT uses gluing on cone couplings in Proposition 11.7; and the Gromov–Wasserstein metric theorem, Theorem 13.20, is the intrinsic-space analogue where couplings are glued while controlling pairwise distortions. In this sense, gluing is the metric engine behind most extensions of $\mathcal{W}_p$, not merely a technical lemma for (4.3).

The Euclidean quadratic case separates the displacement of the means from the transport of the centered shapes. For general p, the same argument still gives a lower bound by the mean displacement.

Proposition 4.9 (Euclidean mean bound and quadratic decomposition)**.** *Let $p \geqslant 1$ and $\alpha,\beta \in \mathcal{P}_p(\mathbb{R}^d)$, and denote their means by $m_\alpha := \int_{\mathbb{R}^d} x \,\mathrm{d}\alpha(x)$ and $m_\beta := \int_{\mathbb{R}^d} x \,\mathrm{d}\beta(x)$. Then $\|m_\alpha - m_\beta\| \leqslant \mathcal{W}_p(\alpha,\beta)$. For $p=2$, define the centered translates $\alpha^0 := (x \mapsto x - m_\alpha)_\sharp\alpha$ and $\beta^0 := (y \mapsto y - m_\beta)_\sharp\beta$. Then the following exact decomposition holds:*

$$\mathcal{W}_2^2(\alpha,\beta) = \|m_\alpha - m_\beta\|^2 + \mathcal{W}_2^2(\alpha^0,\beta^0).$$

Proof. For every $\pi \in \mathcal{U}(\alpha,\beta)$, $m_\alpha - m_\beta = \int_{\mathbb{R}^d\times\mathbb{R}^d}(x-y)\,\mathrm{d}\pi(x,y)$. Jensen's inequality therefore gives $\|m_\alpha - m_\beta\| \leqslant (\int \|x-y\|^p \,\mathrm{d}\pi(x,y))^{1/p}$. Taking the infimum over π proves the first claim.

For $p=2$, set $\pi^0 := ((x,y)\mapsto(x-m_\alpha, y-m_\beta))_\sharp\pi$. This translation defines a bijection from $\mathcal{U}(\alpha,\beta)$ onto $\mathcal{U}(\alpha^0,\beta^0)$. Writing $u = x - m_\alpha$ and $v = y - m_\beta$, one has

$$\int \|x-y\|^2 \,\mathrm{d}\pi = \|m_\alpha - m_\beta\|^2 + \int \|u-v\|^2 \,\mathrm{d}\pi^0(u,v),$$

because $\int (u-v)\,\mathrm{d}\pi^0(u,v) = 0$. Taking the infimum over the corresponding couplings proves the decomposition. □

Applications of Wasserstein distance. Representing structured data as probability measures turns the Wasserstein distance into a geometry-aware comparison tool. The following are two representative application domains: single-cell biology and natural-language processing.

Example 4.10 (Application to gene-expression distances between two cells). A single cell can be encoded as a measure over genes,

$$\alpha_{\text{cell}} = \sum_g e_g \delta_{\varphi(g)},$$

where $e_g \geqslant 0$, $\sum_g e_g = 1$, and $\varphi(g)$ is a gene embedding or annotation vector. Wasserstein distances then compare cells by moving expression mass between genes. The choice of $d(\varphi(g),\varphi(g'))$ is biologically meaningful: it can come from annotations, pathways or a learned ground metric. This is

the idea behind the Gene Mover's Distance and related metric-learning approaches for single-cell data [33, 282]; it is the single-cell analogue of the cost-learning question revisited in Section 12.5.

Example 4.11 (Application to word embeddings and documents). A document can similarly be viewed as the probability measure $\alpha_{\text{doc}} = \sum_{w \in \text{doc}} a_w \delta_{e_w}$ on a word-embedding space. Here a_w are normalized word frequencies and e_w are word embeddings. The Word Mover's Distance is the Wasserstein distance between such document measures [322]. It compares bags of words through the geometry learned by the embedding, so that replacing a word by a nearby synonym is less costly than replacing it by an unrelated word. When two embedding spaces are not already aligned, Gromov–Wasserstein variants compare their intrinsic neighborhood geometry instead of relying on a common coordinate system [10]; this is the intrinsic-space viewpoint developed in Section 13.3.

Interpolation induced by an optimal plan. The quadratic Wasserstein distance does not only compare two endpoint measures. An optimal plan also says how to move mass between them: each active pair (x, y) travels along the segment joining x to y. This turns an optimal coupling into a curve of measures.

Definition 4.12 ($\mathcal{W}_2$ geodesic induced by an optimal plan)**.** Let $\alpha_0, \alpha_1 \in \mathcal{P}_2(\mathbb{R}^d)$, and let $\pi^\star \in \mathcal{U}(\alpha_0, \alpha_1)$ be optimal for $\mathcal{W}_2^2(\alpha_0, \alpha_1)$. For $t \in [0, 1]$, define

$$e_t(x, y) := (1 - t)x + ty, \qquad \alpha_t := (e_t)_\sharp \pi^\star.$$

The curve $(\alpha_t)_{t \in [0,1]}$ is the displacement, or McCann, $\mathcal{W}_2$ geodesic induced by $\pi^\star$.

In the discrete case, each mass $\mathrm{P}_{i,j}$ moves from x_i to y_j along its own segment. When the optimal plan is not induced by a map, one source atom can split into several moving atoms. If the optimal plan is not unique, different optimal plans may also induce different $\mathcal{W}_2$ geodesics.

Proposition 4.13 (Optimal-plan interpolation is a $\mathcal{W}_2$ geodesic)**.** *Let $(\alpha_t)_{t \in [0,1]}$ be defined by Definition 4.12. Then, for every $0 \leqslant s \leqslant t \leqslant 1$,*

$$\mathcal{W}_2(\alpha_s, \alpha_t) = (t - s)\, \mathcal{W}_2(\alpha_0, \alpha_1).$$

Thus $t \mapsto \alpha_t$ is a constant-speed geodesic for the metric $\mathcal{W}_2$.

Proof. Push the optimal plan $\pi^\star$ forward by (e_s, e_t). This gives a coupling $\gamma_{s,t} \in \mathcal{U}(\alpha_s, \alpha_t)$, and

$$\int \|z - z'\|^2 \mathrm{d}\gamma_{s,t}(z, z') = \int \|e_t(x, y) - e_s(x, y)\|^2 \mathrm{d}\pi^\star(x, y) = (t - s)^2\, \mathcal{W}_2^2(\alpha_0, \alpha_1).$$

Hence $\mathcal{W}_2(\alpha_s, \alpha_t) \leqslant (t - s)\, \mathcal{W}_2(\alpha_0, \alpha_1)$. Applying this upper bound to the three pairs $(0, s)$, (s, t) and $(t, 1)$, and using the triangle inequality of Proposition 4.7, gives

$$\mathcal{W}_2(\alpha_0, \alpha_1) \leqslant \mathcal{W}_2(\alpha_0, \alpha_s) + \mathcal{W}_2(\alpha_s, \alpha_t) + \mathcal{W}_2(\alpha_t, \alpha_1) \leqslant \mathcal{W}_2(\alpha_0, \alpha_1).$$

All inequalities are therefore equalities, in particular the middle segment has the claimed length. □

Figure 4.2 visualizes the construction when the optimal plan splits mass, so that the intermediate measure is obtained by moving every coupled pair along its Euclidean segment.

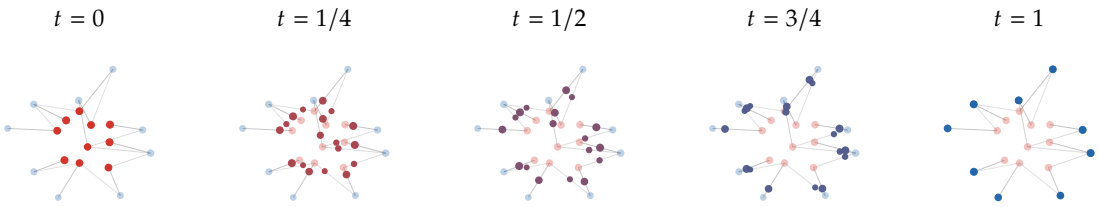


Figure 4.2: McCann interpolation induced by a non-deterministic optimal transport plan. In every panel, the red and blue endpoint measures are shown with low opacity, thin gray segments display the support $\mathrm{P}_{i,j} > \text{tol}$ of the coupling, and the moving atoms are colored from red to blue along the interpolation.

Remark 4.14 (Interpolation on a general geodesic space). For Dirac masses in Euclidean space, the $\mathcal{W}_2$ geodesic from δ_x to δ_y is $t \mapsto \delta_{(1-t)x+ty}$. The same idea extends to any geodesic metric space

$(\mathcal{X}, d)$, meaning that each pair of points can be joined by a constant-speed metric geodesic. For each pair (x, y), one replaces the Euclidean segment by a curve $\gamma^{x,y} : [0,1] \to \mathcal{X}$ such that $\gamma_0^{x,y} = x$, $\gamma_1^{x,y} = y$, and $d(\gamma_s^{x,y}, \gamma_t^{x,y}) = |t-s| d(x,y)$. If this geodesic is unique and depends measurably on (x, y), one defines $e_t(x, y) = \gamma_t^{x,y}$ and sets $\alpha_t = (e_t)_\sharp \pi^\star$ for an optimal coupling $\pi^\star$. When geodesics are not unique, there is no canonical interpolation of a pair of Diracs unless a choice is made: one may select a particular geodesic between x and y, or randomize among several such geodesics. The intrinsic formulation is to choose a probability measure η on the path space of constant-speed geodesics, called a dynamical optimal plan, such that $(e_0, e_1)_\sharp \eta$ is an optimal coupling, and to set $\alpha_t = (e_t)_\sharp \eta$. Different measurable choices, or different conditional distributions over geodesics with the same endpoints, can give different $\mathcal{W}_2$ geodesics; the constant-speed identity remains the same. This path-space viewpoint is standard in the general theory of Wasserstein spaces [12, 539, 469].

Comparison with Monge. This distance $\mathcal{W}_p$ defined through the Kantorovich problem (4.3) should be contrasted with the directed distance $\tilde{\mathcal{W}}_p$ obtained using Monge's problem (2.6). The Kantorovich feasible set is never empty, since it contains the product coupling, although the p-cost may still be infinite without moment assumptions on non-compact spaces. By contrast, Monge's constraint set $\{T \,;\, T_\sharp \alpha = \beta\}$ can be empty. When an optimal Monge map exists, Kantorovich gives the same value by choosing the graph coupling $(\mathrm{Id}, T)_\sharp \alpha$.

The next proposition makes precise one important sense in which Kantorovich is the relaxation of Monge: when the source is atomless, allowing arbitrary couplings does not change the infimum value, although it may create an optimizer when no optimal map exists.

Proposition 4.15 (Kantorovich as the plan-space relaxation of Monge)**.** *Let $(\mathcal{X}, d)$ be a compact metric space, let $p \geqslant 1$, and let $\alpha, \beta \in \mathcal{P}(\mathcal{X})$ with α atomless. Then*

$$\tilde{\mathcal{W}}_p(\alpha, \beta) = \mathcal{W}_p(\alpha, \beta).$$

This is an equality of infimal values: the Kantorovich minimum is attained, whereas the infimum defining $\tilde{\mathcal{W}}_p$ in (2.6) need not be attained by a transport map.

Proof. Let $\pi \in \mathcal{U}(\alpha, \beta)$. Choose finite Borel partitions $(A_i)_i$ and $(B_j)_j$ of $\mathcal{X}$ whose cells have diameter at most ε. Set $m_{i,j} = \pi(A_i \times B_j)$. Since α is atomless and $\sum_j m_{i,j} = \alpha(A_i)$, each A_i can be partitioned into Borel sets $A_{i,j}$ with $\alpha(A_{i,j}) = m_{i,j}$. If $m_{i,j} > 0$, then $\beta(B_j) > 0$, and Proposition 2.16 gives a measurable map $T_{i,j} : A_{i,j} \to B_j$ satisfying $(T_{i,j})_\sharp(\alpha|_{A_{i,j}}/m_{i,j}) = \beta|_{B_j}/\beta(B_j)$. Define T by $T = T_{i,j}$ on $A_{i,j}$, ignoring null cells. Then $T_\sharp \alpha = \beta$, because the contributions landing in B_j sum to $\beta|_{B_j}$, and the graph coupling $(\mathrm{Id}, T)_\sharp \alpha$ has the same masses $m_{i,j}$ as π on all rectangles $A_i \times B_j$.

Let $h \in C(\mathcal{X} \times \mathcal{X})$. By uniform continuity on the compact product, the oscillation of h on every rectangle $A_i \times B_j$ goes to zero as $\varepsilon \to 0$. Since π and $(\mathrm{Id}, T)_\sharp \alpha$ have the same mass on each rectangle, their integrals against h differ by at most this maximal oscillation. Taking partitions with mesh tending to zero therefore gives graph couplings $(\mathrm{Id}, T_k)_\sharp \alpha$ converging weakly to π. Applying this to the continuous function $(x, y) \mapsto d(x, y)^p$ gives the convergence of costs.

It remains to identify the relaxed functional. Define

$$\mathcal{G}(\alpha, \beta) := \{(\mathrm{Id}, T)_\sharp \alpha \,;\, T_\sharp \alpha = \beta\} \subset \mathcal{U}(\alpha, \beta), \qquad F_p(\pi) := \int_{\mathcal{X} \times \mathcal{X}} d(x, y)^p \, \mathrm{d}\pi(x, y),$$

and let G be the Monge graph functional which equals F_p on $\mathcal{G}(\alpha, \beta)$ and $+\infty$ otherwise. The function F_p is weakly continuous on $\mathcal{U}(\alpha, \beta)$ and satisfies $F_p \leqslant G$. If H is any weakly lower-semicontinuous function with $H \leqslant G$, then for the graph approximation above,

$$H(\pi) \leqslant \liminf_k H((\mathrm{Id}, T_k)_\sharp \alpha) \leqslant \lim_k F_p((\mathrm{Id}, T_k)_\sharp \alpha) = F_p(\pi).$$

Thus F_p is the largest weakly lower-semicontinuous minorant of G, that is, the lower-semicontinuous envelope. Minimizing this envelope over all couplings gives the Kantorovich value, and the density argument gives the same infimum when restricting to graph couplings. □

Since F_p is affine in the plan variable and $\mathcal{U}(\alpha, \beta)$ is convex, this envelope is also the closed convex relaxation of the Monge graph problem in the space of transport plans.

At the level of endpoint measures, this gives the following literal lower-semicontinuous-envelope interpretation for the Monge p-cost, whenever source measures can be regularized into atomless ones.

Corollary 4.16 (Lower-semicontinuous envelope of the Monge p-cost)**.** *Let $(\mathcal{X}, d)$ be a compact metric space and assume that atomless probability measures are dense in $\mathcal{P}(\mathcal{X})$ for the $\mathcal{W}_p$ topology. Then $(\alpha, \beta) \mapsto \mathcal{W}_p(\alpha, \beta)^p$ is the lower-semicontinuous envelope, on $\mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{X})$ for the product $\mathcal{W}_p$ topology, of the extended Monge cost*

$$\tilde{\mathcal{W}}_p(\alpha, \beta)^p = \inf_{\substack{T:\mathcal{X}\to\mathcal{X} \\ T_\sharp\alpha=\beta}} \int_{\mathcal{X}} d(x, T(x))^p \mathrm{d}\alpha(x),$$

with the convention $\tilde{\mathcal{W}}_p(\alpha, \beta)^p = +\infty$ if no admissible map exists.

Proof. For every admissible map T, the graph plan $(\mathrm{Id}, T)_\sharp\alpha$ is a coupling, hence $\mathcal{W}_p(\alpha, \beta)^p \leqslant \tilde{\mathcal{W}}_p(\alpha, \beta)^p$. Since $\mathcal{W}_p^p$ is continuous for the product $\mathcal{W}_p$ topology, it is a lower-semicontinuous minorant of the extended Monge cost.

Conversely, let H be any lower-semicontinuous minorant of the extended Monge cost. Fix (α, β) and choose atomless $\alpha_k \to \alpha$ in $\mathcal{W}_p$. By Proposition 4.15, $\tilde{\mathcal{W}}_p(\alpha_k, \beta)^p = \mathcal{W}_p(\alpha_k, \beta)^p$. Therefore

$$H(\alpha, \beta) \leqslant \liminf_k H(\alpha_k, \beta) \leqslant \lim_k \mathcal{W}_p(\alpha_k, \beta)^p = \mathcal{W}_p(\alpha, \beta)^p.$$

Thus no larger lower-semicontinuous minorant exists. □

The extra density assumption in the corollary is essential. If α has atoms, the graph-density statement can fail dramatically: for instance a single source Dirac mass cannot be mapped to two target Dirac masses. On finite spaces, the topology is discrete and this obstruction cannot be removed by closure. In such cases the Kantorovich formulation is not merely a closure of existing maps with the same marginals; it genuinely adds the possibility of splitting atomic mass.

4.2 Topology and Applications

This section shifts from metric axioms to topology and uses. Wasserstein distances metrize weak convergence under moment control, sit between weak and strong topologies, and provide quantitative estimates in probability and robust optimization.

Convergence in law topology. On a bounded metric space, all $\mathcal{W}_p$ distances define the same topology, although they are not equivalent as distances.

Proposition 4.17 (Comparison of Wasserstein distances on bounded spaces)**.** *Let $(\mathcal{X}, d)$ be a bounded Polish metric space, let $1 \leqslant p \leqslant q < \infty$, and let $\alpha, \beta \in \mathcal{P}_q(\mathcal{X})$. Then*

$$\mathcal{W}_p(\alpha, \beta) \leqslant \mathcal{W}_q(\alpha, \beta) \leqslant \mathrm{diam}(\mathcal{X})^{\frac{q-p}{q}} \mathcal{W}_p(\alpha, \beta)^{\frac{p}{q}}, \qquad \mathrm{diam}(\mathcal{X}) := \sup_{x,y\in\mathcal{X}} d(x, y).$$

Proof. The left inequality follows from Jensen's inequality, applied to any coupling π and to the convex function $r \mapsto r^{q/p}$:

$$\left(\int d(x, y)^p \mathrm{d}\pi(x, y)\right)^{\frac{q}{p}} \leqslant \int d(x, y)^q \mathrm{d}\pi(x, y).$$

Taking the infimum gives $\mathcal{W}_p \leqslant \mathcal{W}_q$. The right inequality follows from $d(x, y)^q \leqslant \mathrm{diam}(\mathcal{X})^{q-p} d(x, y)^p$ and taking the infimum over couplings once more. □

The Wasserstein distance $\mathcal{W}_p$ is a weak distance: it compares singular distributions, such as discrete measures, and quantifies spatial shifts between supports. Its topology is studied in detail in [539, 469, 11], while empirical rates are quantified in [192, 220, 548, 71, 72].

Definition 4.18 (Weak, or narrow, topology)**.** A sequence $(\alpha_k)_k$ converges weakly, or narrowly, to α in $\mathcal{M}_+^1(\mathcal{X})$, denoted $\alpha_k \rightharpoonup \alpha$, if and only if

$$\int_{\mathcal{X}} f \mathrm{d}\alpha_k \longrightarrow \int_{\mathcal{X}} f \mathrm{d}\alpha \qquad \text{for every } f \in C_b(\mathcal{X}).$$

On a compact space, $C_b(\mathcal{X}) = C(\mathcal{X})$ and this is precisely weak* convergence in the duality between measures and continuous functions. The term weak* is therefore also used informally in the book

when no confusion can arise.

Remark 4.19 (A Riemann-sum weak limit). On $\mathcal{X} = \mathbb{R}$, the empirical measures on a regular grid satisfy $n^{-1}\sum_{k=1}^{n}\delta_{k/n} \rightharpoonup \mathcal{U}_{[0,1]}$. Indeed, for every continuous bounded function f, $n^{-1}\sum_{k=1}^{n} f(k/n) \to \int_0^1 f(x)\mathrm{d}x$, which is precisely the convergence of Riemann sums. This convergence is weak but not strong: for every n, the discrete measure and the uniform density are mutually singular, hence their total variation distance is equal to 2.

Remark 4.20 (Weak convergence for discrete measures). In the special case of a single Dirac, $\delta_{x^{(n)}} \rightharpoonup \delta_x$ is equivalent to $\int f\mathrm{d}\delta_{x^{(n)}} = f(x^{(n)}) \to \int f\mathrm{d}\delta_x = f(x)$ for any continuous f. This in turn is equivalent to $x^{(n)} \to x$. For a fixed number of atoms, if $\alpha_n = \sum_{i=1}^{N} a_i^{(n)}\delta_{x_i^{(n)}}$ and, after extracting a subsequence and relabeling, $a_i^{(n)} \to a_i$ and $x_i^{(n)} \to x_i$, then α_n converges weakly to $\sum_i a_i\delta_{x_i}$, with atoms at identical limits merged. Without a uniform bound on the number of atoms, weak limits of discrete measures can be non-discrete; empirical measures are the standard example.

Strong versus weak topology. The total variation norm was introduced in Definition 2.7 and Proposition 2.8. Its induced topology is often called the "strong" topology on measures. In the present section we only use the recall that, for a signed difference $\alpha - \beta$,

$$\|\alpha - \beta\|_{\mathrm{TV}} = |\alpha - \beta|(\mathcal{X}),$$

so densities give an L^1 norm and discrete signed measures give an ℓ^1 norm of the signed weights.

The following proposition shows that the TV norm can be seen as a Wasserstein distance, but for a "degenerate" $0/1$ metric.

Proposition 4.21 (Total variation as Wasserstein for the discrete metric)**.** *Let $\mathcal{X}$ be a standard Borel space, let α and β be probability measures on $\mathcal{X}$, and equip $\mathcal{X}$ with the $0/1$ cost $d_0(x,x) = 0$ and $d_0(x,y) = 1$ for $x \neq y$. Then, for every $p \geqslant 1$,*

$$\inf_{\pi\in\mathcal{U}(\alpha,\beta)} \int d_0(x,y)^p \mathrm{d}\pi(x,y) = \frac{1}{2}\|\alpha - \beta\|_{\mathrm{TV}}.$$

Whenever the $0/1$ metric is used as the ground metric, the left-hand side is $\mathcal{W}_p(\alpha,\beta)^p$.

Proof. Let $\lambda = \alpha + \beta$, write $a = \mathrm{d}\alpha/\mathrm{d}\lambda$ and $b = \mathrm{d}\beta/\mathrm{d}\lambda$, and define the common part η by $\frac{\mathrm{d}\eta}{\mathrm{d}\lambda} = \min(a,b)$. Its residuals $\alpha' = \alpha - \eta$ and $\beta' = \beta - \eta$ are mutually singular and have the same mass

$$r = 1 - \eta(\mathcal{X}) = \frac{1}{2}\int |a - b|\mathrm{d}\lambda = \frac{1}{2}\|\alpha - \beta\|_{\mathrm{TV}}.$$

For any $\pi \in \mathcal{U}(\alpha,\beta)$, define its diagonal submeasure by $\zeta(A) = \pi\big(\{(x,x) : x \in A\}\big)$. Then $\zeta \leqslant \alpha$ and $\zeta \leqslant \beta$, so its density with respect to λ is bounded by $\min(a,b)$. Consequently $\zeta \leqslant \eta$, hence $\pi(\{x = y\}) \leqslant \eta(\mathcal{X})$, and $\int d_0(x,y)^p\mathrm{d}\pi(x,y) = 1 - \pi(\{x = y\}) \geqslant r$. If $r = 0$, the diagonal coupling attains this bound. If $r > 0$, let $\Delta(x) = (x,x)$ and set $\pi^\star = \Delta_\sharp\eta + r^{-1}\alpha' \otimes \beta'$. This is a coupling of α and β. Since $\alpha' \perp \beta'$, the product term is concentrated off the diagonal, so its $0/1$ cost is exactly its total mass r. Thus $\pi^\star$ attains the lower bound. □

Probabilistic interpretation. Probability theory gives weak convergence its most direct interpretation: it compares distributions rather than samplewise realizations. In terms of random vectors, if $X_n \sim \alpha_n$ and $X \sim \alpha$ (not necessarily defined on the same probability space), then $\alpha_n \rightharpoonup \alpha$ means precisely that X_n converges in law to X.

Convergence in law should be distinguished from stronger notions that compare random variables on a common probability space. In that setting, $X_n \to X$ almost surely means pointwise convergence outside a null set, while convergence in probability means

$$\forall\,\varepsilon > 0, \qquad \mathbb{P}(\|X_n - X\| > \varepsilon) \to 0.$$

Almost-sure convergence implies convergence in probability, which in turn implies convergence in law. The last notion depends only on the marginal laws, since it is exactly the weak convergence $(X_n)_\sharp\mathbb{P} \rightharpoonup X_\sharp\mathbb{P}$, and therefore does not require a common probability space.

Convergence in law should also be distinguished from strong convergence of measures. Total variation convergence controls the mass assigned to every measurable set, not only averages against continuous test functions, and therefore implies weak convergence. The converse fails, notably for empirical approximations of continuous laws.

Central limit theorem and OT. Limit theorems produce canonical convergent sequences of probability measures, and Wasserstein distances turn their qualitative conclusions into metric error bounds. The central limit theorem concerns sums of independent random variables; because sums of independent variables correspond to convolutions of their laws, we first introduce this operation.

Definition 4.22 (Convolution of measures)**.** For probability measures α, β on $\mathbb{R}^d$, their convolution is

$$\alpha * \beta := \mathrm{add}_\sharp(\alpha \otimes \beta), \qquad \mathrm{add}(x, y) = x + y.$$

Equivalently, for every bounded continuous φ,

$$\int \varphi \, \mathrm{d}(\alpha * \beta) = \iint \varphi(x + y) \, \mathrm{d}\alpha(x) \mathrm{d}\beta(y).$$

Thus $\alpha * \beta$ is the law of $X + Y$ when X and Y are independent with laws α and β. When $\alpha = \rho_\alpha \mathrm{d}x$ and $\beta = \rho_\beta \mathrm{d}x$, their convolution has density

$$\rho_{\alpha*\beta}(z) = (\rho_\alpha * \rho_\beta)(z) = \int_{\mathbb{R}^d} \rho_\alpha(x) \rho_\beta(z - x) \mathrm{d}x.$$

Remark 4.23 (Central limit theorem). The central limit theorem states that if $(X_i)_{i \geqslant 1}$ are i.i.d. random vectors with finite second moments, $\mathbb{E}(X_i) = 0$, and $\mathbb{E}(X_i X_i^\top) = \mathrm{Id}$, then the normalized sum

$$Z_n := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i$$

converges in law toward the standard Gaussian $\mathcal{N}(0, \mathrm{Id})$. In the terminology recalled above, this means that the measure α_n representing the law of Z_n converges weakly toward the centered Gaussian measure $\alpha = \mathcal{N}(0, \mathrm{Id})$.
Equivalently, this is a statement about the convolution of Definition 4.22. If α is the common law of the variables X_i, writing α^{*n} for the n-fold convolution of α with itself, and denoting by $D_a(x) = ax$ the dilation map, the law of Z_n is

$$\alpha_n = (D_{1/\sqrt{n}})_\sharp \alpha^{*n}.$$

The CLT therefore says that the normalized n-fold convolution $(D_{1/\sqrt{n}})_\sharp \alpha^{*n}$ converges weakly to the Gaussian $\mathcal{N}(0, \mathrm{Id})$.

Figure 4.3 makes this qualitative weak convergence visible in the elementary Bernoulli case: every finite-n law is discrete, yet the normalized atom heights approach the Gaussian density.

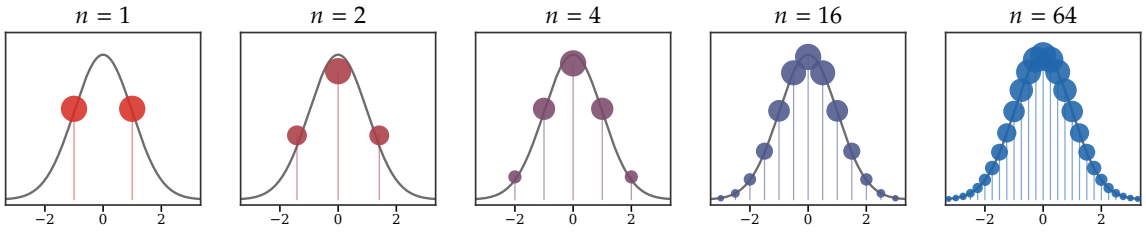


Figure 4.3: Central-limit theorem for normalized Bernoulli sums. Starting from $\alpha_0 = \frac{1}{2}(\delta_{-1} + \delta_1)$, the law of $Z_n = n^{-1/2} \sum_{i=1}^n X_i$ remains discrete, but its rescaled atom heights approach the standard Gaussian density shown in gray. Proposition 10.2 later quantifies this weak convergence in $\mathcal{W}_1$.

Remark 4.24 (A quantitative CLT in Wasserstein form). The metric viewpoint on weak convergence is not only topological. It also turns some classical limit theorems into quantitative metric estimates. For instance, the Berry–Esseen theorem can be stated as a bound on the $\mathcal{W}_1$ distance between the law of the normalized sum $n^{-1/2} \sum_i X_i$ and the limiting Gaussian. By Kantorovich–Rubinstein duality, this is exactly a uniform control of the CLT error over all 1-Lipschitz test functions. This application is developed later in Section 10.1, see in particular Proposition 10.2.

Wasserstein metrizes weak convergence. As explained in Remark 4.20, in the special case of Diracs, $\delta_{x_n} \rightharpoonup \delta_x$ is equivalent to $x_n \to x$. One can then contrast the strong topology with the Wasserstein distance if $x_n \neq x$,

$$\|\delta_{x_n} - \delta_x\|_{\mathrm{TV}} = 2 \quad \text{and} \quad \mathcal{W}_p(\delta_{x_n}, \delta_x) = d(x_n, x).$$

Thus a moving Dirac sequence can converge in total variation only if it is eventually stationary, whereas it converges in Wasserstein distance whenever its locations converge. On an unbounded space, weak convergence alone does not prevent a vanishing amount of mass from escaping to infinity. Wasserstein convergence is exactly weak convergence supplemented by control of this moment tail.

Proposition 4.25 (Wasserstein convergence on Polish spaces)**.** *Let $(\mathcal{X}, d)$ be a Polish metric space, let $1 \leqslant p < +\infty$, and let $\alpha_k, \alpha \in \mathcal{P}_p(\mathcal{X})$. For any reference point $x_0 \in \mathcal{X}$, the following are equivalent:*

1. *$\mathcal{W}_p(\alpha_k, \alpha) \to 0$;*
2. *$\alpha_k \rightharpoonup \alpha$ and the pth moments converge,*

$$\int d(x, x_0)^p \, \mathrm{d}\alpha_k(x) \longrightarrow \int d(x, x_0)^p \, \mathrm{d}\alpha(x);$$

3. *$\alpha_k \rightharpoonup \alpha$ and the pth moments are uniformly integrable,*

$$\lim_{R \to +\infty} \sup_k \int_{\{d(x,x_0) > R\}} d(x, x_0)^p \, \mathrm{d}\alpha_k(x) = 0.$$

These conditions do not depend on the choice of x_0.

Proof. Assume first that $\mathcal{W}_p(\alpha_k, \alpha) \to 0$. Since $\mathcal{W}_1 \leqslant \mathcal{W}_p$, integration against every bounded 1-Lipschitz function converges; such functions determine weak convergence on Polish spaces. For any coupling $\pi \in \mathcal{U}(\alpha_k, \alpha)$, the reverse triangle inequality in $L^p(\pi)$ gives

$$\left| \left(\int d(x, x_0)^p \, \mathrm{d}\alpha_k(x) \right)^{1/p} - \left(\int d(y, x_0)^p \, \mathrm{d}\alpha(y) \right)^{1/p} \right| \leqslant \left(\int d(x, y)^p \, \mathrm{d}\pi(x, y) \right)^{1/p}.$$

Taking the infimum over π proves convergence of the pth moments, hence (ii).

Conversely, assume (ii). By the Skorokhod representation theorem, there exist random variables $X_k \sim \alpha_k$ and $X \sim \alpha$ on a common probability space such that $X_k \to X$ almost surely. Consequently $d(X_k, x_0)^p \to d(X, x_0)^p$ almost surely, and the assumed convergence of their expectations implies, by Scheffé's lemma, convergence in L^1. In particular, the family $(d(X_k, x_0)^p)_k$ is uniformly integrable. Since

$$d(X_k, X)^p \leqslant 2^{p-1} \big(d(X_k, x_0)^p + d(X, x_0)^p \big),$$

the variables $d(X_k, X)^p$ are uniformly integrable and converge almost surely to zero. Vitali's theorem therefore gives $\mathbb{E}[d(X_k, X)^p] \to 0$. The law of (X_k, X) is an admissible coupling, so

$$\mathcal{W}_p(\alpha_k, \alpha)^p \leqslant \mathbb{E}[d(X_k, X)^p] \longrightarrow 0.$$

It remains to compare (ii) and (iii). The preceding Skorokhod argument shows that (ii) makes $d(X_k, x_0)^p$ uniformly integrable, which is exactly the tail condition in (iii). Conversely, assume (iii) and define the bounded continuous truncation $h_R(x) = \min\{d(x, x_0)^p, R^p\}$. Weak convergence gives $\int h_R \mathrm{d}\alpha_k \to \int h_R \mathrm{d}\alpha$ for every R. The uniform tail bound in (iii), together with the integrability of $d(\cdot, x_0)^p$ under α, then lets $R \to +\infty$ and proves convergence of the untruncated moments. This establishes all three equivalences. Since (i) is independent of the reference point, so are (ii) and (iii) [539, 469, 12]. □

On a compact metric space, the moment function $d(\cdot, x_0)^p$ is bounded and continuous, so weak convergence automatically implies moment convergence. Proposition 4.25 therefore recovers the familiar fact that $\mathcal{W}_p$ metrizes weak convergence on $\mathcal{P}(\mathcal{X})$.

On a finite metric space, the strong and weak topologies coincide. The following quantitative comparison explains this equivalence.

Proposition 4.26 (Comparison with total variation on finite spaces)**.** *Let $(\mathcal{X}, d)$ be a finite metric space and define*

$$d_{\min} := \min_{x \neq y} d(x,y) > 0, \qquad d_{\max} := \max_{x,y} d(x,y) < +\infty.$$

Then, for every $\alpha, \beta \in \mathcal{P}(\mathcal{X})$,

$$\frac{d_{\min}}{2}\|\alpha - \beta\|_{\mathrm{TV}} \leqslant \mathcal{W}_1(\alpha,\beta) \leqslant \frac{d_{\max}}{2}\|\alpha - \beta\|_{\mathrm{TV}}.$$

Proof. We denote $d_0(x,y)$ the distance such that $d_0(x,x) = 0$ and $d_0(x,y) = 1$ for $x \neq y$. One has $d_{\min} d_0(x,y) \leqslant d(x,y) \leqslant d_{\max} d_0(x,y)$, so that integrating this against any $\pi \in \mathcal{U}(\alpha,\beta)$ and taking the minimum among those π gives the result using Proposition 4.21. □

The constants are optimal: for $\alpha = \delta_x$ and $\beta = \delta_y$ with $x \neq y$, the comparison reduces to

$$d_{\min} \leqslant d(x,y) \leqslant d_{\max}.$$

Choosing pairs that attain $d_{\min}$ or $d_{\max}$ gives equality in the corresponding bound. The comparison deteriorates as $d_{\max}/d_{\min}$ grows. On spaces with arbitrarily close distinct points, $d_{\min} = 0$ and no such two-sided equivalence is possible, consistently with the difference between strong and weak convergence.

4.3 Distributional Robustness and $\mathcal{W}_\infty$

The preceding sections developed $\mathcal{W}_p$ for $1 \leqslant p < \infty$, where transport displacements are averaged through an L^p cost. We now study the endpoint $p = \infty$, which controls the largest displacement and arises naturally in distributional robustness. We first recall Wasserstein ambiguity sets and their dual formulation, then contrast their average-budget geometry with the worst-displacement geometry of $\mathcal{W}_\infty$.

DRO ambiguity sets. Wasserstein distances are also used to define ambiguity sets around an empirical law. Given samples z_i and $\widehat{\alpha}_n = \frac{1}{n}\sum_i \delta_{z_i}$, a distributionally robust optimization (DRO) problem replaces the empirical risk $\frac{1}{n}\sum_i \ell_\theta(z_i)$ by

$$\sup_{\beta : \mathcal{W}_p(\beta, \widehat{\alpha}_n) \leqslant \rho} \int \ell_\theta(z) \mathrm{d}\beta(z),$$

or, in Lagrangian penalized form, by $\sup_\beta \int \ell_\theta \mathrm{d}\beta - \lambda\, \mathcal{W}_p(\beta, \widehat{\alpha}_n)^p$. The constrained and penalized formulations are linked by the choice of multiplier λ, but are not the same problem for an arbitrary fixed λ. Both ask for performance against distributions that can be reached by transporting the empirical mass within a budget. The radius ρ is expressed in the geometry of the data space, so it can encode feature perturbations, domain shift or model misspecification [386, 63, 233].

The basic computational reason for the popularity of Wasserstein DRO is a dual reformulation. Under the usual upper-semicontinuity and growth assumptions on the loss, one has

$$\sup_{\beta : \mathcal{W}_p(\beta, \widehat{\alpha}_n)^p \leqslant \rho^p} \int \ell_\theta \mathrm{d}\beta = \inf_{\lambda \geqslant 0} \lambda \rho^p + \frac{1}{n}\sum_{i=1}^n \sup_z \big\{ \ell_\theta(z) - \lambda d(z, z_i)^p \big\}. \tag{4.4}$$

Thus the robust risk is an empirical risk in which each sample is replaced by its worst penalized perturbation. For $p = 1$ and an L_θ-Lipschitz loss, the Kantorovich–Rubinstein dual gives the transparent upper bound

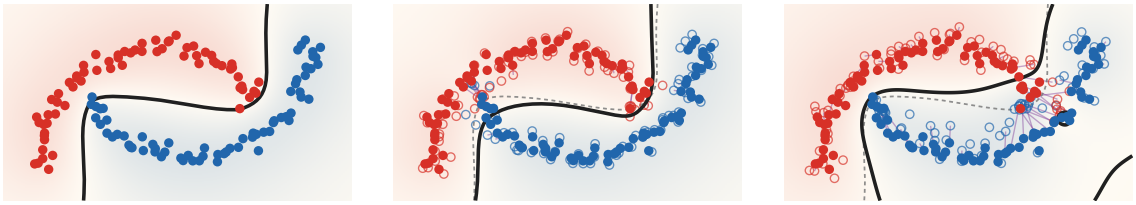
$$\sup_{\beta : \mathcal{W}_1(\beta, \widehat{\alpha}_n) \leqslant \rho} \int \ell_\theta \mathrm{d}\beta \leqslant \frac{1}{n}\sum_i \ell_\theta(z_i) + \rho L_\theta,$$

which exhibits Wasserstein robustness as a Lipschitz regularizer. The bound is sharp for worst-case optimization over a full Lipschitz ball of losses, while a fixed loss may not realize all extremal Lipschitz directions.

In machine learning, the inner supremum in (4.4) has the form of an adversarial perturbation problem: the adversary moves each datum away from z_i but pays a transport penalty. This connects Wasserstein DRO to adversarial training and certified robustness [489, 485]. In sequential decision problems, the

same idea places Wasserstein ambiguity sets around transition kernels or rewards, producing robust Bellman operators and robust reinforcement-learning models [555, 556].

Figure 4.4 gives the corresponding geometric picture for nonlinear binary classification. The score belongs to the reproducing-kernel Hilbert space of the Laplacian kernel $k_h(x,z) = e^{-\|x-z\|/h}$. The two classes form noisy interlocking crescents whose opposing tips overlap locally. The nominal boundary winds through the narrow channel between the moons; increasing the Wasserstein radius moves samples toward high-loss regions and visibly widens and reorganizes this fragile separator.



$\rho = 0$ $\rho = 0.055$ $\rho = 0.11$

Figure 4.4: Wasserstein robustness reshapes the separator between two noisy interlocking moons. Filled dots are the observed classes, the black curve is the learned zero-score boundary, and the background encodes the logistic probability. The opposing tips of the red and blue crescents overlap locally, forcing the nominal classifier to wind through a narrow nonlinear channel. In the robust panels, the dashed gray curve recalls this nominal boundary; hollow dots show all adversarial locations, while violet segments display a representative half of their transports. As ρ increases, the separator moves away from the most fragile overlap regions and trades clean fit for wider margins. The root-mean-square displacement equals the global quadratic Wasserstein budget ρ.

Applied to $c = d^p$, Proposition 3.32 shows that $(\alpha,\beta) \mapsto \mathcal{W}_p(\alpha,\beta)^p$ is jointly convex. Thus $\mathcal{W}_1$ itself is jointly convex, but $\mathcal{W}_p$ need not be convex for $p > 1$: on the real line,

$$\mathcal{W}_p\left((1-t)\delta_0 + t\delta_1, \delta_0\right) = t^{1/p},$$

which is strictly concave for $p > 1$.

This convexity is useful when distributions themselves are decision variables. In the usual DRO problem, however, the ambiguity set is fixed once the data are fixed. Therefore, if $z \mapsto \ell_\theta(z)$ is measurable and $\theta \mapsto \ell_\theta(z)$ is convex for every z, then

$$\theta \mapsto \sup_{\beta:\, \mathcal{W}_p(\beta,\widehat{\alpha}_n)\leqslant\rho} \int \ell_\theta \, \mathrm{d}\beta$$

is convex as a supremum of convex functions of θ. This explains why Wasserstein DRO can preserve convex learning formulations while still modeling adversarial distributional shifts.

$\mathcal{W}_\infty$ robustness. Worst-case displacement requires an essential-supremum counterpart of the finite-p Wasserstein distances.

Definition 4.27 ($\mathcal{W}_\infty$ distance)**.** For probability measures α, β on a metric space $(\mathcal{X}, d)$, define

$$\mathcal{W}_\infty(\alpha,\beta) := \inf_{\pi\in\mathcal{U}(\alpha,\beta)} \operatorname*{ess\,sup}_{(x,y)\sim\pi} d(x,y) \tag{4.5}$$

with value $+\infty$ if no finite-displacement coupling exists. Equivalently,

$$\mathcal{W}_\infty(\alpha,\beta) \leqslant r \iff \exists\, \pi \in \mathcal{U}(\alpha,\beta) \text{ supported on } \{(x,y) : d(x,y) \leqslant r\}.$$

This is the limit of $\mathcal{W}_p(\alpha,\beta)$ as $p \to \infty$ on bounded spaces, not the limit of the linear objectives defining $\mathcal{W}_p^p$. It minimizes the worst displacement rather than an average displacement. Although the essential-supremum objective is not linear, Definition 4.27 shows that each sublevel test is a convex feasibility problem. Thus $\mathcal{W}_\infty$ can be approached by threshold search over support-constrained coupling problems. It remains less convenient than a single linear OT problem, but it is natural in robust formulations where every transported sample must stay within a prescribed perturbation radius.

Proposition 4.28 ($\mathcal{W}_\infty$ robust envelope around an empirical law)**.** *Let $(\mathcal{Z}, d)$ be a Polish metric space. Let $\widehat{\alpha} = \sum_{i=1}^n a_i \delta_{z_i}$ with distinct z_i, positive a_i, and $\sum_i a_i = 1$, and assume that the closed balls $\overline{B}(z_i, \rho)$ are compact. Repeated atoms can always be merged to meet this convention. For any real-valued upper-semicontinuous loss ℓ,*

$$\sup_{\beta:\mathcal{W}_\infty(\beta,\widehat{\alpha})\leqslant\rho} \int \ell(z)\mathrm{d}\beta(z) = \sum_{i=1}^n a_i \sup_{z\in\overline{B}(z_i,\rho)} \ell(z).$$

Proof. If $\mathcal{W}_\infty(\beta, \widehat{\alpha}) \leqslant \rho$, then, by symmetry, there are couplings $\pi_m \in \mathcal{U}(\widehat{\alpha}, \beta)$ whose essential displacements are at most $\rho + 1/m$. Since the two marginals are fixed probability measures on a Polish space, $\mathcal{U}(\widehat{\alpha}, \beta)$ is tight and closed, hence weakly compact by Prokhorov's theorem. After extraction, $\pi_m \rightharpoonup \pi \in \mathcal{U}(\widehat{\alpha}, \beta)$. For every $\eta > 0$, the closed set $F_\eta = \{(x, z) : d(x, z) \leqslant \rho + \eta\}$ has $\pi_m(F_\eta) = 1$ for all sufficiently large m. Portmanteau's theorem gives $\pi(F_\eta) = 1$. Letting $\eta \downarrow 0$ along a countable sequence gives $\pi(\{(x, z) : d(x, z) \leqslant \rho\}) = 1$. Disintegrating π with respect to the first marginal gives $\pi = \sum_i a_i \delta_{z_i} \otimes \nu_i$, where each ν_i is supported in the closed ball $\overline{B}(z_i, \rho)$ and $\beta = \sum_i a_i \nu_i$. Hence

$$\int \ell \,\mathrm{d}\beta = \sum_i a_i \int \ell \,\mathrm{d}\nu_i \leqslant \sum_i a_i \sup_{\overline{B}(z_i,\rho)} \ell.$$

The reverse inequality follows by choosing, for each i, a maximizer $z_i^\star \in \overline{B}(z_i, \rho)$ and setting $\beta = \sum_i a_i \delta_{z_i^\star}$. The coupling $\sum_i a_i \delta_{(z_i, z_i^\star)}$ has essential displacement at most ρ, so this β is feasible and attains the displayed value. □

4.4 Measure-to-Vector Maps on Wasserstein Space

Many losses, statistics and observables depend on a probability law rather than on a single point. This section formalizes such measure-to-vector maps, introduces their natural metric regularity and shows how finite particle samples generate constructive polynomial approximations. Chapter 15 returns to scalar functionals as energies to be minimized over Wasserstein space.

Functionals and Wasserstein–Hölder regularity. A scalar functional is a map

$$f : \mathcal{P}_p(\mathcal{X}) \longrightarrow \mathbb{R}.$$

Because the approximation results below concern uniform convergence, we equip bounded functionals with the uniform norm.

Definition 4.29 (Uniform norm on functionals)**.** For a bounded functional $f : \mathcal{P}_p(\mathcal{X}) \to \mathbb{R}$, define

$$\|f\|_\infty := \sup_{\alpha\in\mathcal{P}_p(\mathcal{X})} |f(\alpha)|.$$

There is no loss of principle in starting with scalar values. If $F = (f_1, \dots, f_s) : \mathcal{P}_p(\mathcal{X}) \to \mathbb{R}^s$, the definitions and constructions below apply coordinatewise, and the resulting bounds combine in any norm on $\mathbb{R}^s$. For the quantitative approximation results, assume from now on that $\mathcal{X} \subset \mathbb{R}^d$ is compact; in particular, $\mathcal{P}_p(\mathcal{X}) = \mathcal{P}(\mathcal{X})$.

Definition 4.30 (Wasserstein–Hölder functional)**.** Let $p \geqslant 1$ and $0 < \eta \leqslant 1$. A functional $f : \mathcal{P}(\mathcal{X}) \to \mathbb{R}$ is η-Hölder for $\mathcal{W}_p$, with constant L, if

$$|f(\alpha) - f(\beta)| \leqslant L\,\mathcal{W}_p(\alpha, \beta)^\eta \qquad \text{for all } \alpha, \beta \in \mathcal{P}(\mathcal{X}). \tag{4.6}$$

The endpoint $\eta = 1$ is Wasserstein–Lipschitz regularity. It is the natural stability assumption for a functional: perturbing the input law by a small transport cost changes its output by at most a proportional amount. The restriction $\eta \leqslant 1$ is intrinsic on a geodesic space. Indeed, subdividing a constant-speed geodesic into N pieces shows that a globally Hölder functional with exponent $\eta > 1$ is constant along that geodesic, hence constant when every pair of points can be joined by one.

Examples and non-examples. Basic OT quantities already provide useful examples. For a fixed $\beta \in \mathcal{P}(\mathcal{X})$, the triangle inequality shows that $\alpha \mapsto \mathcal{W}_p(\alpha, \beta)$ is 1-Lipschitz. More generally, $\alpha \mapsto \mathcal{W}_p(\alpha, \beta)^\eta$ is η-Hölder with constant one, since $|r^\eta - s^\eta| \leqslant |r - s|^\eta$ for $r, s \geqslant 0$. If $V : \mathcal{X} \to \mathbb{R}$ is L-Lipschitz, then Kantorovich–Rubinstein duality (6.21) and the comparison $\mathcal{W}_1 \leqslant \mathcal{W}_p$ from Proposition 4.17 give

$$f(\alpha) = \int_{\mathcal{X}} V \, \mathrm{d}\alpha, \qquad |f(\alpha) - f(\beta)| \leqslant L \, \mathcal{W}_p(\alpha, \beta).$$

Pointwise mass is a simple non-example. If x_0 is a non-isolated point of $\mathcal{X}$, choose $x_n \to x_0$ with $x_n \neq x_0$. Although $\delta_{x_n} \to \delta_{x_0}$ in every $\mathcal{W}_p$, the functional $\alpha \mapsto \alpha(\{x_0\})$ jumps from zero to one. Relative entropy fails even more sharply on the full Wasserstein space. If β is nonatomic and $\widehat{\beta}_n$ is its empirical measure, then $\widehat{\beta}_n \to \beta$ in $\mathcal{W}_p$ almost surely, whereas $\mathrm{KL}(\widehat{\beta}_n | \beta) = +\infty$. Density-based functionals therefore require additional regularity assumptions or smoothing before one can expect Wasserstein–Hölder control.

Interaction polynomials. Finite-body interactions provide the natural polynomial functions of a measure.

Definition 4.31 (Wasserstein polynomial)**.** For an integer $m \geqslant 1$, a functional on $\mathcal{P}(\mathcal{X})$ is a *Wasserstein polynomial of order at most m* if it can be written, for some bounded continuous symmetric kernel $\varphi_m : \mathcal{X}^m \to \mathbb{R}$, as

$$\mathsf{P}_{\varphi_m}(\alpha) \coloneqq \int_{\mathcal{X}^m} \varphi_m(x_1, \dots, x_m) \, \mathrm{d}\alpha(x_1) \cdots \mathrm{d}\alpha(x_m). \tag{4.7}$$

Any kernel can be symmetrized without changing this functional. The order counts the number of particles jointly inspected by φ_m; it is unrelated to an ordinary spatial polynomial degree. Order one recovers linear observables $\int V \, \mathrm{d}\alpha$. Order two includes pairwise interaction energies and, after expansion around a fixed target, the squared MMD of Definition 7.9.

Empirical particle approximation. A natural first question is whether Wasserstein polynomials can approximate an arbitrary continuous functional of a probability measure. On a compact ground space, the answer follows from the Stone–Weierstrass theorem.

Proposition 4.32 (Density of Wasserstein polynomials)**.** *Let $\mathcal{X}$ be a compact metric space and $p \geqslant 1$. For every $\mathcal{W}_p$-continuous functional $f : \mathcal{P}(\mathcal{X}) \to \mathbb{R}$ and every $\varepsilon > 0$, there exists a Wasserstein polynomial P such that*

$$\|f - \mathsf{P}\|_\infty = \sup_{\alpha \in \mathcal{P}(\mathcal{X})} |f(\alpha) - \mathsf{P}(\alpha)| < \varepsilon.$$

Proof. Since $\mathcal{X}$ is compact, $\mathcal{P}(\mathcal{X})$ is compact for the weak topology, which coincides with the topology induced by $\mathcal{W}_p$. Wasserstein polynomials are continuous on this space. They contain the constants and are stable under linear combinations, after lifting lower-order kernels by adding dummy variables. They are also stable under products: if P_φ and P_ψ have orders m and r, then

$$\mathsf{P}_\varphi(\alpha) \mathsf{P}_\psi(\alpha) = \int_{\mathcal{X}^{m+r}} \varphi(x_1, \dots, x_m) \psi(x_{m+1}, \dots, x_{m+r}) \, \mathrm{d}\alpha^{\otimes(m+r)},$$

which is a Wasserstein polynomial after symmetrizing its kernel. This algebra separates points: if $\alpha \neq \beta$, there exists $V \in C(\mathcal{X})$ such that $\int V \, \mathrm{d}\alpha \neq \int V \, \mathrm{d}\beta$, and $\alpha \mapsto \int V \, \mathrm{d}\alpha$ is an order-one polynomial. The real Stone–Weierstrass theorem therefore gives the claimed uniform density in $C(\mathcal{P}(\mathcal{X}))$. □

This result is qualitative: it neither constructs the approximating polynomial nor controls the order required to reach a prescribed error. Wasserstein–Hölder regularity makes both aspects quantitative. The simple recipe is to evaluate f on an empirical measure and average over the sampled particles.

Definition 4.33 (Particle polynomial)**.** Let $f : \mathcal{P}(\mathcal{X}) \to \mathbb{R}$ be bounded and continuous and let $n \geqslant 1$. Its symmetric n-particle kernel is

$$\varphi_n^f(x_1, \dots, x_n) \coloneqq f\left(\frac{1}{n} \sum_{i=1}^n \delta_{x_i}\right),$$

and its associated *particle polynomial* is

$$\mathsf{B}_n f(\alpha) := \int_{\mathcal{X}^n} \varphi_n^f(x_1, \dots, x_n)\, \mathrm{d}\alpha^{\otimes n}(x_1, \dots, x_n) = \mathbb{E} f(\widehat{\alpha}_n), \tag{4.8}$$

where $\widehat{\alpha}_n = n^{-1} \sum_{i=1}^n \delta_{X_i}$ for i.i.d. random variables $X_i \sim \alpha$.

On a finite $\mathcal{X}$, B_n is the multivariate Bernstein operator on the probability simplex; on a general compact $\mathcal{X}$, it is the Bernstein–Schnabl construction on $\mathcal{P}(\mathcal{X})$ [478].

By Proposition 10.1, $\widehat{\alpha}_n \to \alpha$ in $\mathcal{W}_p$ almost surely; bounded continuity and dominated convergence therefore give $\mathsf{B}_n f(\alpha) = \mathbb{E} f(\widehat{\alpha}_n) \to f(\alpha)$ pointwise. Wasserstein–Hölder regularity upgrades this qualitative law of large numbers to the following uniform rate.

Proposition 4.34 (Particle-polynomial approximation of Hölder functionals)**.** *Let $\mathcal{X} \subset \mathbb{R}^d$ be compact and let $f : \mathcal{P}(\mathcal{X}) \to \mathbb{R}$ be η-Hölder for $\mathcal{W}_p$, with $p \geqslant 1$, $0 < \eta \leqslant 1$ and constant L. Then, for every $\alpha \in \mathcal{P}(\mathcal{X})$,*

$$|\mathsf{B}_n f(\alpha) - f(\alpha)| \leqslant L\, \mathbb{E}\big[\mathcal{W}_p(\widehat{\alpha}_n, \alpha)^\eta\big]. \tag{4.9}$$

Moreover,

$$\|\mathsf{B}_n f - f\|_\infty \leqslant C_{\mathcal{X},p,d,\eta} L\, r_{n,p,d}^\eta, \tag{4.10}$$

where the empirical Wasserstein scale $r_{n,p,d}$ is defined in (10.3) of Section 10.2.

Proof. The empirical representation (4.8) and Hölder regularity give

$$|\mathsf{B}_n f(\alpha) - f(\alpha)| \leqslant \mathbb{E}|f(\widehat{\alpha}_n) - f(\alpha)| \leqslant L\, \mathbb{E}\big[\mathcal{W}_p(\widehat{\alpha}_n, \alpha)^\eta\big],$$

which proves (4.9). The empirical Wasserstein moment bound (10.4) gives

$$\sup_{\alpha \in \mathcal{P}(\mathcal{X})} \mathbb{E}[\mathcal{W}_p(\widehat{\alpha}_n, \alpha)^p] \leqslant C_{\mathcal{X},p,d} r_{n,p,d}^p.$$

For any nonnegative random variable Z and exponents $0 < \eta \leqslant p$, Lyapunov's moment inequality states that $\mathbb{E}[Z^\eta]^{1/\eta} \leqslant \mathbb{E}[Z^p]^{1/p}$. It is the monotonicity of L^q norms on a probability space and follows directly from Jensen's inequality applied to the concave function $s \mapsto s^{\eta/p}$; see, for instance, [69]. Applying it to $Z = \mathcal{W}_p(\widehat{\alpha}_n, \alpha)$ gives

$$\mathbb{E}\big[\mathcal{W}_p(\widehat{\alpha}_n, \alpha)^\eta\big] \leqslant \mathbb{E}\big[\mathcal{W}_p(\widehat{\alpha}_n, \alpha)^p\big]^{\eta/p},$$

and taking the supremum over α concludes the proof. □

In particular, $\mathsf{B}_n f \to f$ uniformly. The rate cleanly separates the Hölder exponent from the empirical Wasserstein scale, but it does not capture the full computational complexity of the approximation. Indeed, implementing a generic polynomial of order m also requires approximating its kernel φ_m by a finite-dimensional model with, say, M parameters. Without additional tensor, sparsity or low-dimensional structure, approximating a function on the product space $\mathcal{X}^m$ typically requires M to grow exponentially with m. Thus m grows at best logarithmically with the parameter budget, and a rate that is polynomial in m becomes only an inverse power of $\log M$. This extremely slow worst-case behavior reflects the infinite-dimensional nature of Wasserstein space; useful practical rates require additional structure on either the functional or its interaction kernels.

4.5 Measure-to-Measure Maps on Wasserstein Space

The preceding section mapped a probability law into a finite-dimensional vector space, where approximation can be performed coordinatewise. A measure-valued output is more demanding: the codomain is itself a Wasserstein space, so there is no analogous coordinatewise reduction. This section distinguishes transformations that move particles without splitting them from transformations that are intrinsically diffusive.

Maps on Wasserstein space. Once Wasserstein distances provide a topology on probability measures, it is natural to study transformations of probability measures as maps

$$\Phi : \mathcal{P}_p(\mathcal{X}) \longrightarrow \mathcal{P}_p(\mathcal{X})$$

on Wasserstein space. Later chapters use such maps repeatedly: flow matching and diffusion models evolve laws during sampling, one-step transportation methods learn maps between latent and data distributions, and transformers update the empirical law of their tokens; see Chapter 16 and Section 16.4. Two questions are especially useful. The structural question asks whether Φ preserves a discrete particle representation. The metric question asks whether Φ is stable, for instance Lipschitz, for $\mathcal{W}_p$.

Particle-preserving transport representations. The deterministic case is obtained by pushing each input measure through a map that may itself depend on that input measure:

$$\Phi(\alpha) = \Gamma[\alpha]_\sharp \alpha, \qquad \Gamma[\alpha] : \mathcal{X} \to \mathcal{X}. \tag{4.11}$$

Then, for every discrete measure,

$$\alpha = \sum_{i=1}^n a_i \delta_{x_i} \quad \Longrightarrow \quad \Phi(\alpha) = \sum_{i=1}^n a_i \delta_{\Gamma[\alpha](x_i)}. \tag{4.12}$$

Thus the weights and the number of particles are preserved, up to possible collisions between images. This is the natural structure behind deterministic particle methods: particles move, but they do not split. Lavenant and Savaré [325] study when transformations of measures admit transport representatives of the form (4.11), the obstructions to choosing representatives continuously, and the additional regularity available when Φ is Wasserstein-Lipschitz.

When $\mathcal{X} \subset \mathbb{R}^d$, the family $\alpha \mapsto \Gamma[\alpha]$ can equivalently be written as the jointly defined vector-valued map

$$\Gamma : \mathcal{P}_p(\mathcal{X}) \times \mathcal{X} \to \mathbb{R}^d, \qquad \Gamma(\alpha, x) \coloneqq \Gamma[\alpha](x). \tag{4.13}$$

Approximating a particle-preserving map Φ can therefore be separated into approximating $\Gamma(\alpha, x)$ and applying the resulting map to each particle. For each fixed x, the dependence $\alpha \mapsto \Gamma(\alpha, x)$ is a vector-valued functional of the type studied in Section 4.4, so the particle-polynomial construction (4.8) applies coordinatewise. A useful quantitative hypothesis is the uniform estimate

$$\sup_{x \in \mathcal{X}} \|\Gamma(\alpha, x) - \Gamma(\beta, x)\| \leqslant L \mathcal{W}_p(\alpha, \beta)^\eta.$$

At the ambient Euclidean level, applying B_n to $\alpha \mapsto \Gamma(\alpha, x)$ then gives an error bounded uniformly in x by the right-hand side of (4.9). The additional variable x still requires a spatial approximation; once an $\mathcal{X}$-valued representative has been constructed, stability of its push-forward is controlled by Proposition 4.35.

Wasserstein stability. Regularity of Φ is a stability requirement: small perturbations of the input law should not create large changes of the output law. For transport representations, the following elementary estimate separates the spatial Lipschitz constant of the map from its sensitivity to the input measure.

Proposition 4.35 (Wasserstein stability of transport representations)**.** *Let $(\mathcal{X}, d)$ be Polish, let $p \geqslant 1$, let $E \subset \mathcal{X}$, and assume that, for all $\alpha, \beta \in \mathcal{P}_p(\mathcal{X})$ supported in E, the maps $T[\alpha] : E \to \mathcal{X}$ satisfy*

$$d(T[\alpha](x), T[\alpha](y)) \leqslant L_x d(x, y) \quad \text{for all } x, y \in E,$$

and

$$\sup_{y \in E} d(T[\alpha](y), T[\beta](y)) \leqslant L_{\mathrm{law}} \mathcal{W}_p(\alpha, \beta).$$

Then $\Phi(\alpha) = T[\alpha]_\sharp \alpha$ is $(L_x + L_{\mathrm{law}})$-Lipschitz on probability measures supported in E:

$$\mathcal{W}_p(\Phi(\alpha), \Phi(\beta)) \leqslant (L_x + L_{\mathrm{law}}) \mathcal{W}_p(\alpha, \beta).$$

Proof. Let π be an optimal coupling between α and β. The measure $(T[\alpha], T[\beta])_\sharp \pi$ is a coupling between $\Phi(\alpha)$ and $\Phi(\beta)$, hence

$$\mathcal{W}_p(\Phi(\alpha), \Phi(\beta)) \leqslant \left(\int d(T[\alpha](x), T[\beta](y))^p \, \mathrm{d}\pi(x, y) \right)^{1/p}.$$

By the triangle inequality,

$$d(T[\alpha](x), T[\beta](y)) \leqslant L_x d(x, y) + L_{\mathrm{law}} \mathcal{W}_p(\alpha, \beta).$$

Minkowski's inequality gives the claim. □

The fixed-map case is the basic corollary, obtained with $E = \mathcal{X}$ and $L_{\mathrm{law}} = 0$. If $T : \mathcal{X} \to \mathcal{X}$ is L-Lipschitz, then

$$\mathcal{W}_p(T_\sharp \alpha, T_\sharp \beta) \leqslant L \, \mathcal{W}_p(\alpha, \beta).$$

This is the metric counterpart of the elementary push-forward operation introduced in Definition 2.10.

Mean-field attention. Transformers were introduced as sequence-to-sequence architectures driven by self-attention [531] and have become central to language and vision models [83, 191]. After tokenization, embedding and any positional encoding, an input is represented by vectors $x_1, \dots, x_n \in \mathbb{R}^d$, called *tokens*. A transformer manipulates these vectors through a succession of attention, tokenwise neural-network and normalization layers. Attention is the principal mixing operation: it associates with each token a data-dependent weighted average of the value vectors carried by the full token population.

Consider one attention head with parameters $\theta = (Q, K, V)$, suppressing biases and the final output projection. Here $Q, K : \mathbb{R}^d \to \mathbb{R}^r$ are the query and key maps and $V : \mathbb{R}^d \to \mathbb{R}^d$ is the value map. Thus Qx_i is the query carried by token i, while Kx_j and Vx_j are the key and value carried by token j. In the non-causal, unmasked setting used here, its discrete output is

$$\mathrm{Att}_\theta^{(n)}(x_1, \dots, x_n)_i := \sum_{j=1}^n a_{i,j}(x_1, \dots, x_n) V x_j, \qquad a_{i,j}(x_1, \dots, x_n) := \frac{\exp(\langle Qx_i, Kx_j \rangle)}{\sum_{\ell=1}^n \exp(\langle Qx_i, Kx_\ell \rangle)}, \tag{4.14}$$

Thus every query token may attend to every key token. Causal attention instead masks future indices; the resulting order dependence is deliberately omitted here. The conventional factor $1/\sqrt{r}$ in the score can be absorbed into Q or K.

The map (4.14) is permutation equivariant. Indeed, if σ is a permutation and $x_i^\sigma = x_{\sigma(i)}$, reindexing the denominator gives $a_{i,j}(x_1^\sigma, \dots, x_n^\sigma) = a_{\sigma(i),\sigma(j)}(x_1, \dots, x_n)$, and therefore

$$\mathrm{Att}_\theta^{(n)}(x_1^\sigma, \dots, x_n^\sigma)_i = \mathrm{Att}_\theta^{(n)}(x_1, \dots, x_n)_{\sigma(i)}.$$

Relabeling the input tokens therefore only relabels the outputs. This statement concerns the already encoded token vectors: a causal mask or external position labels break this plain symmetry unless the position is retained as part of the token state. In the unmasked setting, permutation equivariance suggests replacing the ordered tuple by its empirical measure $\alpha = n^{-1} \sum_{i=1}^n \delta_{x_i}$.

Definition 4.36 (Mean-field attention map)**.** Fix $p \geqslant 1$, let $\theta = (Q, K, V)$, and let $\alpha \in \mathcal{P}_p(\mathbb{R}^d)$. Assume that, for α-almost every x,

$$0 < \int e^{\langle Qx, Kz \rangle} \, \mathrm{d}\alpha(z) < +\infty \quad \text{and} \quad \int e^{\langle Qx, Kz \rangle} \|Vz\| \, \mathrm{d}\alpha(z) < +\infty.$$

The *mean-field attention field* is

$$\Gamma_\theta[\alpha](x) := \frac{\int e^{\langle Qx, Kz \rangle} Vz \, \mathrm{d}\alpha(z)}{\int e^{\langle Qx, Kz \rangle} \, \mathrm{d}\alpha(z)} \tag{4.15}$$

and, whenever $\Gamma_\theta[\alpha] \in L^p(\alpha)$, the induced attention map on measures is

$$\mathrm{Att}_\theta(\alpha) := (\Gamma_\theta[\alpha])_\sharp \alpha.$$

These integrability conditions are automatic on compact supports. For an empirical token measure, the construction is an exact rewriting of discrete attention:

$$\Gamma_\theta[\alpha](x_i) = \mathrm{Att}_\theta^{(n)}(x_1, \dots, x_n)_i, \qquad \mathrm{Att}_\theta(\alpha) = \frac{1}{n}\sum_{i=1}^{n} \delta_{\mathrm{Att}_\theta^{(n)}(x_1,\dots,x_n)_i}.$$

For a nonuniform discrete law, the same formula gives weighted attention. Beyond discrete inputs, it defines a continuum extension to arbitrary token distributions satisfying the stated integrability conditions. This viewpoint underlies analyses of attention as a Lipschitz or interacting-particle operator [544, 241], as well as recent mean-field and token-sampling results [117, 70]. The next proposition gives the stability needed to pass from converging empirical token laws to this continuum operator.

Proposition 4.37 (Compact-support attention stability)**.** *Assume $\mathcal{X} = \mathbb{R}^d$. Fix $R > 0$ and let $\mathcal{P}_R$ be the set of probability measures supported in the Euclidean ball $B(0, R)$. For every $p \geqslant 1$, there is a constant $C_{\theta,p}(R)$ such that*

$$\mathcal{W}_p(\mathrm{Att}_\theta(\alpha), \mathrm{Att}_\theta(\beta)) \leqslant C_{\theta,p}(R)\, \mathcal{W}_p(\alpha, \beta), \qquad \alpha, \beta \in \mathcal{P}_R.$$

Writing $A_R = \|Q\|_{\mathrm{op}}\|K\|_{\mathrm{op}}R^2$, one can take a constant bounded, up to polynomial factors in R and the operator norms of Q, K, V, by

$$C_{\theta,p}(R) \lesssim e^{2A_R}.$$

This is the exponential-in-score-radius behavior, equivalently $e^{O(R^2)}$ for dot-product scores on $B(0, R)$, refined in [118].

Proof. Write $A_R = \|Q\|_{\mathrm{op}}\|K\|_{\mathrm{op}}R^2$, and introduce the normalizer and numerator

$$Z_\alpha(x) = \int e^{\langle Qx, Kz\rangle}\, \mathrm{d}\alpha(z), \qquad N_\alpha(x) = \int e^{\langle Qx, Kz\rangle} Vz\, \mathrm{d}\alpha(z).$$

For $x, z \in B(0, R)$, one has $e^{-A_R} \leqslant e^{\langle Qx, Kz\rangle} \leqslant e^{A_R}$, and hence $Z_\alpha(x) \geqslant e^{-A_R}$ and $\|\Gamma_\theta[\alpha](x)\| \leqslant \|V\|_{\mathrm{op}}R$. As functions of z, the scalar kernel $e^{\langle Qx, Kz\rangle}$ and the vector-valued kernel $e^{\langle Qx, Kz\rangle}Vz$ have Lipschitz constants bounded by e^{A_R} times polynomial factors in R and the operator norms of Q, K, V. Kantorovich–Rubinstein duality, applied to the vector kernel after pairing it with arbitrary unit vectors, therefore gives

$$|Z_\alpha(x) - Z_\beta(x)| + \|N_\alpha(x) - N_\beta(x)\| \leqslant C_\theta(R) e^{A_R}\, \mathcal{W}_1(\alpha, \beta).$$

Using

$$\Gamma_\theta[\alpha](x) - \Gamma_\theta[\beta](x) = \frac{N_\alpha(x) - N_\beta(x)}{Z_\alpha(x)} + \Gamma_\theta[\beta](x)\frac{Z_\beta(x) - Z_\alpha(x)}{Z_\alpha(x)}$$

and the preceding bounds yield

$$\sup_{x \in B(0,R)} \|\Gamma_\theta[\alpha](x) - \Gamma_\theta[\beta](x)\| \leqslant C_\theta(R) e^{2A_R}\, \mathcal{W}_1(\alpha, \beta).$$

For spatial regularity, introduce the attention law

$$\mathrm{d}\omega_{\alpha,x}(z) := \frac{e^{\langle Qx, Kz\rangle}}{Z_\alpha(x)}\, \mathrm{d}\alpha(z).$$

Differentiating (4.15) gives

$$D_x\Gamma_\theta[\alpha](x) = \int \big(Vz - \Gamma_\theta[\alpha](x)\big)(Q^\top Kz)^\top\, \mathrm{d}\omega_{\alpha,x}(z),$$

so $\|D_x\Gamma_\theta[\alpha](x)\|_{\mathrm{op}} \leqslant 2\|V\|_{\mathrm{op}}\|Q\|_{\mathrm{op}}\|K\|_{\mathrm{op}}R^2$, uniformly in $\alpha \in \mathcal{P}_R$. Finally, $\mathcal{W}_1 \leqslant \mathcal{W}_p$ on probability measures and the output is supported in $B(0, \|V\|_{\mathrm{op}}R)$. Proposition 4.35, applied with $E = B(0, R)$, proves the claim. □

Consequently, if $\alpha_n, \alpha \in \mathcal{P}_R$ and $\mathcal{W}_p(\alpha_n, \alpha) \to 0$, then $\mathcal{W}_p(\mathrm{Att}_\theta(\alpha_n), \mathrm{Att}_\theta(\alpha)) \to 0$. In particular, empirical measures formed from i.i.d. tokens with a compactly supported law consistently approximate the corresponding mean-field attention output.

Mass-splitting Markov maps. The preceding maps transport each input particle to one output location. A Markov map belongs to a different class: one input particle may generate an entire output distribution. Let K be a Markov kernel on $\mathcal{X}$, so that $K(y,\cdot)$ is a probability measure for each y. To obtain a map on $\mathcal{P}_p(\mathcal{X})$, assume the finite-moment bound

$$\int_{\mathcal{X}} d(x,x_0)^p \, K(y,\mathrm{d}x) \leqslant C\big(1 + d(y,x_0)^p\big)$$

for some $x_0 \in \mathcal{X}$. The associated linear map is defined weakly by

$$\int_{\mathcal{X}} f(x)\,\mathrm{d}\Psi_K(\alpha)(x) \coloneqq \int_{\mathcal{X}}\int_{\mathcal{X}} f(x)\,K(y,\mathrm{d}x)\,\mathrm{d}\alpha(y). \tag{4.16}$$

Integrating the moment bound against α shows that Ψ_K maps $\mathcal{P}_p(\mathcal{X})$ into itself. Unless $K(y,\cdot)$ is a Dirac mass, it does not preserve particles: $\Psi_K(\delta_y) = K(y,\cdot)$ is generally diffuse. Heat flows, noising steps in diffusion models and other smoothing mechanisms belong to this mass-splitting class.

The Wasserstein regularity of the output kernel propagates directly to the induced map on laws.

Proposition 4.38 (Wasserstein stability of Markov maps)**.** *Let $(\mathcal{X},d)$ be Polish, let $p \geqslant 1$, and let K satisfy the preceding moment bound. If*

$$\mathcal{W}_p\big(K(y,\cdot),K(y',\cdot)\big) \leqslant L d(y,y') \qquad \text{for all } y,y' \in \mathcal{X},$$

then the map (4.16) *is L-Lipschitz:*

$$\mathcal{W}_p\big(\Psi_K(\alpha),\Psi_K(\beta)\big) \leqslant L\,\mathcal{W}_p(\alpha,\beta) \qquad \text{for all } \alpha,\beta \in \mathcal{P}_p(\mathcal{X}).$$

Proof. Fix $\pi \in \mathcal{U}(\alpha,\beta)$. On Polish spaces, the measurable-selection theorem for optimal transport [539] provides a probability kernel $(y,y') \mapsto q_{y,y'}$ such that

$$q_{y,y'} \in \mathcal{U}\big(K(y,\cdot),K(y',\cdot)\big) \quad \text{and} \quad \int d(x,x')^p \,\mathrm{d}q_{y,y'}(x,x') = \mathcal{W}_p\big(K(y,\cdot),K(y',\cdot)\big)^p.$$

Integrating this kernel against π defines a coupling

$$q(\mathrm{d}x,\mathrm{d}x') \coloneqq \int_{\mathcal{X}\times\mathcal{X}} q_{y,y'}(\mathrm{d}x,\mathrm{d}x')\,\mathrm{d}\pi(y,y')$$

between $\Psi_K(\alpha)$ and $\Psi_K(\beta)$. Therefore

$$\mathcal{W}_p\big(\Psi_K(\alpha),\Psi_K(\beta)\big)^p \leqslant \int \mathcal{W}_p\big(K(y,\cdot),K(y',\cdot)\big)^p \,\mathrm{d}\pi(y,y') \leqslant L^p \int d(y,y')^p \,\mathrm{d}\pi(y,y').$$

Taking the infimum over π proves the claim. □

Example 4.39 (Convolution is Wasserstein non-expansive). Let $\mathcal{X} = \mathbb{R}^d$ and let $\kappa \in \mathcal{P}_p(\mathbb{R}^d)$. With $\tau_y(z) = y + z$, set $K(y,\cdot) = (\tau_y)_\sharp\kappa$. Then $\Psi_K(\alpha) = \alpha * \kappa$, with convolution as in Definition 4.22. Coupling the translates $y+Z$ and $y'+Z$ with the same $Z \sim \kappa$ gives

$$\mathcal{W}_p\big(K(y,\cdot),K(y',\cdot)\big) \leqslant \|y - y'\|.$$

The reverse inequality follows from $\mathcal{W}_1 \leqslant \mathcal{W}_p$ and Kantorovich–Rubinstein duality applied to the linear function directed along $y - y'$. Hence equality holds, so the sharp kernel constant is $L = 1$, and Proposition 4.38 gives

$$\mathcal{W}_p(\alpha * \kappa, \beta * \kappa) \leqslant \mathcal{W}_p(\alpha,\beta).$$

The Wasserstein viewpoint developed in this chapter turns transport costs into a geometry with meaningful notions of interpolation, convergence and stability. The next chapter adds the complementary dual description, in which marginal constraints are represented by Kantorovich potentials.

CHAPTER 5

Dual Problem

Duality turns the transport problem into a search for potentials rather than couplings. This chapter explains why potentials certify optimality, how c-transforms regularize them, and why the quadratic case reveals convex analysis behind Brenier maps. Linear-programming duality gives the discrete picture [52], while the continuous form is one of the central theorems of OT [538, 469].

5.1 Discrete dual

The discrete dual gives finite-dimensional certificates of optimality. Its complementary slackness conditions identify where an optimal coupling can put mass.

The Kantorovich problem (3.2) is a linear program so that one can equivalently compute its value by solving a dual linear program.

Definition 5.1 (Admissible potentials)**.** For a discrete problem with marginal sizes n, m and cost matrix $C \in \mathbb{R}^{n\times m}$, a pair $(f, g) \in \mathbb{R}^n \times \mathbb{R}^m$ is admissible if it lies below the cost:

$$R(C) := \left\{(f, g) \in \mathbb{R}^n \times \mathbb{R}^m \;;\; \forall (i, j) \in [\![n]\!] \times [\![m]\!], f_i + g_j \leqslant C_{i,j}\right\}. \tag{5.1}$$

Equivalently, $f \oplus g \leqslant C$ entrywise. In particular, the feasible set depends on the cost matrix, not on the marginal weights.

The two vectors play the role of source and target prices; admissibility means that no transported pair is priced above its travel cost.

Proposition 5.2 (Discrete Kantorovich duality)**.** *One has*

$$L_C(a, b) = \max_{(f,g)\in R(C)} \langle f, a\rangle + \langle g, b\rangle. \tag{5.2}$$

Proof. For completeness, introduce multipliers (f, g) for the two marginal constraints. The Lagrangian is

$$\mathcal{L}(P, f, g) = \langle C, P\rangle + \langle a - P\mathbb{1}_m, f\rangle + \langle b - P^\top \mathbb{1}_n, g\rangle. \tag{5.3}$$

For fixed potentials, the dual function is

$$\inf_{P\geqslant 0} \mathcal{L}(P, f, g) = \langle a, f\rangle + \langle b, g\rangle + \inf_{P\geqslant 0}\langle C - (f \oplus g), P\rangle.$$

The last infimum is

$$\inf_{P\geqslant 0}\langle Q, P\rangle = \begin{cases} 0 & \text{if } Q \geqslant 0, \\ -\infty & \text{otherwise,} \end{cases}$$

and is therefore finite exactly when $(f, g) \in R(C)$. The primal polytope is nonempty because it contains the product coupling $a\otimes b$, and it is compact; hence the primal value is finite and attained. Finite-dimensional linear-programming strong duality then identifies this value with the maximum of the dual function and also gives dual attainment. □

The exact primal-dual gap makes the support restriction transparent.

Proposition 5.3 (Discrete complementary slackness)**.** *For every* $P \in U(a, b)$ *and* $(f, g) \in R(C)$,

$$\langle C, P\rangle - \langle f, a\rangle - \langle g, b\rangle = \sum_{i,j} P_{i,j}\big(C_{i,j} - f_i - g_j\big) \geqslant 0.$$

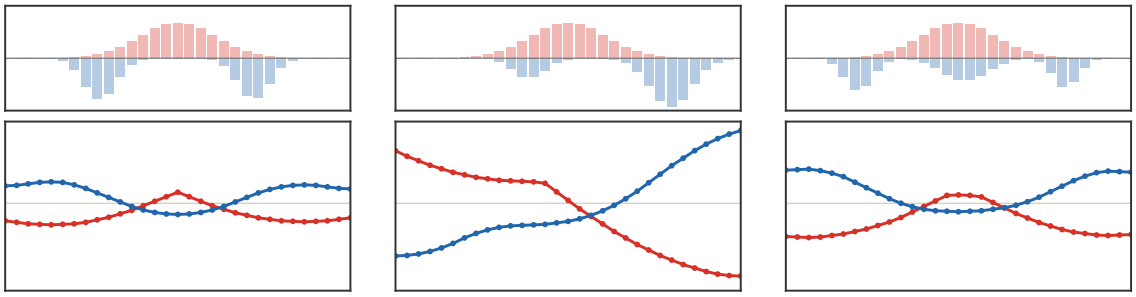
The plan and potentials are respectively primal and dual optimal if and only if

$$\mathrm{Supp}(P) \subset \left\{(i, j) \in [\![n]\!] \times [\![m]\!] \;;\; f_i + g_j = C_{i,j}\right\}. \tag{5.4}$$

Proof. The marginal constraints give $\langle \mathrm{f}, \mathrm{a}\rangle + \langle \mathrm{g}, \mathrm{b}\rangle = \sum_{i,j} \mathrm{P}_{i,j}(\mathrm{f}_i + \mathrm{g}_j)$, which proves the identity and nonnegativity. If both feasible objects are optimal, their values agree by Proposition 5.2; every nonnegative summand with $\mathrm{P}_{i,j} > 0$ must therefore vanish. Conversely, the support condition makes the gap zero, so weak duality forces both objects to be optimal. □

Thus an optimal transport plan can put positive mass only at contact entries of the dual inequality.

Figure 5.1 shows these finite-dimensional certificates on a one-dimensional quadratic problem. The potentials are not transport maps themselves; rather, their contact set with the cost matrix is where an optimal coupling is allowed to put mass.



balanced target | strongly shifted target | three-mode target

Figure 5.1: Discrete Kantorovich dual potentials for the quadratic cost $\mathrm{C}_{i,j} = |x_i - y_j|^2$. In each panel the upper strip shows the fixed source histogram in red and the target histogram in blue: a balanced bimodal target, a more strongly shifted target, and a three-component mixture. The lower strip shows an optimal pair of dual vectors (f, g), with gauge chosen so that $\langle \mathrm{f}, \mathrm{a}\rangle = 0$. Complementary slackness states that mass can be transported only through entries where $\mathrm{f}_i + \mathrm{g}_j = \mathrm{C}_{i,j}$.

The formulation (5.2) shows that $(\mathrm{a}, \mathrm{b}) \mapsto \mathrm{L}_{\mathrm{C}}(\mathrm{a}, \mathrm{b})$ is a convex function (as a supremum of linear functions). From the primal problem (3.2), one also sees that $\mathrm{C} \mapsto \mathrm{L}_{\mathrm{C}}(\mathrm{a}, \mathrm{b})$ is concave.

5.2 General formulation

The continuous dual is the analytic counterpart of the discrete linear program. It uses continuous potentials because measures are probed through integration.

Continuous duality. The finite-dimensional price vectors become continuous test functions paired with the two marginals. To extend this primal-dual construction to arbitrary measures, it is important to realize that measures are naturally paired in duality with continuous functions, using the pairing $\langle f, \alpha\rangle := \int f \mathrm{d}\alpha$. The continuous marginal constraints are encoded by the following feasible set.

Definition 5.4 (Continuous admissible potentials)**.** For $c \in C(\mathcal{X} \times \mathcal{Y})$, define

$$\mathcal{R}(c) := \{(f, g) \in C(\mathcal{X}) \times C(\mathcal{Y}) \,;\; f(x) + g(y) \leqslant c(x, y) \text{ for all } (x, y)\}\,. \tag{5.5}$$

Its elements are admissible continuous potentials.

Proposition 5.5 (Kantorovich duality)**.** *Assume that $\mathcal{X}$ and $\mathcal{Y}$ are compact metric spaces and that $c \in C(\mathcal{X} \times \mathcal{Y})$. Then*

$$\mathcal{L}_c(\alpha, \beta) = \max_{(f,g)\in\mathcal{R}(c)} \int_{\mathcal{X}} f(x)\mathrm{d}\alpha(x) + \int_{\mathcal{Y}} g(y)\mathrm{d}\beta(y), \tag{5.6}$$

Here, (f, g) is a pair of continuous functions, often called "Kantorovich potentials". The same formula extends under the usual lower-semicontinuity and integrability assumptions, replacing maxima by suprema when dual optimizers need not exist.

Proof. Weak duality is immediate: if $f(x) + g(y) \leqslant c(x, y)$ and $\pi \in \mathcal{U}(\alpha, \beta)$, then

$$\int f\mathrm{d}\alpha + \int g\mathrm{d}\beta = \int (f(x) + g(y))\mathrm{d}\pi(x, y) \leqslant \int c\mathrm{d}\pi.$$

Taking the supremum over admissible potentials and the infimum over couplings gives the first inequality.

For the reverse inequality, write $V = \mathcal{L}_c(\alpha, \beta)$ and consider the convex cone

$$\mathcal{K} := \left\{(\pi_1, \pi_2, \int c\mathrm{d}\pi + r) : \pi \in \mathcal{M}_+(\mathcal{X} \times \mathcal{Y}),\ r \geqslant 0\right\}$$

in $\mathcal{M}(\mathcal{X}) \times \mathcal{M}(\mathcal{Y}) \times \mathbb{R}$, equipped with the weak topologies on the measure factors. This cone is closed. Indeed, convergence of the first marginal implies a uniform bound on the masses of the underlying nonnegative measures. Compactness of $\mathcal{X} \times \mathcal{Y}$ then gives a convergent subnet $\pi_k \rightharpoonup \pi$; continuity of marginalization and of $\pi \mapsto \int c\mathrm{d}\pi$ identifies the limit with an element of $\mathcal{K}$, while the limiting epigraph remainder remains nonnegative.

For every $t < V$, the point $z_t = (\alpha, \beta, t)$ does not belong to $\mathcal{K}$. Strong separation of this point from the closed convex cone gives a continuous linear functional. Because $\mathcal{K}$ is a cone containing the origin, the functional can be oriented so that, for some $F \in C(\mathcal{X})$, $G \in C(\mathcal{Y})$ and $\lambda \in \mathbb{R}$,

$$\int F\mathrm{d}\alpha + \int G\mathrm{d}\beta + \lambda t < 0 \leqslant \int F\mathrm{d}\eta + \int G\mathrm{d}\zeta + \lambda s \qquad \text{for all } (\eta, \zeta, s) \in \mathcal{K}.$$

Testing $(0, 0, r) \in \mathcal{K}$ gives $\lambda \geqslant 0$, while testing the element generated by $\pi = \delta_{(x,y)}$ and $r = 0$ gives $F(x) + G(y) + \lambda c(x, y) \geqslant 0$. The multiplier cannot vanish: if $\lambda = 0$, integrating this last inequality against $\alpha \otimes \beta$ would contradict the strict inequality at z_t. Thus $\lambda > 0$, and $f := -F/\lambda, \qquad g := -G/\lambda$ satisfy $f \oplus g \leqslant c$ and $\int f\mathrm{d}\alpha + \int g\mathrm{d}\beta > t$. Letting $t \uparrow V$ proves the reverse inequality.

It remains to prove that the dual supremum is attained. Take a maximizing sequence (f_k, g_k). Successively replace it by

$$\widetilde{g}_k(y) = \min_x\{c(x, y) - f_k(x)\}, \qquad \widetilde{f}_k(x) = \min_y\{c(x, y) - \widetilde{g}_k(y)\}.$$

Each replacement preserves feasibility and cannot decrease the objective. Fixing $x_0 \in \mathcal{X}$, use the gauge freedom $(\widetilde{f}_k, \widetilde{g}_k) \mapsto (\widetilde{f}_k - a_k, \widetilde{g}_k + a_k)$ to impose $\widetilde{f}_k(x_0) = 0$. Uniform continuity of c shows that $\widetilde{g}_k$ inherits a common modulus in the second variable and $\widetilde{f}_k$ a common modulus in the first. The gauge gives a uniform bound on $\widetilde{f}_k$, and the first envelope then gives one on $\widetilde{g}_k$. Arzelà–Ascoli yields uniformly convergent subsequences. Their limit remains admissible and attains the dual supremum. □

> **Remark 5.6 (Dual attainment from c-transforms).** The final part of the proof uses precisely the c-transforms introduced in Section 5.3. A continuous cost on a compact product has uniform moduli of continuity in both variables, and transformed potentials inherit those moduli by Proposition 5.10. If the cost is Lipschitz, this specializes to a common Lipschitz bound; Lipschitz continuity is not required for attainment.

Complementary slackness. As in the finite problem, equality of primal and dual values localizes the support of every optimal coupling.

Proposition 5.7 (Continuous complementary slackness)**.** *For $\pi \in \mathcal{U}(\alpha, \beta)$ and $(f, g) \in \mathcal{R}(c)$,*

$$\int c\mathrm{d}\pi - \int f\mathrm{d}\alpha - \int g\mathrm{d}\beta = \int \big(c(x, y) - f(x) - g(y)\big)\mathrm{d}\pi(x, y) \geqslant 0.$$

The coupling and potentials are respectively primal and dual optimal if and only if this gap vanishes. In that case,

$$\mathrm{Supp}(\pi) \subset \{(x, y) \in \mathcal{X} \times \mathcal{Y}\ ;\ f(x) + g(y) = c(x, y)\}\,. \tag{5.7}$$

Proof. The marginal identities give the equality, and admissibility gives nonnegativity. A zero gap is equivalent to equality of the feasible primal and dual values, hence to joint optimality by Proposition 5.5. Finally, the slack is continuous and nonnegative. If it were positive at a point of $\mathrm{Supp}(\pi)$, it would be bounded below by a positive constant on a neighborhood of positive π-mass, contradicting its zero integral. □

The discrete case (5.2) corresponds to the dual vectors being samples of the continuous potentials, namely $(\mathrm{f}_i, \mathrm{g}_j) = (f(x_i), g(y_j))$. Proposition 5.7 is the exact analogue of Proposition 5.3. For the one-dimensional quadratic cost, the continuous potentials can be read from the monotone map $T = F_\beta^{-1} \circ F_\alpha$: on the active graph, $f'(x) = 2(x - T(x))$ and $g = f^c$.

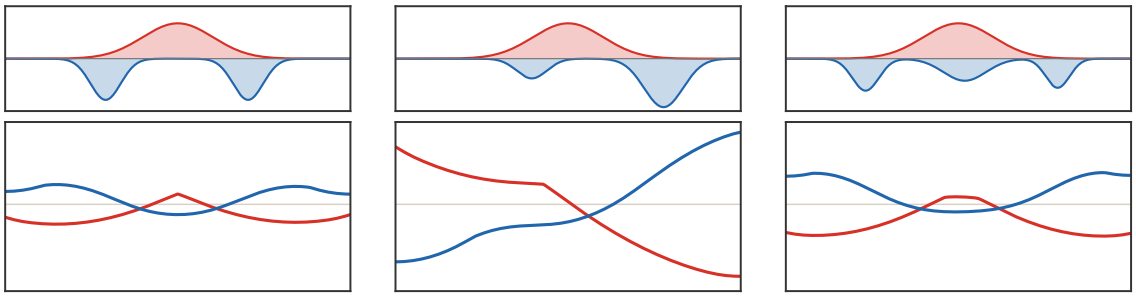


Figure 5.2: Continuous Kantorovich potentials for the same source and target families as Figure 5.1. The upper strips show the source density α in red and the target density β in blue, including the strongly shifted and three-mode targets. The lower strips show potentials f and $g = f^c$ for the quadratic cost $c(x, y) = |x - y|^2$, with the same gauge convention as in the discrete figure. The equality set $f(x) + g(y) = c(x, y)$ contains the monotone transport graph.

Figure 5.2 shows how these continuous potentials adapt to increasingly multimodal source and target densities while retaining the same complementary-slackness interpretation.

The same optimality condition is even more explicit if one plots the *slack*

$$s(x, y) := c(x, y) - f(x) - g(y).$$

Dual feasibility is the statement $s \geqslant 0$, while complementary slackness says that an optimal coupling can only live where this slack vanishes. Figure 5.3 shows this zero set for increasingly structured one-dimensional transports, including a Gaussian source mapped to a three-component target mixture: the transported graph is not guessed from the potentials separately, but appears as the contact valley of the dual inequality.

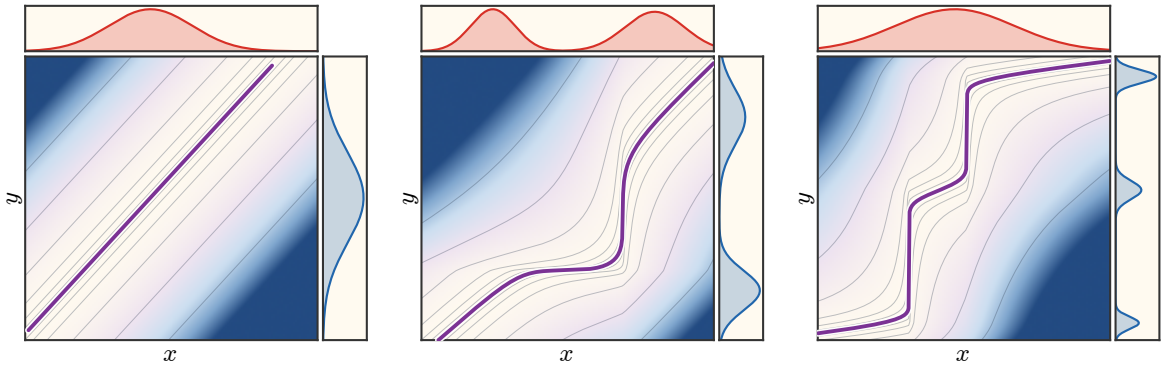


Figure 5.3: Complementary slackness contacts for one-dimensional quadratic OT. From left to right: transport between two Gaussian densities, transport between two-component Gaussian-mixture densities, and transport from one Gaussian density to a separated three-component mixture. Each central heatmap displays the nonnegative dual slack $s(x, y) = |x - y|^2 - f(x) - g(y)$: pale colors mark the contact valley and darker blue means larger strict slack. The red top strip and blue side strip show the source and target densities. The violet curve is the quantile transport contact graph $(x, T(x))$, where $s(x, T(x)) = 0$ and the optimal Monge plan is supported.

In contrast to the primal problem (3.7), dual attainment is not immediate because $\mathcal{R}(c)$ is neither compact nor controlled by a coercive objective. The c-transform resolves this issue: it selects canonical representatives that inherit the modulus of continuity of the cost, and a gauge condition then makes the family compact.

5.3 c-transforms

The c-transform is the operation that improves potentials without changing feasibility. It is both a proof device for dual attainment and the route from duality to Brenier's convex potentials.

Best-response potentials and the c-transform. Keeping a dual potential f fixed, one can maximize in closed form over the second potential in the dual problem (5.6), which leads one to consider

$$\sup_{g \in C(\mathcal{Y})} \left\{ \int g \mathrm{d}\beta \; ; \; \forall (x,y), g(y) \leqslant c(x,y) - f(x) \right\}.$$

For each y, feasibility imposes the upper bound $g(y) \leqslant \inf_x \{c(x,y) - f(x)\}$. Since β is nonnegative, an optimal g must saturate this bound β-almost everywhere. This pointwise best response leads to the following definition.

Definition 5.8 (c-transform)**.** For a real-valued function $f : \mathcal{X} \to \mathbb{R}$, its c-transform is the extended-real-valued function

$$\forall\, y \in \mathcal{Y}, \quad f^c(y) := \inf_{x \in \mathcal{X}} c(x,y) - f(x). \tag{5.8}$$

For a real-valued function $g : \mathcal{Y} \to \mathbb{R}$, the $\bar{c}$-transform associated with $\bar{c}(y,x) = c(x,y)$ is

$$\forall\, x \in \mathcal{X}, \quad g^{\bar{c}}(x) := \inf_{y \in \mathcal{Y}} c(x,y) - g(y).$$

When $\mathcal{X}, \mathcal{Y}$ are compact, c is continuous and the input potential is continuous, these infima are finite minima and the transformed potential is continuous.

Proposition 5.9 (c-transforms solve the semi-relaxed problems)**.** *Assume the compactness and continuity hypotheses of Proposition 5.5. For fixed $f \in C(\mathcal{X})$, the maximizers of the dual objective over all $g \in C(\mathcal{Y})$ such that $f \oplus g \leqslant c$ are exactly the functions satisfying $g = f^c$ β-almost everywhere. Symmetrically, for fixed g, the maximizers over f are exactly the functions satisfying $f = g^{\bar{c}}$ α-almost everywhere.*

Proof. The constraint $f(x) + g(y) \leqslant c(x,y)$ for every x is equivalent to $g(y) \leqslant \inf_x \{c(x,y) - f(x)\} = f^c(y)$. Thus f^c is feasible and pointwise dominates every feasible g. Integrating against β proves optimality, and equality holds exactly when $g = f^c$ β-almost everywhere. The other assertion follows by exchanging the two marginals. □

These transforms also give canonical extensions of dual solutions beyond the active supports.

Proposition 5.10 (Modulus stability of c-transforms)**.** *Suppose that ω is a modulus such that*

$$|c(x,y) - c(x,y')| \leqslant \omega\big(d_{\mathcal{Y}}(y,y')\big) \qquad \text{for all } x, y, y'.$$

Then every finite f^c has modulus ω. Symmetrically, $g^{\bar{c}}$ inherits any uniform modulus of c in its first variable. In particular, a uniform L-Lipschitz bound on the cost yields an L-Lipschitz bound on the corresponding transform.

Proof. For each x, set $F_x(y) = c(x,y) - f(x)$ and $F(y) = f^c(y) = \inf_x F_x(y)$. Since all the functions F_x share the modulus ω,

$$|F(y) - F(y')| = \left| \inf_x F_x(y) - \inf_x F_x(y') \right| \leqslant \sup_x |F_x(y) - F_x(y')| \leqslant \omega\big(d_{\mathcal{Y}}(y,y')\big).$$

The proof in the first variable is identical. □

This stability is precisely the compactness mechanism used to prove dual attainment in Proposition 5.5; see also Remark 5.6. On compact spaces, continuity of c supplies uniform moduli, while sequential closure and a harmless additive gauge give the uniform bounds required by the Arzelà–Ascoli theorem.

Discrete c-transform. If $\alpha = \sum_{i=1}^n a_i \delta_{x_i}$ and $\beta = \sum_{j=1}^m b_j \delta_{y_j}$, the same definition applies to the finite spaces $\mathcal{X} = \{x_i\}_i$ and $\mathcal{Y} = \{y_j\}_j$. The dual potentials are then vectors $(\mathrm{f}, \mathrm{g}) \in \mathbb{R}^n \times \mathbb{R}^m$, as in Section 5.1, and the cost is the matrix $\mathrm{C} \in \mathbb{R}^{n \times m}$ with entries

$$\mathrm{C}_{i,j} = c(x_i, y_j).$$

With the present sign convention $f \oplus g \leqslant c$, the discrete transform of f is the vector $\mathrm{f}^{\mathrm{C}} \in \mathbb{R}^m$ defined by

$$(\mathrm{f}^{\mathrm{C}})_j = \min_{1 \leqslant i \leqslant n} \mathrm{C}_{i,j} - \mathrm{f}_i.$$

This is a minimum because feasibility imposes the upper bound $g_j \leqslant C_{i,j} - f_i$ for every i. Thus the largest feasible best response is the minimum of these upper bounds; formulas with a maximum correspond to a different sign convention, not to the dual constraint (5.2). With $\bar{C} := C^\top$, the opposite transform of g is

$$g^{\bar{C}} := g^{C^\top} \in \mathbb{R}^n, \qquad (g^{\bar{C}})_i = \min_{1 \leqslant j \leqslant m} C_{i,j} - g_j.$$

Thus the continuous best-response operation reduces exactly to taking column minima for f^C, or row minima for $g^{\bar{C}}$, in shifted cost matrices. Chapter 6 studies the hybrid case where α has a density and β is discrete; then $g^{\bar{c}}(x) = \min_j c(x, y_j) - g_j$ is a lower envelope over target atoms, and the regions where each atom realizes the minimum are the Laguerre cells for the quadratic cost.

Figure 5.4 makes the best-response operation geometric: the discrete c-transform is the lower envelope of the shifted cost functions.

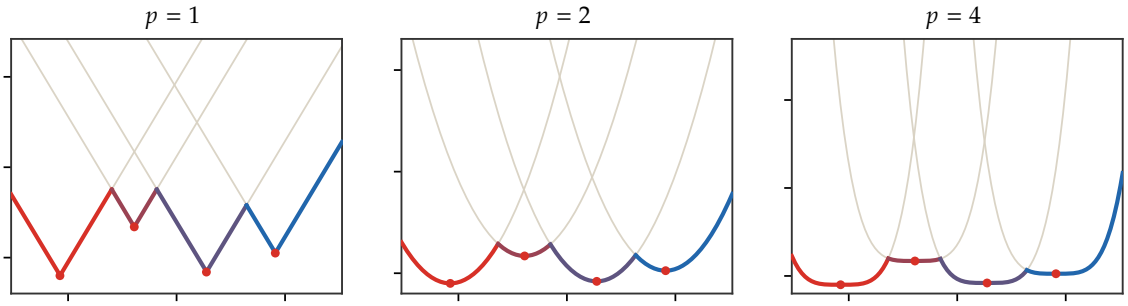


Figure 5.4: Discrete c-transform as a lower envelope for costs $c_p(x, y) = |x - y|^p$. The red circles are four source atoms x_i with potential values f_i; the gray curves are the translated functions $y \mapsto c_p(x_i, y) - f_i$; the colored curve is their lower envelope $f^c(y) = \min_i c_p(x_i, y) - f_i$. This is the semi-discrete situation where $\mathcal{X}$ is finite, equivalently the source measure α is discrete; Chapter 6 studies the complementary case where the eliminated potential is supported on finitely many target atoms.

Euclidean case. The Euclidean quadratic cost is the model case where c-transforms become ordinary convex conjugates after removing the quadratic terms. This is the algebraic bridge between Kantorovich duality and Brenier maps.

It is convenient to normalize the quadratic cost as $c(x, y) = \frac{1}{2}\|x - y\|^2$. For any $\pi \in \mathcal{U}(\alpha, \beta)$,

$$\int c(x, y) \mathrm{d}\pi(x, y) = \frac{1}{2} \int \|x\|^2 \mathrm{d}\alpha(x) + \frac{1}{2} \int \|y\|^2 \mathrm{d}\beta(y) - \int \langle x, y \rangle \mathrm{d}\pi(x, y).$$

The required convex-analytic operation is the Legendre–Fenchel transform.

Definition 5.11 (Legendre–Fenchel transform and biconjugate)**.** For $h : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$, define

$$h^*(y) := \sup_{x \in \mathbb{R}^d} \{\langle x, y \rangle - h(x)\}, \qquad h^{**} := (h^*)^*.$$

The first two terms depend only on the marginals, so the nonseparable part of quadratic transport is the bilinear term $-\langle x, y \rangle$. More precisely, if $\varphi(x) = \frac{1}{2}\|x\|^2 - f(x)$, then

$$f^c(y) = \frac{1}{2}\|y\|^2 - \varphi^*(y), \qquad f^{c\bar{c}}(x) = \frac{1}{2}\|x\|^2 - \varphi^{**}(x).$$

On the full Euclidean space, the c-closed functions are therefore those for which $x \mapsto \frac{1}{2}\|x\|^2 - f(x)$ is lower-semicontinuous and convex; equivalently, f is 1-semiconcave. The double transform is the smallest $\bar{c}$-concave majorant of f. On restricted compact domains, the same statement holds with the additional restriction that the supporting slopes belong to the opposite domain.

Remark 5.12 (Proof of Brenier's theorem). Let $\alpha, \beta \in \mathcal{P}_2(\mathbb{R}^d)$ and assume that α has a density. For

$c(x,y) = \frac{1}{2}\|x-y\|^2$, let (u,v) be a closed optimal dual pair and set

$$\varphi(x) \coloneqq \frac{1}{2}\|x\|^2 - u(x), \qquad \psi(y) \coloneqq \frac{1}{2}\|y\|^2 - v(y).$$

The constraint $u(x)+v(y) \leqslant c(x,y)$ becomes $\varphi(x)+\psi(y) \geqslant \langle x,\, y\rangle$. Since the pair is closed,

$$v(y) = \inf_x\{c(x,y)-u(x)\}, \qquad u(x) = \inf_y\{c(x,y)-v(y)\}.$$

The preceding conjugacy calculation gives $\psi = \varphi^*$ and $\varphi = \psi^*$, so φ is convex. Complementary slackness then shows that an optimal plan satisfies

$$\mathrm{Supp}(\pi) \subset \{(x,y)\,;\ \varphi(x)+\varphi^*(y) = \langle x,\, y\rangle\}\,.$$

By the Fenchel equality criterion, the contact condition is equivalent to $y \in \partial\varphi(x)$. Convex functions are differentiable Lebesgue-almost everywhere, hence α-almost everywhere, so $\partial\varphi(x)$ is a singleton for α-almost every x. This yields the Brenier map $T = \nabla\varphi$ and explains why the optimal coupling is concentrated on a graph. Multiplying the cost by two rescales the dual potentials but leaves this map unchanged, recovering the convention $c(x,y) = \|x-y\|^2$ used elsewhere in the book.

Why hard alternating optimization stops. Alternating best responses must be applied sequentially. Starting from an admissible pair, the updates

$$(f,g) \longmapsto (f,f^c) \longmapsto (f^{c\bar{c}}, f^c)$$

preserve feasibility and cannot decrease the dual objective. By contrast, the simultaneous replacement $(f,g) \mapsto (g^{\bar{c}}, f^c)$ need not be feasible because its two components are computed against different old coordinates. For example, if $c \equiv 0$ and $f = g \equiv -1$, the original pair is feasible, whereas both transforms equal 1 and their simultaneous replacement is not.

A crucial property of the Legendre transform is that $f^{***} = f^*$, while f^{**} is the lower-semicontinuous convex envelope of a proper function f. Analogous identities explain why exact alternating c-transform updates stop after one full cycle. The required convex-analytic background is standard in Rockafellar's theory [452].

Proposition 5.13 (Algebra of c-transforms)**.** *Writing $f^{c\bar{c}} \coloneqq (f^c)^{\bar{c}}$, the following pointwise identities hold whenever the transforms are well defined:*

$$\text{(i) } f \leqslant f' \Rightarrow f^c \geqslant f'^c, \qquad \text{(ii) } f^{c\bar{c}} \geqslant f, \qquad \text{(iii) } g^{\bar{c}c} \geqslant g, \qquad \text{(iv) } f^{c\bar{c}c} = f^c.$$

Proof. The order reversal in (i) follows directly from the minus sign in the definition. For every x and y,

$$f^c(y) = \inf_{x'}\{c(x',y) - f(x')\} \leqslant c(x,y) - f(x),$$

so $c(x,y) - f^c(y) \geqslant f(x)$. Taking the infimum over y proves (ii), and exchanging the two spaces proves (iii). Finally, (ii) and the order reversal imply $f^c \geqslant f^{c\bar{c}c}$, whereas (iii) applied to $g = f^c$ gives $f^{c\bar{c}c} \geqslant f^c$. This proves (iv). □

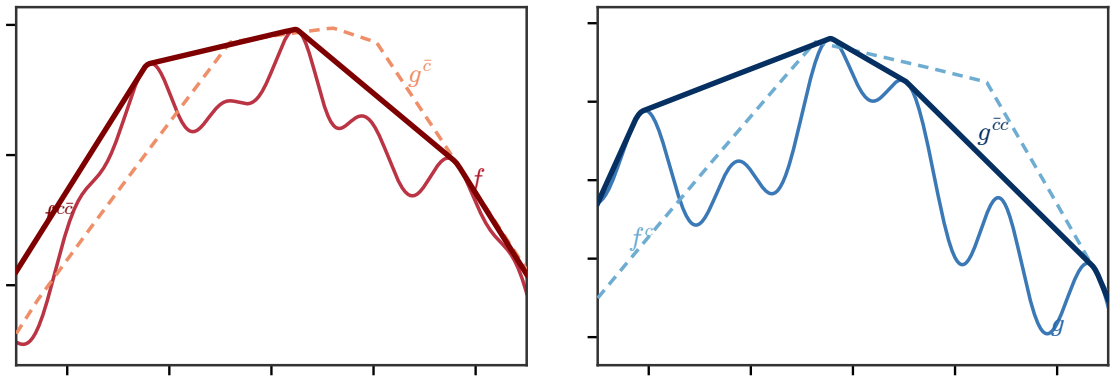
Identity (iv) shows that the next best response leaves the pair obtained after one full cycle unchanged:

$$(f^{c\bar{c}}, f^c) \longmapsto (f^{c\bar{c}}, f^{c\bar{c}c}) = (f^{c\bar{c}}, f^c). \tag{5.9}$$

Thus exact block maximization reaches a coordinatewise fixed point after one full cycle. The resulting pair is dual feasible, but it need not maximize the joint dual objective because its value still depends on the arbitrary initialization f.

The issue is therefore not that the best responses are inaccurate, but that hard closure becomes idempotent before solving the joint optimization problem. Entropic regularization changes these hard minima into soft log-sum-exp responses. Their alternating updates are no longer exhausted by one closure cycle and lead to the Sinkhorn scaling algorithm studied in Chapter 8.

For the bilinear cost $c(x,y) = -xy$ on a compact interval, $f^{c\bar{c}}$ is the smallest concave majorant representable with slopes in the opposite interval. In Figure 5.5, the displayed interval contains all relevant supporting slopes, so this restricted closure agrees with the ordinary concave envelope. A hard transform removes non-concave oscillations in one step rather than producing a gradual ascent.



source-side closure target-side closure

Figure 5.5: Hard c-transforms for the bilinear cost $c(x, y) = -xy$. The source-side panel uses a reddish palette and starts from a sharply oscillatory potential f; the target-side panel uses a blueish palette and starts from a visually different potential g. Dark curves are the double-transform closures $f^{c\bar{c}}$ and $g^{\bar{c}c}$, while dashed lighter curves are the one-sided best responses $g^{\bar{c}}$ and f^c after a harmless vertical gauge shift. The plotted domains contain the relevant supporting slopes, so the restricted closures coincide here with ordinary concave majorants. The figure illustrates why exact alternating best responses are useful for dual certificates but do not give the smooth iterative dynamics later provided by entropic regularization.

c-concave functions. The ranges of the two c-transform operators define the natural classes of dual potentials.

Definition 5.14 (c-concave functions)**.** A function on $\mathcal{Y}$ is c-concave if it equals f^c for some function f on $\mathcal{X}$. A function on $\mathcal{X}$ is $\bar{c}$-concave if it equals $g^{\bar{c}}$ for some function g on $\mathcal{Y}$.

These classes have considerably more structure than their range characterization alone suggests.

Definition 5.15 (Finite-valued c-concave class)**.** The finite-valued c-concave class on $\mathcal{Y}$ is

$$\mathrm{Conc}_c(\mathcal{Y}) := \{u : \mathcal{Y} \to \mathbb{R} \; ; \; u = f^c \text{ for some } f : \mathcal{X} \to \mathbb{R} \cup \{-\infty\}\}.$$

The value $f(x) = -\infty$ suppresses the section $c(x, \cdot) - f(x)$ and is needed for exact full-space characterizations. The defining infimum makes this class stable under min-plus combinations.

Proposition 5.16 (Stability under tropical barycenters)**.** *If $u_r = f_r^c \in \mathrm{Conc}_c(\mathcal{Y})$ and $a_r \in \mathbb{R}$ for $1 \leqslant r \leqslant R$, then*

$$\min_{1 \leqslant r \leqslant R} \{u_r + a_r\} = \left(\max_{1 \leqslant r \leqslant R} \{f_r - a_r\} \right)^c \in \mathrm{Conc}_c(\mathcal{Y}).$$

Proof. Set $f = \max_r (f_r - a_r)$. For every $y \in \mathcal{Y}$,

$$f^c(y) = \inf_x \min_r \{c(x, y) - f_r(x) + a_r\} = \min_r \inf_x \{c(x, y) - f_r(x) + a_r\} = \min_r \{u_r(y) + a_r\}.$$

The middle equality holds because the minimum ranges over finitely many indices: both sides equal the infimum over $(x, r) \in \mathcal{X} \times \{1, \dots, R\}$. □

Thus $\mathrm{Conc}_c(\mathcal{Y})$ is a min-plus, or tropical, convex cone: tropical addition is pointwise minimum, while tropical scalar multiplication adds a constant. This universal closure does not imply stability under ordinary averages.

Three model costs distinguish the possibilities. On $\mathcal{X} = \mathcal{Y} = \mathbb{R}^d$, the cost $c(x, y) = -\langle x, y \rangle$ gives exactly the convex cone of finite concave functions. For $c_\tau(x, y) = \|x - y\|^2/(2\tau)$, one has

$$u \in \mathrm{Conc}_{c_\tau}(\mathbb{R}^d) \quad \Longleftrightarrow \quad y \mapsto u(y) - \frac{\|y\|^2}{2\tau} \text{ is concave,}$$

so the class is an ordinary convex set but not a cone: for instance, $y \mapsto \|y\|^2/(2\tau)$ belongs to it, whereas twice this function does not. Finally, on any metric space $(\mathcal{X}, d)$,

$$\mathrm{Conc}_d(\mathcal{X}) = \left\{u : \mathcal{X} \to \mathbb{R}\,;\ \mathrm{Lip}(u) \leqslant 1\right\},$$

which becomes the convex unit ball of the Lipschitz norm after quotienting by constants. Proposition 6.16, in the section on the $\mathcal{W}_1$ norm, proves this identity and the sharper relation $u^c = -u$.

Ordinary convexity beyond these examples is a geometric property of the cost. Under the standard smooth twist, nondegeneracy, compactness and domain c-convexity hypotheses, Figalli, Kim and McCann show that the class of c-concave potentials is convex if and only if c has *nonnegative cross-curvature*, a strengthening of the Ma–Trudinger–Wang condition developed by Kim and McCann [304, 215]. A finite product M of round spheres has this property for $c = d_M^2/2$, and hence $\mathrm{Conc}_c(M)$ is convex [305, 216]. The factor $1/2$ is immaterial here: for every $\eta > 0$, $\mathrm{Conc}_{\eta c} = \eta\,\mathrm{Conc}_c$.

Léger, Todeschi and Vialard extend nonnegative cross-curvature to nonsmooth and infinite-dimensional costs [332]. In that generality, the exact convexity statement concerns potentials admitting a c-support at every point; for finite continuous costs on compact spaces this is precisely the class above. Under their standard finiteness hypotheses, they also prove stability under Kantorovich lifting: if c has nonnegative cross-curvature, then so does the cost $(\alpha, \beta) \mapsto \mathcal{L}_c(\alpha, \beta)$ from (3.7). For $c = d^2$, this lift is $\mathcal{W}_2^2$ by (4.3); iteration connects the result to the Wasserstein towers $\mathcal{P}_2(\mathcal{X}), \mathcal{P}_2(\mathcal{P}_2(\mathcal{X})), \ldots$ of Section 13.2.

Example 5.17 (Neural parameterizations of Monge maps). Convex stability can be enforced by the architecture rather than by a penalty. For $c(x, y) = \|x - y\|^2/2$, a source dual potential u is c-concave exactly when $\varphi(x) = \|x\|^2/2 - u(x)$ is convex, and the associated map is $T = \nabla\varphi = \mathrm{Id} - \nabla u$; see Remark 5.12. Input-convex neural networks enforce convexity through convex nondecreasing activations and nonnegative hidden-to-hidden weights [13]. Makkuva, Taghvaei, Oh and Lee parameterize the convex potentials by such networks and learn the resulting Monge map [365]. On a sphere or a product of spheres, symmetry of the cost makes the finite envelopes

$$u_\theta(x) = \min_{1 \leqslant r \leqslant R} \left\{\frac{1}{2} d_M(x, y_r)^2 + a_r\right\}$$

c-concave by Proposition 5.16. At differentiability points they generate the intrinsic map $T_\theta(x) = \exp_x(-\nabla u_\theta(x))$ from Section 2.5. Cohen, Amos and Lipman use these finite envelopes as universal approximants of manifold c-concave potentials and, on spheres and their products, take learned ordinary convex combinations without leaving the c-concave class [150]. Their hard-min construction therefore preserves exact c-concavity; replacing the minimum by a soft minimum, as is convenient for training, no longer carries this guarantee.

CHAPTER 6

Semi-discrete and $\mathcal{W}_1$

This chapter develops three computational consequences of duality. Eliminating one potential gives the semi-dual; for discrete measures, this viewpoint leads to auction algorithms, while a continuous source and discrete target lead to Laguerre-cell geometry. The final part specializes duality to $\mathcal{W}_1$, where Lipschitz functions and flow fields replace convex potentials. The material connects auction and network-flow methods [50, 51], computational geometry [24, 380, 381], and the Kantorovich–Rubinstein and Beckmann formulations [295, 31].

6.1 Semi-dual

The semi-dual eliminates one potential by an exact c-transform. It preserves concavity while removing the explicit pointwise inequality constraint.

General measure semi-dual. For arbitrary measures, partial maximization converts the constrained two-potential dual into an unconstrained optimization over one function.

The full objective and the result of eliminating its source potential will be used repeatedly in this chapter. We write $\mathcal{D}_0(f, g)$ for the hard two-potential dual objective and $\mathcal{E}_0(g)$ for its one-potential semi-dual; the subscript 0 records the absence of entropic regularization. Section 9.5 uses the parallel notation $\mathcal{D}_\varepsilon(f, g)$ and $\mathcal{E}_\varepsilon(g)$ at positive temperature.

Definition 6.1 (Full dual and semi-dual functionals)**.** Under the compactness and continuity assumptions of Chapter 5, define

$$\mathcal{D}_0(f, g) := \begin{cases} \int_X f \mathrm{d}\alpha + \int_Y g \mathrm{d}\beta, & (f, g) \in \mathcal{R}(c), \\ -\infty, & \text{otherwise.} \end{cases} \tag{6.1}$$

$$\mathcal{E}_0(g) := \mathcal{D}_0(g^{\bar{c}}, g) = \int_X g^{\bar{c}} \mathrm{d}\alpha + \int_Y g \mathrm{d}\beta. \tag{6.2}$$

Thus the dual problem (5.6) is $\mathcal{L}_c(\alpha, \beta) = \max_{f,g} \mathcal{D}_0(f, g)$. For fixed g, feasibility is equivalent to $f \leqslant g^{\bar{c}}$. Since α is nonnegative, the largest admissible choice $f = g^{\bar{c}}$ maximizes the objective. Consequently,

$$\mathcal{L}_c(\alpha, \beta) = \sup_{g \in C(Y)} \mathcal{E}_0(g), \qquad \mathcal{E}_0(g) = \sup_{f \in C(X)} \mathcal{D}_0(f, g).$$

Partial maximization of a jointly concave objective preserves concavity, so $\mathcal{E}_0$ is concave. The semi-dual is invariant under $g \mapsto g + s$: indeed, $(g + s)^{\bar{c}} = g^{\bar{c}} - s$ and both measures have unit mass. Potentials are therefore determined only up to an additive constant, exactly as in the full dual. The main computational advantage is that the semi-dual is unconstrained.

Discrete semi-dual. When both measures are discrete, the same elimination produces a finite-dimensional concave objective whose active affine pieces are selected by the discrete c-transform.

Let $\alpha = \sum_{i=1}^n \mathrm{a}_i \delta_{x_i}$, $\beta = \sum_{j=1}^m \mathrm{b}_j \delta_{y_j}$, and $\mathrm{C}_{i,j} = c(x_i, y_j)$, where $\mathrm{a} \in \mathbb{R}_+^n$ and $\mathrm{b} \in \mathbb{R}_+^m$ have the same total mass M. We use the same notation for functions and vectors: in the discrete setting,

$$\mathcal{D}_0(\mathrm{f}, \mathrm{g}) := \begin{cases} \langle \mathrm{f}, \mathrm{a} \rangle + \langle \mathrm{g}, \mathrm{b} \rangle, & \mathrm{f} \oplus \mathrm{g} \leqslant \mathrm{C}, \\ -\infty, & \text{otherwise.} \end{cases}$$

Eliminating the source potential through the discrete $\bar{\mathrm{C}}$-transform described in Section 5.3 gives

$$\mathrm{L}_\mathrm{C}(\mathrm{a}, \mathrm{b}) = \max_{\mathrm{g} \in \mathbb{R}^m} \mathcal{E}_0(\mathrm{g}), \qquad \mathcal{E}_0(\mathrm{g}) := \mathcal{D}_0(\mathrm{g}^{\bar{\mathrm{C}}}, \mathrm{g}) = \sum_{i=1}^n \mathrm{a}_i (\mathrm{g}^{\bar{\mathrm{C}}})_i + \sum_{j=1}^m \mathrm{b}_j \mathrm{g}_j, \tag{6.3}$$

with

$$(g^{\bar{C}})_i = \min_{1 \leqslant j \leqslant m} (C_{i,j} - g_j).$$

The function $\mathcal{E}_0$ is concave and piecewise affine. It is invariant under $g \mapsto g + s\mathbb{1}_m$, because the two marginals have equal mass. If a tie-breaking rule selects

$$\sigma_g(i) \in \operatorname{argmin}_j (C_{i,j} - g_j),$$

the induced target masses determine a supergradient explicitly.

Proposition 6.2 (Supergradients of the discrete semi-dual)**.** *For any selection $\sigma_g(i) \in \operatorname{argmin}_j (C_{i,j} - g_j)$, define*

$$\widehat{b}_j(g) := \sum_{i:\, \sigma_g(i)=j} a_i.$$

Then $b - \widehat{b}(g)$ is a supergradient of $\mathcal{E}_0$ at g. If every row has a unique minimizer, $\mathcal{E}_0$ is differentiable at g and

$$\nabla \mathcal{E}_0(g) = b - \widehat{b}(g).$$

Proof. Let $h \in \mathbb{R}^m$. Since $\sigma_g(i)$ is active at g,

$$\min_j (C_{i,j} - h_j) \leqslant C_{i,\sigma_g(i)} - h_{\sigma_g(i)} = \min_j (C_{i,j} - g_j) - \big(h_{\sigma_g(i)} - g_{\sigma_g(i)}\big).$$

Multiplying by a_i, summing over i, and adding $\langle b,\, h - g\rangle$ gives

$$\mathcal{E}_0(h) \leqslant \mathcal{E}_0(g) \langle b - \widehat{b}(g),\, h - g\rangle,$$

which is the supergradient inequality. Uniqueness of all active indices makes the same affine branch active in a neighborhood of g, yielding the gradient formula. □

Thus the semi-dual supergradient is exactly the mismatch between prescribed target mass and mass currently attracted by each target coordinate. At ties, convex combinations of active selections, equivalently splitting each source mass among its tied minimizers, give the full superdifferential. The auction method developed next specializes this same target-potential formulation to uniform square assignments and couples targeted dual updates with ownership changes.

6.2 Auction Algorithm

The auction algorithm is derived from coordinate maximization of the semi-dual of the linear assignment problem. Its practical form uses bidder-specific dual-weight updates that cross the selected row's next indifference threshold by ε; this controlled relaxation prevents jamming at nonsmooth ties. We follow the particularly transparent account of Mérigot and Thibert [383, Sec. 3.2], itself based on Bertsekas' auction algorithm and its ε-scaling refinement [49, 51, 50]. Throughout this section, $C \in \mathbb{R}^{n \times n}$ and $n \geqslant 2$; the case $n = 1$ is immediate. A permutation matrix $P \in \mathcal{P}_n^{\mathrm{perm}}$, as defined in Definition 3.11, represents the probability coupling P/n.

Coordinate ascent and discrete Laguerre cells. Auction dual weights give the discrete semi-dual a geometric partition whose cell cardinalities are its coordinate derivatives.

Write g_j for the target Kantorovich potential. Specializing (6.3) to uniform weights gives

$$\mathcal{E}_0(g) = \frac{1}{n}\sum_{i=1}^{n} (g^{\bar{C}})_i + \frac{1}{n}\sum_{j=1}^{n} g_j, \qquad (g^{\bar{C}})_i = \min_{1 \leqslant j \leqslant n} (C_{i,j} - g_j). \tag{6.4}$$

Thus $(g^{\bar{C}}, g)$ is dual feasible. This is precisely the discrete $\bar{C}$-transform described in Section 5.3, with the same sign convention as in the general and semi-discrete formulations.

The discrete Laguerre cell of target j is

$$\mathbb{L}_j^{\mathrm{D}}(g) := \left\{ i \in [\![n]\!] \;;\; C_{i,j} - g_j \leqslant C_{i,k} - g_k \text{ for every } k \right\}. \tag{6.5}$$

These cells can overlap at ties. They are the finite counterpart of the Laguerre cells introduced for the general semi-discrete problem in Definition 6.7: specifically, (6.5) is that definition restricted to the source indices. When every row has a unique minimizer, differentiation of (6.4) gives

$$\frac{\partial}{\partial g_j}\mathcal{E}_0(g) = \frac{1 - \left|\mathbb{L}_j^{\mathrm{D}}(g)\right|}{n}. \tag{6.6}$$

Hence an overfull cell calls for a decrease of its dual weight. More generally, Proposition 1.7 shows that a perfect matching in the contact graph $i \in \mathbb{L}_j^{\mathrm{D}}(g)$ certifies optimality of g in (6.4); conversely, assignment duality and complementary slackness produce such a matching at every maximizer. Auction searches for this certificate while moving one coordinate of g at a time.

For $i \in \mathbb{L}_j^{\mathrm{D}}(g)$, define the bid needed to make row i indifferent between j and its best alternative,

$$\mathrm{bid}_j(g, i) := \min_{k \neq j}(C_{i,k} - g_k) - (C_{i,j} - g_j). \tag{6.7}$$

When $\mathbb{L}_j^{\mathrm{D}}(g)$ is nonempty, the largest maximizing decrement along the negative jth coordinate is given by $\max_{i \in \mathbb{L}_j^{\mathrm{D}}(g)} \mathrm{bid}_j(g, i)$ [383, Lem. 20]. This formula explains both the coordinate-ascent interpretation and its weakness: at a tie, even this largest maximizing decrement can vanish, and naive coordinate ascent can jam before reaching a dual maximizer.

Bids and relaxed contacts. Exact coordinate maximization can stall at a nonsmooth tie, so auction deliberately moves a bidder ε beyond its next indifference point. Suppose row i is currently unassigned. Let j_0 and j_1 be its best and second-best targets for the current dual weights,

$$j_0 \in \operatorname{argmin}_j(C_{i,j} - g_j), \qquad j_1 \in \operatorname{argmin}_{j \neq j_0}(C_{i,j} - g_j), \tag{6.8}$$

with fixed deterministic tie-breaking. The bid decreases only the winning dual weight by

$$\Delta_i := \mathrm{bid}_{j_0}(g, i) + \varepsilon = (C_{i,j_1} - g_{j_1}) - (C_{i,j_0} - g_{j_0}) + \varepsilon, \qquad g_{j_0} \leftarrow g_{j_0} - \Delta_i. \tag{6.9}$$

Unlike exact coordinate maximization, this update uses the bid of the selected row rather than the maximum bid over the whole cell. Since $\Delta_i \geqslant \varepsilon$, every selected weight decreases by at least ε. After the update, the reduced cost of j_0 for row i is exactly ε above that of the unchanged alternative j_1. The row nevertheless takes j_0; if that target already has an owner, the former owner becomes unassigned. Because of this overshoot, a bid need not increase the nonsmooth semi-dual. Coordinate ascent motivates the update, but ε-complementary slackness, rather than monotonicity of the objective, is the invariant that drives the convergence proof.

Definition 6.3 (ε-complementary slackness)**.** A partial permutation matrix $P \in \{0,1\}^{n \times n}$ has at most one nonzero entry in each row and column. Such a matrix and target potential g satisfy ε-complementary slackness if

$$P_{i,j} = 1 \quad \Longrightarrow \quad C_{i,j} - g_j \leqslant (g^{\bar{C}})_i + \varepsilon. \tag{6.10}$$

The bid (6.9) gives the newly assigned row this property. Decreasing g_{j_0} can only make j_0 less attractive to rows assigned elsewhere, and its previous owner is removed. Therefore every iteration preserves ε-complementary slackness.

Figure 6.1 shows actual iterates of Algorithm 6.1 on the planar point clouds of Figure 1.5. To expose the dual mechanism geometrically, define the auction reduced costs

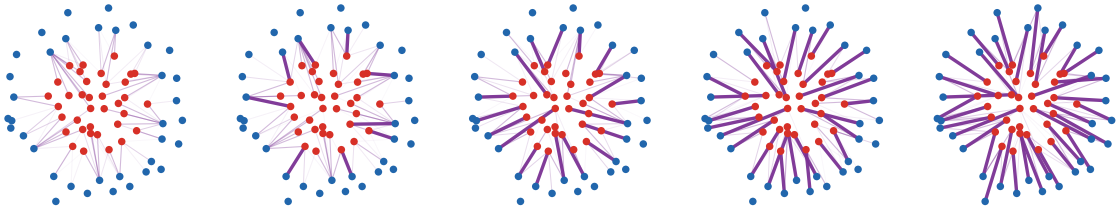
$$r_{i,j}(g) := C_{i,j} - g_j - (g^{\bar{C}})_i \geqslant 0.$$

At any intermediate state, write $M := \{(i, j) : P_{i,j} = 1\}$ for the current ownership matching. Exact zeros of $r_{i,j}$ are the discrete Laguerre contacts of row i, whereas an edge in M only needs to satisfy $r_{i,j}(g) \leqslant \varepsilon$ by ε-complementary slackness.

Algorithm 6.1: Bertsekas auction

Input: Cost matrix $C \in \mathbb{R}^{n\times n}$, tolerance $\varepsilon > 0$, initial target potential $g \in \mathbb{R}^n$ (default $g = 0$).
Output: Permutation matrix P and target potential g; the source potential is $g^{\bar{C}}$.
Initialize: Set $P = 0$.
While some row i satisfies $\sum_j P_{i,j} = 0$ **do**:
  Choose any such row i.
  Set $j_0 \in \operatorname{argmin}_j (C_{i,j} - g_j)$ and $j_1 \in \operatorname{argmin}_{j\neq j_0} (C_{i,j} - g_j)$.
  Set $\Delta \leftarrow (C_{i,j_1} - g_{j_1}) - (C_{i,j_0} - g_{j_0}) + \varepsilon$.
  Set $g_{j_0} \leftarrow g_{j_0} - \Delta$.
  If $P_{i_0,j_0} = 1$ for some i_0 **then**:
    Set $P_{i_0,j_0} \leftarrow 0$.
  Set $P_{i,j_0} \leftarrow 1$.

Return P and g.



$|M| = 0$, 0 bids | $|M| = 9$, 10 bids | $|M| = 18$, 40 bids | $|M| = 27$, 98 bids | $|M| = 36$, 505 bids

Figure 6.1: Geometric progression of the unit-mass transportation auction. Thick violet segments form the current partial ownership matching M, while the thin translucent segments are the $2n$ unmatched edges with smallest auction reduced costs; exact zero-reduced-cost alternatives are slightly stronger. The panels record the first states with the indicated $|M|$ and report the cumulative number of bids. For $\varepsilon = 2\times 10^{-3}$, the final assignment is reached after 505 bids and coincides with the exact squared-distance optimum.

Fixed tolerance. For a prescribed tolerance, the relaxed contact condition gives both an optimality certificate and a finite bound on the number of bids.

Proposition 6.4 (Fixed-ε auction convergence and complexity)**.** *Let*

$$R_C := \max_{i,j} C_{i,j} - \min_{i,j} C_{i,j}.$$

Started from $g = 0$, Algorithm 6.1 terminates after at most $n\big(\lfloor R_C/\varepsilon \rfloor + 1\big)$ bids. It returns a permutation matrix P satisfying ε-complementary slackness and

$$0 \leqslant \frac{1}{n}\langle C, P\rangle - \min_{P'\in \mathcal{P}_n^{\mathrm{perm}}} \frac{1}{n}\langle C, P'\rangle \leqslant \varepsilon. \tag{6.11}$$

With dense scans for the two smallest reduced costs, the algorithm requires $O\big(n^2(1 + R_C/\varepsilon)\big)$ arithmetic and comparison operations and $O(n^2)$ storage, including C. If C has integer entries and $\varepsilon < 1/n$, the returned assignment is exactly optimal.

Proof. At every bid, (6.9) decreases one target potential by at least ε. Once a target is assigned, it remains assigned: a later bid can change its owner but never leaves it empty. While the algorithm is incomplete, there is therefore an unassigned target whose potential is still zero. If the selected target j_0 is already assigned, comparison with such a zero-potential target gives $g_{j_0} \geqslant -R_C$; if it is unassigned, its potential is itself zero. Thus a target can be selected at most $\lfloor R_C/\varepsilon \rfloor + 1$ times. Summing over the n targets proves finite termination and the bid bound. At termination, every row and column has exactly one nonzero entry, so P is a permutation matrix.

The ε-complementary-slackness invariant was established above. Summing it over the returned permutation gives the first inequality below. The second is the dual lower bound of Proposition 1.7,

applied to the feasible pair $(g^{\bar{C}}, g)$:

$$\frac{1}{n}\langle C, P\rangle \leqslant \frac{1}{n}\sum_i (g^{\bar{C}})_i + \frac{1}{n}\sum_j g_j + \varepsilon \leqslant \min_{P' \in \mathcal{P}_n^{\mathrm{perm}}} \frac{1}{n}\langle C, P'\rangle + \varepsilon,$$

which proves (6.11). Each bid scans one row of C and otherwise uses constant work if the current owner of every target is stored, proving the complexity bound. For integer costs, the unnormalized assignment gap is an integer bounded by $n\varepsilon < 1$, hence it vanishes. □

ε-scaling. The cold-start estimate in Proposition 6.4 exposes the limitation of using a single tolerance. A large ε terminates quickly but gives only a coarse certificate, whereas the small ε needed for high accuracy can produce a bid count proportional to R_C/ε. The remedy is continuation: first learn the rough dual-potential landscape with a coarse auction, then decrease the tolerance while retaining that potential.

Algorithm 6.2 starts at the cost scale $\max\{R_C, \eta\}$ and halves the tolerance until it reaches the requested value η. Each phase rebuilds the ownership matrix but warm-starts g with the preceding phase. Intuitively, the previous complete assignment already certifies approximate contact for this potential; the next phase only needs to refine that contact rather than rediscover the dual landscape from $g = 0$.

Algorithm 6.2: Auction with ε-scaling

Input: Cost matrix $C \in \mathbb{R}^{n\times n}$, final tolerance $\eta > 0$.
Output: Permutation matrix P and target potential g; the source potential is $g^{\bar{C}}$.
Initialize: Set $R_C \leftarrow \max_{i,j} C_{i,j} - \min_{i,j} C_{i,j}$ and $g \leftarrow 0$.
If $R_C = 0$ **then**:
  Return any permutation matrix P and $g = 0$.

Set $\varepsilon \leftarrow \max\{R_C, \eta\}$.
While $\varepsilon > \eta$ **do**:
  Set $(P, g) \leftarrow \mathrm{Auction}(C, \varepsilon, g)$.
  Set $\varepsilon \leftarrow \max\{\varepsilon/2, \eta\}$.

Set $(P, g) \leftarrow \mathrm{Auction}(C, \eta, g)$.
Return P and g.

Proposition 6.5 (Complexity of ε-scaling)**.** *For a final tolerance $\eta > 0$, Algorithm 6.2 returns a permutation matrix satisfying η-complementary slackness and (6.11) with ε replaced by η. If $R_C > 0$, it uses at most $1 + \lceil \log_2^+(R_C/\eta)\rceil$ auction phases, where $\log_2^+(s) := \max\{0, \log_2 s\}$. Its dense worst-case complexity is*

$$O\left(n^3\left(1 + \log_+ \frac{R_C}{\eta}\right)\right), \qquad \log_+(s) := \max\{0, \log s\}.$$

When $R_C = 0$, the algorithm terminates immediately. If C has integer entries and $\eta < 1/n$, the returned assignment is exactly optimal.

Proof. Assume $R_C > 0$. The first tolerance is $\varepsilon_0 = \max\{R_C, \eta\}$. If $\varepsilon_0 = \eta$, Proposition 6.4 directly gives the claim with one phase. Otherwise, the first phase starts from the zero potential and costs $O(n^2)$ operations because $R_C/\varepsilon_0 = 1$.

Consider a later phase with tolerance ε, initialized from a target potential produced at tolerance λ. The preceding complete assignment satisfies λ-complementary slackness for this potential. Expressed in the present dual-potential convention, the warm-start estimate of [383, Lem. 25] shows that the new phase decreases each component of g by at most $n(\lambda + \varepsilon)$. Since every bid decreases one component by at least ε, the phase performs at most $n^2\left(1 + \frac{\lambda}{\varepsilon}\right)$ bids. Consecutive tolerances in Algorithm 6.2 satisfy $\lambda/\varepsilon \leqslant 2$, so a warm-started phase uses $O(n^2)$ bids and $O(n^3)$ dense operations.

Halving from R_C to η gives the announced phase count. The final phase enforces η-complementary slackness, and Proposition 6.4 gives the η-optimality certificate. Its integer-cost conclusion also applies with $\varepsilon = \eta$. □

Compared with a single η-auction started from zero, whose dense bound grows as $O(n^2(1 + R_C/\eta))$, scaling replaces the linear accuracy factor by a logarithmic number of $O(n^3)$ phases. This is a high-accuracy guarantee rather than an unconditional speedup: at coarse tolerances the cold-start bound can be smaller, while the scaling bound improves it once R_C/η is large compared with $n\big(1 + \log_+(R_C/\eta)\big)$. For integer costs, choosing $\eta < 1/n$ gives an exact assignment with complexity $O\big(n^3(1 + \log_+(nR_C))\big)$.

> **Remark 6.6 (ε-scaling and relation with Sinkhorn).** Sinkhorn uses a superficially similar continuation parameter, but its role is different. Entropy replaces the hard minimum in the semi-dual by a smooth log-sum-exp, whereas auction retains the hard discrete transform and relaxes only the contact condition. Sinkhorn evolves a dense coupling by smooth block maximization; auction maintains a partial permutation matrix through dual-weight bids and ownership changes.

6.3 Semi-discrete

The semi-discrete case is the setting where dual potentials become weights of Laguerre cells. This gives both geometry and algorithms for quantization and density fitting.

Discrete target and Laguerre cells. A case of particular interest is when $\beta = \sum_{j=1}^m \mathrm{b}_j \delta_{y_j}$ has distinct atoms and positive weights. Zero-weight atoms can simply be removed. The same construction applies when α is discrete by exchanging the roles of α and β. One can adapt Definition 5.8 of the $\bar{c}$-transform to this setting by restricting the minimization to the finite support $\{y_j\}_{j=1}^m$ of β. Equivalently, one applies the same definition with the discrete target space $\mathcal{Y} = \{y_j\}_{j=1}^m$ and identifies a vector $\mathrm{g} \in \mathbb{R}^m$ with the function $g : \mathcal{Y} \to \mathbb{R}$ defined by $g(y_j) = \mathrm{g}_j$:

$$\forall\, \mathrm{g} \in \mathbb{R}^m,\ \forall\, x \in \mathcal{X}, \quad \mathrm{g}^{\bar{c}}(x) \coloneqq \min_{j \in [\![m]\!]} c(x, y_j) - \mathrm{g}_j. \tag{6.12}$$

This transform maps a vector g to a continuous function $\mathrm{g}^{\bar{c}} \in C(\mathcal{X})$ because it is the minimum of finitely many continuous functions.

Crucially, using the discrete $\bar{c}$-transform, when β is a discrete measure, yields a finite-dimensional optimization,

$$\mathcal{L}_c(\alpha, \beta) = \max_{\mathrm{g} \in \mathbb{R}^m} \mathcal{E}_0(\mathrm{g}) \coloneqq \mathcal{D}_0(\mathrm{g}^{\bar{c}}, \mathrm{g}) = \int_{\mathcal{X}} \mathrm{g}^{\bar{c}}(x) \mathrm{d}\alpha(x) + \sum_j \mathrm{g}_j \mathrm{b}_j. \tag{6.13}$$

The objective satisfies $\mathcal{E}_0(\mathrm{g} + s\mathbb{1}) = \mathcal{E}_0(\mathrm{g})$ for every $s \in \mathbb{R}$. One may therefore impose the gauge $\sum_j \mathrm{g}_j = 0$ without changing the problem. The geometric object encoded by the dual weights is a weighted nearest-neighbor diagram: each source point is assigned to a target atom realizing the discrete $\bar{c}$-transform.

> **Definition 6.7** (Laguerre cells and power diagrams)**.** For sites $(y_j)_{j=1}^m$ and weights $\mathrm{g} \in \mathbb{R}^m$, the Laguerre cell associated with y_j is
>
> $$\mathbb{L}_j(\mathrm{g}) \coloneqq \left\{ x \in \mathcal{X} \,;\, \forall\, j' \neq j, c(x, y_j) - \mathrm{g}_j \leqslant c(x, y_{j'}) - \mathrm{g}_{j'} \right\}. \tag{6.14}$$
>
> The cells cover $\mathcal{X}$; after arbitrary tie-breaking on common boundaries, they induce a disjoint partition. When $c(x, y) = \|x - y\|^2$, this Laguerre decomposition is also called a power diagram. If g is constant, it reduces to the ordinary Voronoi diagram.

For quadratic costs, varying the dual weights moves the walls between adjacent cells while keeping them parallel; this is the geometric mechanism by which the cell masses are adjusted. Figure 6.2 follows this adjustment from unweighted Voronoi cells to a power diagram whose cell masses match the prescribed discrete target weights.

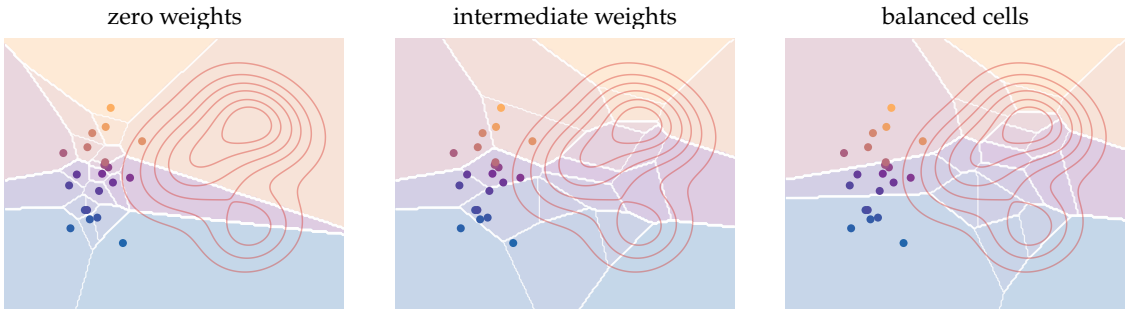


Figure 6.2: Laguerre cells for semi-discrete quadratic transport. The red contours show a continuous source density α given by a three-component Gaussian mixture on the right. The twenty-one colored circular sites are the atoms of the discrete target β sampled from a compact Gaussian cloud on the left; each site color matches its Laguerre cell. Starting from ordinary Voronoi cells, semi-dual weight updates deform the cells so that the α-mass captured by each cell approaches the prescribed target mass.

Mass balance. This gives the convenient semi-dual form

$$\mathcal{E}_0(\mathrm{g}) = \sum_{j=1}^{m} \int_{\mathbb{L}_j(\mathrm{g})} \big(c(x, y_j) - \mathrm{g}_j\big)\, \mathrm{d}\alpha(x) + \langle \mathrm{g}, \mathrm{b} \rangle. \tag{6.15}$$

The following proposition provides a formula for the gradient of this concave function.

Proposition 6.8 (Gradient of the semi-discrete dual)**.** *If the minimizing index in (6.12) is unique for α-almost every x, then $\mathcal{E}_0$ is differentiable at g and*

$$\forall\, j \in [\![m]\!], \quad \nabla\mathcal{E}_0(\mathrm{g})_j = \mathrm{b}_j - \int_{\mathbb{L}_j(\mathrm{g})} \mathrm{d}\alpha.$$

Proof. For α-almost every x, the minimizing index in $\min_j c(x, y_j) - \mathrm{g}_j$ is unique. If this index is $j(x)$, then the directional derivative in a direction $h \in \mathbb{R}^m$ is

$$\frac{\mathrm{d}}{\mathrm{d}t}\bigg|_{t=0} \min_j \big(c(x, y_j) - (\mathrm{g}_j + t h_j)\big) = -h_{j(x)}.$$

The corresponding difference quotients are bounded by $\|h\|_\infty$. Dominated convergence therefore gives

$$\mathrm{d}\mathcal{E}_0(\mathrm{g})[h] = -\sum_j h_j \int_{\mathbb{L}_j(\mathrm{g})} \mathrm{d}\alpha + \sum_j h_j \mathrm{b}_j,$$

which is the announced gradient formula. □

Because the gradient components sum to zero, gradient ascent preserves the chosen gauge. At a differentiable maximizer, first-order optimality gives $\alpha(\mathbb{L}_j(\mathrm{g})) = \mathrm{b}_j$ for every j. Conversely, if these identities hold, the piecewise-constant map $T(x) = y_j$ on $\mathbb{L}_j(\mathrm{g})$ pushes α onto β. Its graph lies in the contact set $\mathrm{g}^{\bar{c}}(x) + \mathrm{g}_j = c(x, y_j)$, so complementary slackness, Proposition 5.7, proves that T and g are optimal. For the quadratic cost, uniqueness of the map follows from Brenier's theorem when α has a density.

Figure 6.3 makes the sign of this gradient geometric. Increasing a weight lowers the corresponding power distance and expands the associated Laguerre cell; decreasing it shrinks the cell. The dotted outline marks the balanced cell, so the semi-dual ascent can be read as a mass-balancing procedure on a power diagram.

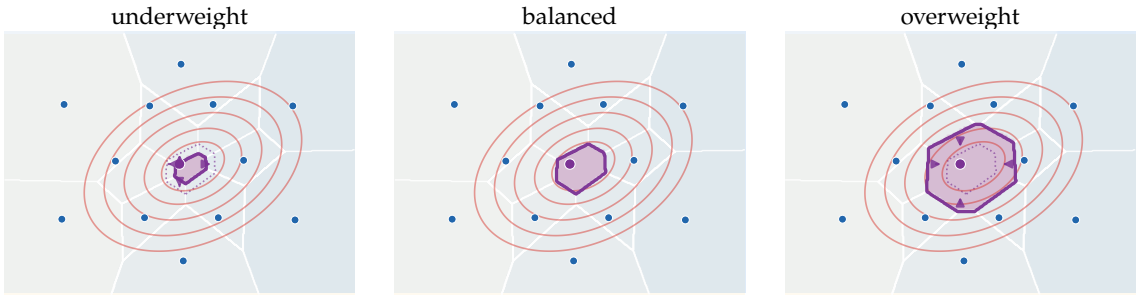


Figure 6.3: Dual weights control Laguerre cell masses in the semi-discrete quadratic problem. The same blue target sites and red Gaussian source density are used in all panels; only the highlighted violet weight is changed. The dotted violet outline is the balanced cell. If the highlighted cell has too little source mass, then $\mathrm{b}_j - \alpha(\mathbb{L}_j(\mathrm{g})) > 0$ and the ascent update increases the weight, expanding the cell outward. If it has too much mass, the update decreases the weight, shrinking it inward. At balance, the cell mass matches the prescribed target mass and the first-order update vanishes.

Algorithm 6.3: Semi-discrete Laguerre ascent

Input: Source measure α, target atoms (y_j, b_j), cost c, steps $(\tau^{(\ell)})_{\ell=0}^{L-1}$, tolerance tol, iteration budget $L \geqslant 1$.
Output: Semi-discrete dual weights g and Laguerre cells.
Initialize: Set $\mathrm{g}^{(0)} = 0$.
For $\ell = 0, \dots, L-1$ **do**:

- **Compute cells:** $\mathbb{L}_j(\mathrm{g}^{(\ell)}) = \left\{ x \,;\, c(x, y_j) - \mathrm{g}_j^{(\ell)} \leqslant c(x, y_r) - \mathrm{g}_r^{(\ell)} \quad \forall r \right\}$.
- **Compute masses:** $m_j^{(\ell)} = \int_{\mathbb{L}_j(\mathrm{g}^{(\ell)})} \mathrm{d}\alpha$.
- **If** $\max_j \left| m_j^{(\ell)} - \mathrm{b}_j \right| \leqslant \mathrm{tol}$ **then**:
  - **Return** $\mathrm{g}^{(\ell)}$ and $\left(\mathbb{L}_j(\mathrm{g}^{(\ell)})\right)_{j=1}^m$.
- **Update** $\mathrm{g}^{(\ell+1)} = \mathrm{g}^{(\ell)} + \tau^{(\ell)}\left(\mathrm{b} - m^{(\ell)}\right)$.

Compute $\left(\mathbb{L}_j(\mathrm{g}^{(L)})\right)_{j=1}^m$.
Return $\mathrm{g}^{(L)}$ and these cells.

The quadratic power diagrams of Definition 6.7 have polyhedral cells and can be computed efficiently using computational-geometry algorithms [23, 24, 380]. Expanding the squared distance shows that the active cell at x minimizes the affine function $x \mapsto -2\langle x, y_j \rangle + \|y_j\|^2 - \mathrm{g}_j$. The lower envelope of these affine functions produces the power diagram, while the lower convex hull of the lifted sites $(y_j, \|y_j\|^2 - \mathrm{g}_j)$ produces its dual regular triangulation. For a planar source domain, the lift lives in $\mathbb{R}^3$; Chan's output-sensitive three-dimensional convex-hull algorithm computes a hull with Q vertices in $O(m \log Q)$ time [122]. A three-dimensional source requires a four-dimensional hull and therefore does not fall under that particular complexity bound.

Stochastic optimization. The semi-discrete formulation (6.15) is useful because the objective is an expectation with respect to α,

$$\mathcal{E}_0(\mathrm{g}) = \int_{\mathcal{X}} E(\mathrm{g}, x) \mathrm{d}\alpha(x) = \mathbb{E}_X(E(\mathrm{g}, X)), \qquad E(\mathrm{g}, x) := \mathrm{g}^{\bar{c}}(x) + \langle \mathrm{g}, \mathrm{b} \rangle. \tag{6.16}$$

Here $X \sim \alpha$. Away from cell boundaries, the stochastic gradient of the integrand is

$$\nabla_{\mathrm{g}} E(\mathrm{g}, x) = \left(\mathrm{b}_j - \mathbb{1}_{\mathbb{L}_j(\mathrm{g})}(x) \right)_{j=1}^m,$$

which is an unbiased estimator of $\nabla \mathcal{E}_0(\mathrm{g})$ when cell boundaries have α-measure zero. One can therefore maximize (6.15) without first discretizing α: the measure is used as a black box from which independent samples are drawn, a natural setup in high-dimensional statistics and machine learning.

Starting from $\mathrm{g}^{(0)} = 0$, stochastic gradient ascent draws $x^{(\ell)} \sim \alpha$ and performs

$$\mathrm{g}^{(\ell+1)} := \mathrm{g}^{(\ell)} + \tau^{(\ell)} \nabla_{\mathrm{g}} E(\mathrm{g}^{(\ell)}, x^{(\ell)}). \tag{6.17}$$

Equivalently, if

$$j^{(\ell)} \in \operatorname{argmin}_j \big(c(x^{(\ell)}, y_j) - \mathrm{g}_j^{(\ell)}\big),$$

then the coordinate update is

$$\mathrm{g}_j^{(\ell+1)} = \mathrm{g}_j^{(\ell)} + \tau^{(\ell)} \big(\mathrm{b}_j - \mathbb{1}_{\{j=j^{(\ell)}\}}\big).$$

The stochastic supergradient has zero coordinate sum, so the iterates preserve the gauge $\sum_j \mathrm{g}_j = 0$. For almost-sure stochastic-approximation convergence, one typically chooses

$$\sum_{\ell=0}^{+\infty} \tau^{(\ell)} = +\infty, \qquad \sum_{\ell=0}^{+\infty} (\tau^{(\ell)})^2 < +\infty, \tag{6.18}$$

for instance $\tau^{(\ell)} = \tau^{(0)}(1 + \ell/\ell_0)^{-q}$ with $1/2 < q \leqslant 1$. A finite-horizon $O(L^{-1/2})$ guarantee instead uses averaging and a horizon-dependent step, as made precise next.

The following estimate is the classical averaged stochastic-supergradient bound for concave optimization, instantiated for the semi-dual objective; see Nemirovski, Juditsky, Lan and Shapiro [397] and its OT specialization by Genevay et al. [238].

Proposition 6.9 (Averaged stochastic semi-dual rate). *Let $\mathrm{g}^\star$ maximize $\mathcal{E}_0$, set $R = \|\mathrm{g}^{(0)} - \mathrm{g}^\star\|_2$, and suppose the stochastic supergradients are conditionally unbiased and satisfy $\|\nabla_{\mathrm{g}} E(\mathrm{g}, x)\|_2 \leqslant G$. For a horizon $L \geqslant 1$, use the constant step $\tau = R/(G\sqrt{L})$ and the average $\bar{\mathrm{g}}_L = L^{-1} \sum_{\ell=0}^{L-1} \mathrm{g}^{(\ell)}$. If $R = 0$, the conclusion is immediate; otherwise,*

$$\mathcal{E}_0(\mathrm{g}^\star) - \mathbb{E}\big[\mathcal{E}_0(\bar{\mathrm{g}}_L)\big] \leqslant \frac{RG}{\sqrt{L}}.$$

Proof. Conditionally on $\mathrm{g}^{(\ell)}$, the stochastic supergradient is unbiased. Concavity and the squared-distance recursion give

$$2\tau\, \mathbb{E}\Big[\mathcal{E}_0(\mathrm{g}^\star) - \mathcal{E}_0(\mathrm{g}^{(\ell)})\Big] \leqslant \mathbb{E}\|\mathrm{g}^{(\ell)} - \mathrm{g}^\star\|_2^2 - \mathbb{E}\|\mathrm{g}^{(\ell+1)} - \mathrm{g}^\star\|_2^2 + \tau^2 G^2.$$

Summing from 0 to $L-1$, discarding the final nonnegative squared distance, and using concavity at $\bar{\mathrm{g}}_L$ yields

$$\mathcal{E}_0(\mathrm{g}^\star) - \mathbb{E}\big[\mathcal{E}_0(\bar{\mathrm{g}}_L)\big] \leqslant \frac{R^2}{2\tau L} + \frac{\tau G^2}{2}.$$

The stated choice of τ gives the result. □

Here one may take $G \leqslant 2$ because each sample supergradient is $\mathrm{b} - e_{j^{(\ell)}}$. This stochastic viewpoint is one of the main algorithmic advantages of the semi-discrete formulation [380, 238].

Algorithm 6.4: Stochastic semi-discrete ascent

Input: Source sampler $x \sim \alpha$, target atoms (y_j, b_j), steps $(\tau^{(\ell)})_{\ell=0}^{L-1}$, iteration count L.
Output: Stochastic semi-discrete dual weights g.
Initialize: Set $\mathrm{g}^{(0)} = 0$.
For $\ell = 0, \dots, L-1$ **do**:
  Draw $x^{(\ell)} \sim \alpha$.
  Set $j^{(\ell)} = \min \operatorname{argmin}_j \big(c(x^{(\ell)}, y_j) - \mathrm{g}_j^{(\ell)}\big)$.
  For $j = 1, \dots, m$ **do**:
    $\mathrm{g}_j^{(\ell+1)} = \mathrm{g}_j^{(\ell)} + \tau^{(\ell)} \big(\mathrm{b}_j - \mathbb{1}_{\{j=j^{(\ell)}\}}\big)$.
Return $\bar{\mathrm{g}}_L = L^{-1} \sum_{\ell=0}^{L-1} \mathrm{g}^{(\ell)}$.

6.4 Optimal Quantization

Optimal quantization asks for the best discrete approximation of a measure by m codepoints. It is the geometric core of vector quantization, compression and k-means clustering.

Free masses and prescribed weights. The classical free-mass problem optimizes both the codepoint positions and their probabilities. Prescribing the weights gives a different approximation problem, whose equal-weight specialization is treated below.

The free-mass approximation is the basic quantization problem.

Definition 6.10 (Optimal quantization problem)**.** For $\alpha \in \mathcal{P}_p(\mathbb{R}^d)$ and $m \in \mathbb{N} \setminus \{0\}$, define

$$\mathcal{Q}_m(\alpha) := \min_{Y=(y_j)_{j=1}^m,\ \mathrm{b}\in\Sigma_m} \mathcal{W}_p(\alpha, \sum_{j=1}^m \mathrm{b}_j \delta_{y_j}). \tag{6.19}$$

Zero weights and coincident codepoints are allowed, so the approximating measure has at most m atoms.

This problem is classical in approximation theory and information theory [251, 348].

In (6.19), one optimizes both the codepoint positions $Y = (y_j)_j$ and the weights $\mathrm{b} \in \Sigma_m$. The equal-weight case, $\mathrm{b}_j = 1/m$, is especially important for particle approximations and is isolated at the end of this section.

Proposition 6.11 (Quantization rate and curse of dimensionality)**.** *Let α be supported in a bounded subset $\Omega \subset \mathbb{R}^d$ and assume that $\alpha = \rho\,\mathrm{d}x$ with $\rho \leqslant \rho_+ < +\infty$. Then, for fixed $p \geqslant 1$, there exist constants $0 < c \leqslant C < +\infty$ such that*

$$c\, m^{-1/d} \leqslant \mathcal{Q}_m(\alpha) \leqslant C\, m^{-1/d}.$$

Proof. Enclose Ω in a cube. With $k = \lfloor m^{1/d} \rfloor$, subdivide this cube into $k^d \leqslant m$ congruent subcubes and place one codepoint in each subcube meeting Ω. Every source point then lies within distance $C/k \leqslant C' m^{-1/d}$ of a codepoint. The nearest-codepoint coupling proves the upper bound; unused codepoints may be assigned zero mass.

For the lower bound, fix any set Y of at most m codepoints and write $d_Y(x) = \min_j \|x - y_j\|$. If ω_d is the volume of the Euclidean unit ball, then $\alpha(\{d_Y \leqslant t\}) \leqslant \rho_+ m \omega_d t^d$. Set $t_0 = (2\rho_+ m \omega_d)^{-1/d}$. Then $\alpha(\{d_Y > t\}) \geqslant 1/2$ for $0 < t < t_0$, and the layer-cake formula gives

$$\int d_Y(x)^p \mathrm{d}\alpha(x) = \int_0^{+\infty} p t^{p-1} \alpha(\{d_Y > t\}) \mathrm{d}t \geqslant \frac{t_0^p}{2} \simeq c m^{-p/d}.$$

Taking the p-th root and minimizing over Y proves the lower bound. □

This deterministic rate mirrors the empirical OT sample-complexity rate: both are governed by the spacing $m^{-1/d}$ of points in dimension d. Quantization is best-case and deterministic, while empirical OT is random, but both display the same curse of dimensionality. Zador's theorem refines this comparison by identifying the sharp asymptotic constant and the limiting density of optimal codepoints [251]. For fixed codepoints Y, the powered transport cost $\mathrm{b} \mapsto \mathcal{W}_p^p(\alpha, \sum_j \mathrm{b}_j \delta_{y_j})$ is convex because it is an optimal value of a linear program with affine marginal constraints. This statement need not hold for its p-th root $\mathcal{W}_p$. The dependence on Y is non-convex and is generally computationally hard. The rest of this section distinguishes the free-mass Lloyd reduction from the fixed-weight geometry underlying finite-particle $\mathcal{W}_2$ gradient flows.

Lloyd algorithm. The computational appeal of quantization comes from splitting the non-convex search over sites into two elementary operations. For fixed sites, the optimal assignment is purely local: each point is sent to one of its nearest sites, and the resulting cells are Voronoi cells. This is the assignment step behind Lloyd's algorithm and the k-means method.

Proposition 6.12 (Free masses give Voronoi cells)**.** *For the cost $c(x, y) = d(x, y)^p$, fix distinct codepoints $Y = (y_j)_{j=1}^m$. Duplicate codepoints can be merged beforehand. Minimizing over the weights $\mathrm{b} \in \Sigma_m$ gives*

$$\min_{\mathrm{b}\in\Sigma_m} \mathcal{W}_p^p(\alpha, \sum_j \mathrm{b}_j \delta_{y_j}) = \int_{\mathcal{X}} \min_{1\leqslant j\leqslant m} c(x, y_j) \mathrm{d}\alpha(x).$$

An optimal coupling is induced by sending each x to a nearest codepoint. The corresponding cells are the Voronoi cells

$$\mathbb{V}_j(Y) := \left\{x \ ;\ \forall\, j',\ c(x, y_j) \leqslant c(x, y_{j'})\right\},$$

up to arbitrary tie-breaking on common boundaries.

Proof. For any coupling between α and a measure supported on Y, the conditional destination of a point x belongs to Y, hence its conditional cost is at least $\min_j c(x, y_j)$. Integrating gives the lower bound. Conversely, choose a measurable nearest-codepoint map $T_Y(x) \in \operatorname{argmin}_j c(x, y_j)$, breaking ties measurably, and set $\mathrm{b}_j = \alpha(T_Y^{-1}(y_j))$. Then $(T_Y)_\sharp \alpha = \sum_j \mathrm{b}_j \delta_{y_j}$ and the induced transport reaches the displayed lower bound. □

Consequently, the quantization energy can be written in the nearest-centroid form

$$Q_m(\alpha)^p = \min_Y \mathcal{F}(Y), \qquad \mathcal{F}(Y) := \int_X \min_{1 \leqslant j \leqslant m} c(x, y_j) \mathrm{d}\alpha(x).$$

At a differentiability point of this energy, any local minimizer with nonempty cells satisfies the centroid condition

$$y_j \in \operatorname*{argmin}_y \int_{\mathbb{V}_j(Y)} c(x, y) \mathrm{d}\alpha(x).$$

For the squared Euclidean cost, this becomes the fixed-point equation

$$y_j = \frac{\int_{\mathbb{V}_j(Y)} x \mathrm{d}\alpha(x)}{\int_{\mathbb{V}_j(Y)} \mathrm{d}\alpha}.$$

Lloyd's algorithm, also known as the k-means algorithm, iterates this fixed point: assign points to nearest sites, then replace each site by the centroid of its cell [348]. The assignment step minimizes the cost for fixed sites, and the centroid step minimizes it for the resulting fixed partition; hence the objective cannot increase. Since the problem is non-convex in Y, this descent does not guarantee a global minimizer and may stop at a stationary configuration or local minimum. Good seeding matters. For a finite data set with squared Euclidean loss, k-means++ gives an expected logarithmic approximation guarantee [20].

Continuous Lloyd flow. There is also an infinitesimal version of Lloyd's fixed point, but it should first be understood on finite labelled configurations. Assume that $c(x, y) = \|x - y\|^2$ and that α does not charge Voronoi boundaries. For a configuration Y, define, on nonempty cells,

$$a_j(Y) = \alpha(\mathbb{V}_j(Y)), \qquad b_j(Y) = \frac{1}{a_j(Y)} \int_{\mathbb{V}_j(Y)} x \, \mathrm{d}\alpha(x)$$

as the cell mass and centroid. Empty cells are singular points of the vector field; one either freezes them, as in the algorithm below, or reseeds them. The relaxed Lloyd step

$$y_j^{(\ell+1)} = y_j^{(\ell)} + \tau\big(b_j(Y^{(\ell)}) - y_j^{(\ell)}\big), \qquad 0 < \tau \leqslant 1,$$

is the explicit-Euler discretization of the cell-mass preconditioned gradient flow of the quantization energy $\mathcal{F}$,

$$\dot{y}_j(t) = b_j(Y_t) - y_j(t).$$

Indeed, at differentiability points of $\mathcal{F}$, the envelope theorem gives

$$\nabla_{y_j} \mathcal{F}(Y) = 2 \int_{\mathbb{V}_j(Y)} (y_j - x) \mathrm{d}\alpha(x) = 2 a_j(Y)(y_j - b_j(Y)),$$

so that

$$\dot{y}_j(t) = -\frac{1}{2a_j(Y_t)} \nabla_{y_j} \mathcal{F}(Y_t).$$

Equivalently, this is the gradient flow of $\mathcal{F}$ for the site metric $g_Y(U, V) = 2 \sum_j a_j(Y) \langle u_j, v_j \rangle$; it is not the unweighted Euclidean gradient flow unless the masses are absorbed into the time step. Along smooth portions of the flow,

$$\frac{\mathrm{d}}{\mathrm{d}t} \mathcal{F}(Y_t) = -2 \sum_j a_j(Y_t) \|b_j(Y_t) - y_j(t)\|^2 \leqslant 0.$$

If $\eta_t = \sum_j w_j \delta_{y_j(t)}$ carries fixed positive weights, independent of the Voronoi masses $a_j(Y_t)$, Proposition 14.2 turns this labelled particle ODE into the weak continuity equation

$$\partial_t \eta_t + \operatorname{div}(v_t \eta_t) = 0, \qquad v_t(y_j(t)) = b_j(Y_t) - y_j(t),$$

in the sense of the measure evolutions of Section 14. The weights w_j in this transport equation are auxiliary weights for the moving labelled particles; they are not the Voronoi masses used to define the quantization energy. If one instead records the free-weight projection

$$\nu_{Y_t} = \sum_j a_j(Y_t)\delta_{y_j(t)},$$

then, formally,

$$\partial_t \nu_{Y_t} + \operatorname{div}(v_t \nu_{Y_t}) = \sum_j \dot{a}_j(Y_t)\delta_{y_j(t)}, \qquad v_t(y_j(t)) = \dot{y}_j(t).$$

Thus the free-weight quantizer evolves by a balance equation, not by pure transport. This is why the construction is intrinsically finite-dimensional: Voronoi cells, centroids and labels define the velocity, and a canonical extension to arbitrary measures is not obtained by replacing Y with the support of a measure. Indeed, any measure with dense support would have zero support-distance quantization error.

Mean-field limit and ultrafast diffusion. When the number m of codepoints is large, their empirical distribution $m^{-1}\sum_{j=1}^m \delta_{y_j}$ can be described by a site density σ. Thus $m\sigma(x)\mathrm{d}x$ represents the number of codepoints near x: regions where σ is large are quantized more finely. If $\alpha = \rho \mathrm{d}x$ and $r = p/d$, high-resolution quantization leads, up to a constant depending only on p and d, to the effective energy [251]

$$\mathcal{G}_\rho(\sigma) := \int_\Omega \rho(x)\sigma(x)^{-r}\mathrm{d}x, \qquad \int_\Omega \sigma \,\mathrm{d}x = 1,$$

where $r = 2/d$ for the quadratic cost considered here. Notice that σ describes the codepoints, whereas ρ is the source density being quantized. The discrete-to-continuum limit of the dynamics is rigorous in one dimension [97]. Related works study the resulting one-dimensional diffusion, perturbations of the hexagonal lattice in dimension two, and the weighted PDE in arbitrary dimension [286, 98, 287].

Formally, the $\mathcal{W}_2$-gradient flow of $\mathcal{G}_\rho$ is

$$\partial_t \sigma = -r \operatorname{div}\left(\sigma \nabla\left(\frac{\rho}{\sigma^{r+1}}\right)\right),$$

on a periodic domain or with no-flux boundary conditions. For a uniform source density $\rho = 1$, this becomes $\partial_t \sigma = -(r+1)\Delta(\sigma^{-r})$, an ultrafast-diffusion equation. Its negative power produces a strong redistribution where codepoints are scarce, and its stationary site density satisfies $\sigma_\star \propto \rho^{d/(d+p)}$.

Figure 6.4 shows the relaxed Lloyd flow in a two-dimensional toy problem. The sites are initialized from a red Gaussian mixture, but the centroids are computed with respect to a different blue Gaussian-mixture density. The motion is not a transport interpolation between these two mixtures; it is the descent of the quantization energy for the blue target, starting from a biased initial configuration.

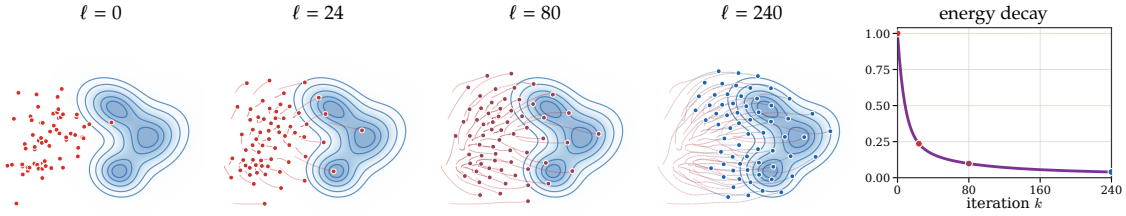


Figure 6.4: Relaxed Lloyd flow from a source Gaussian-mixture initialization toward a different target Gaussian-mixture density. The blue contours and shading show the target density α, while the colored disks are the moving codepoints initialized from the source mixture. The faint curves trace the labelled sites under the explicit-Euler Lloyd ODE. The right panel displays the relative quantization energy $\mathcal{F}(Y^{(\ell)})/\mathcal{F}(Y^{(0)})$, illustrating the monotone decay of the objective along the relaxed iterations.

Algorithm 6.5: Lloyd quantization

Input: Source measure α, initial codepoints $Y^{(0)} = (y_j^{(0)})_{j=1}^m$, squared Euclidean cost, tolerance tol, iteration budget $L \geqslant 1$.
Output: Codepoints $Y = (y_j)_{j=1}^m$.
Initialize: Set $d^{(0)} = +\infty$ and $\ell = 0$.
While $d^{(\ell)} >$ tol and $\ell < L$ **do**:
  Set $\ell \leftarrow \ell + 1$.
  Compute the tie-broken Voronoi partition: $\mathbb{V}_j(Y^{(\ell-1)}) = \left\{ x \,;\, j = \min \operatorname{argmin}_{1 \leqslant r \leqslant m} \|x - y_r^{(\ell-1)}\|^2 \right\}$.
  For each nonempty cell $\mathbb{V}_j$ **do**:
    $y_j^{(\ell)} = \dfrac{\int_{\mathbb{V}_j(Y^{(\ell-1)})} x \, \mathrm{d}\alpha(x)}{\int_{\mathbb{V}_j(Y^{(\ell-1)})} \mathrm{d}\alpha(x)}$.
  For each empty cell $\mathbb{V}_j$ **do**:
    Set $y_j^{(\ell)} = y_j^{(\ell-1)}$.
  Set $d^{(\ell)} = \max_j \|y_j^{(\ell)} - y_j^{(\ell-1)}\|$.
Return $Y^{(\ell)}$.

Figure 6.5 follows the associated Voronoi cells and generators through Lloyd iterations, making the decrease of the quantization energy visible.

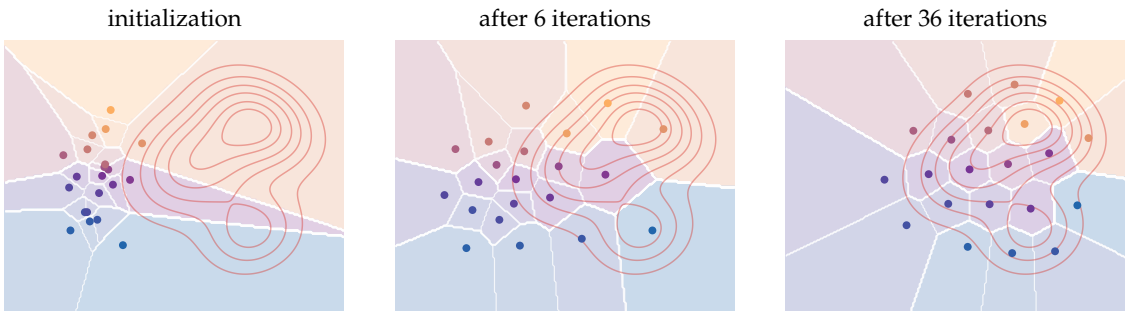


Figure 6.5: Lloyd quantization for the same continuous density and twenty-one initial sites as Figure 6.2. The red contours show the density α, while the colored disks are the current codepoints and have the same colors as their Voronoi cells. The iterations move the initially left-located sites toward the high-density region and reshape the cells according to centroidal Voronoi geometry.

Quantization with fixed equal weights. The free-mass formulation above optimizes the positions and the weights of the atoms. A different problem is obtained by prescribing the weights. In the equal-weight case, set

$$\nu_Y := \frac{1}{m} \sum_{j=1}^m \delta_{y_j}, \qquad \mathcal{F}_{\mathrm{eq}}(Y) := \frac{1}{2} \mathcal{W}_2^2(\alpha, \nu_Y),$$

and minimize only over the positions $Y = (y_j)_j$. Assume that the sites are distinct, that α has a density, and that it does not charge the boundaries of the semi-discrete optimal cells. Let $C_j(Y)$ be the Laguerre cell transported to y_j, so that $\alpha(C_j(Y)) = 1/m$, and define its centroid

$$\bar{x}_j(Y) := m \int_{C_j(Y)} x \mathrm{d}\alpha(x).$$

At differentiability points, the envelope theorem gives

$$\nabla_{y_j} \mathcal{F}_{\mathrm{eq}}(Y) = \frac{1}{m} \big(y_j - \bar{x}_j(Y) \big).$$

As long as the optimal matching between two nearby labelled configurations preserves labels, the $\mathcal{W}_2$ metric on equal-weight empirical measures induces the local particle metric $g_Y(U, V) = m^{-1} \sum_j \langle u_j, v_j \rangle$.

Hence the associated $\mathcal{W}_2$ gradient flow is the coupled system

$$\dot{y}_j(t) = \bar{x}_j(Y_t) - y_j(t), \qquad j = 1, \dots, m.$$

Equivalently, ν_{Y_t} satisfies a continuity equation with velocity $v_t(y_j(t)) = \bar{x}_j(Y_t) - y_j(t)$. This is the so-called finite-particle $\mathcal{W}_2$ gradient-flow viewpoint developed more systematically in Chapter 15; the fixed-weight cells are Laguerre cells rather than the free-mass Voronoi cells used by Lloyd's method.

Equal-weight quantization on the line. The following classical scalar quantization result gives the precise form of the inverse-CDF rule for equal-weight quadratic quantization; see [251]. The atoms are not exactly the midpoint quantiles in general; they are the averages of the quantile function over equal mass bins. Midpoint inverse-CDF samples are nevertheless asymptotically equivalent and are often the most convenient rule in numerical examples. For the uniform law on $[0,1]$, both rules coincide and give the regular grid $y_i = (i - 1/2)/m$.

Proposition 6.13 (One-dimensional equal-weight quantization). *Let $\alpha \in \mathcal{M}_+^1(\mathbb{R})$ have finite second moment and quantile function $Q = C_\alpha^{-1}$. For*

$$\mathcal{Q}_m^{\mathrm{eq}}(\alpha)^2 := \min_{y_1 \le \cdots \le y_m} \mathcal{W}_2^2(\alpha, \frac{1}{m}\sum_{i=1}^m \delta_{y_i}),$$

set $I_i = ((i-1)/m, i/m]$. Then the sorted minimizer is unique and its ith atom is

$$y_i^\star = m \int_{I_i} Q(u)\mathrm{d}u,$$

and

$$\mathcal{Q}_m^{\mathrm{eq}}(\alpha)^2 = \sum_{i=1}^m \int_{I_i} \left|Q(u) - y_i^\star\right|^2 \mathrm{d}u.$$

If $Q \in C^1([0,1])$, then

$$m^2 \mathcal{Q}_m^{\mathrm{eq}}(\alpha)^2 \longrightarrow \frac{1}{12}\int_0^1 |Q'(u)|^2 \,\mathrm{d}u.$$

Proof. After sorting the atoms, the quantile formula for $\mathcal{W}_2$ gives

$$\mathcal{W}_2^2(\alpha, \frac{1}{m}\sum_{i=1}^m \delta_{y_i}) = \sum_{i=1}^m \int_{I_i} |Q(u) - y_i|^2 \,\mathrm{d}u.$$

The minimization therefore decouples over the intervals I_i, and the best constant approximation of Q on I_i in L^2 is its average on I_i. Since Q is nondecreasing, these interval averages are nondecreasing, so they satisfy the sorting constraint. Strict convexity in each scalar y_i proves uniqueness and gives the displayed energy.

Denote by $\bar{Q}_{I_i} = m\int_{I_i} Q(u)\mathrm{d}u$ the interval average. If Q is C^1, set $h = 1/m$ and write $I_i = (a_i, a_i + h]$. Uniform Taylor expansion gives, for $v \in [0,1]$,

$$Q(a_i + hv) = Q(a_i) + hvQ'(a_i) + h\, r_i(v), \qquad \max_i \sup_{v\in[0,1]} |r_i(v)| \to 0.$$

Subtracting the average over $v \in [0,1]$ and integrating gives

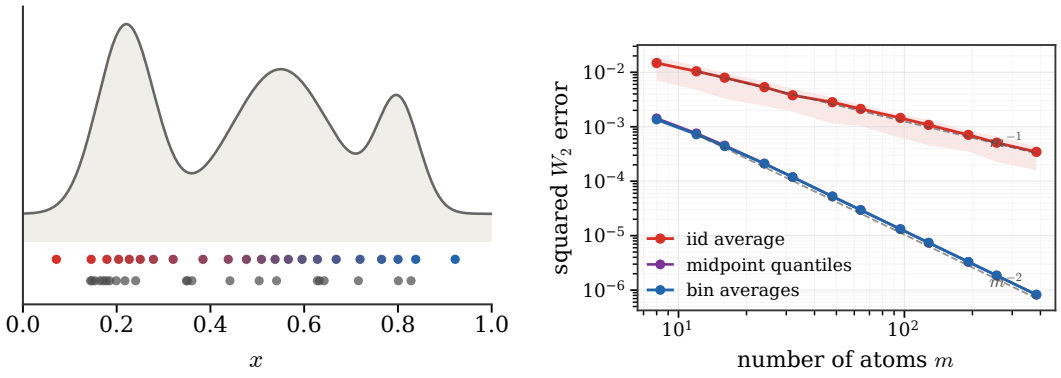
$$\int_{I_i} \left|Q(u) - \bar{Q}_{I_i}\right|^2 \mathrm{d}u = \frac{h^3}{12}\,|Q'(a_i)|^2 + o(h^3),$$

uniformly in i. Summing over i yields a Riemann sum for $\frac{1}{12}\int_0^1 |Q'(u)|^2\,\mathrm{d}u$. □

Thus the common deterministic rule $m^{-1}\sum_i \delta_{Q((i-1/2)/m)}$ should be read as a midpoint approximation of the optimal bin-average formula. It has the same leading squared error under the same smoothness assumptions. Indeed, orthogonal projection onto constants on each I_i gives

$$\int_{I_i} |Q(u) - z|^2 \mathrm{d}u = \int_{I_i} |Q(u) - y_i^\star|^2 \mathrm{d}u + \frac{1}{m}|z - y_i^\star|^2,$$

and $Q((i-1/2)/m) - y_i^\star = o(m^{-1})$ uniformly when $Q \in C^1([0,1])$. Random sampling has a different asymptotic regime.



equal-mass quantile particles deterministic versus i.i.d. decay

Figure 6.6: One-dimensional equal-weight quantization in quantile coordinates. Left: for a smooth positive density on $[0,1]$, the colored atoms are bin averages of the inverse CDF over equal quantile intervals, while the gray atoms show one i.i.d. empirical draw with the same number of particles. Right: expected squared $\mathcal{W}_2$ errors. The deterministic bin averages and midpoint quantiles follow the m^{-2} squared-error law, whereas i.i.d. empirical measures follow the slower m^{-1} expected squared-error law.

Proposition 6.14 (Quantile-process asymptotics for random placement). *Let $\alpha \in \mathcal{M}_+^1(\mathbb{R})$ have quantile $Q \in C^1([0,1])$, and let $\widehat{\alpha}_m = m^{-1}\sum_{i=1}^m \delta_{X_i}$ with X_i i.i.d. with law α. If B denotes the standard Brownian bridge on $[0,1]$, then*

$$m\,\mathcal{W}_2^2(\alpha,\widehat{\alpha}_m) \xrightarrow{\text{law}} \int_0^1 B(u)^2\,|Q'(u)|^2\,\mathrm{d}u,$$

and, in expectation,

$$m\,\mathbb{E}\big[\mathcal{W}_2^2(\alpha,\widehat{\alpha}_m)\big] \longrightarrow \int_0^1 u(1-u)\,|Q'(u)|^2\,\mathrm{d}u.$$

Proof. Let $\widehat{Q}_m$ be the empirical quantile function. The one-dimensional formula gives

$$\mathcal{W}_2^2(\alpha,\widehat{\alpha}_m) = \int_0^1 \big|\widehat{Q}_m(u) - Q(u)\big|^2\,\mathrm{d}u.$$

Write $X_i = Q(U_i)$ with U_i i.i.d. uniform on $[0,1]$, and let $\widehat{U}_m^{-1}$ be the empirical quantile function of the U_i. Then $\widehat{Q}_m = Q \circ \widehat{U}_m^{-1}$. The classical uniform quantile-process theorem [529, 68] gives

$$\sqrt{m}\,(\widehat{U}_m^{-1} - \mathrm{Id}) \xrightarrow{\text{law}} -B \quad \text{in } L^2(0,1).$$

Because Q' is uniformly continuous and $\widehat{U}_m^{-1} \to \mathrm{Id}$ uniformly in probability, the functional delta method gives

$$\sqrt{m}\,(\widehat{Q}_m - Q) \xrightarrow{\text{law}} -B\,Q' \quad \text{in } L^2(0,1).$$

Since B and $-B$ have the same law, the sign disappears in the squared L^2 norm. The continuous mapping theorem gives the distributional convergence. Standard fourth-moment bounds for the uniform quantile process, together with the boundedness of Q', give uniform integrability of the squared L^2 norms and hence convergence of expectations. Since $\mathbb{E}[B(u)^2] = u(1-u)$, Fubini's theorem gives the displayed expectation limit. □

Combining Propositions 6.13 and 6.14, one obtains a sharp contrast between optimal placement and random placement on the line. Deterministic equal-weight quantization has squared error of order m^{-2}, hence $\mathcal{W}_2$ error of order m^{-1}, while i.i.d. empirical sampling has expected squared error of order m^{-1}, hence root-mean-square $\mathcal{W}_2$ error of order $m^{-1/2}$. This one-dimensional comparison is consistent with the broader empirical OT sample-complexity theory [185, 220, 548].

Figure 6.6 illustrates both parts of this comparison: optimal atoms are uniform in quantile coordinates, and their error decays one power of m faster than the root-mean-square empirical error.

6.5 Wasserstein-1 norm

The $\mathcal{W}_1$ distance has an especially transparent dual: the admissible potentials are exactly 1-Lipschitz test functions. This makes $\mathcal{W}_1$ the meeting point between transport, PDE formulations and weak norms on signed measures.

c-transform for $\mathcal{W}_1$. Assume that d is a distance on $\mathcal{X} = \mathcal{Y}$ and take the ground cost $c(x, y) = d(x, y)$. We denote the Lipschitz constant of $f \in \mathcal{C}(\mathcal{X})$ by

Definition 6.15 (Lipschitz constant)**.** For a function $f : \mathcal{X} \to \mathbb{R}$ on a metric space $(\mathcal{X}, d)$, its Lipschitz constant is

$$\mathrm{Lip}(f) := \sup \left\{ \frac{|f(x) - f(y)|}{d(x, y)} \; ; \; x \neq y \right\}. \tag{6.20}$$

The function is 1-Lipschitz when $\mathrm{Lip}(f) \leqslant 1$.

Proposition 6.16 (c-transforms and 1-Lipschitz functions)**.** *Suppose $\mathcal{X} = \mathcal{Y}$ and $c(x, y) = d(x, y)$. Then there exists g such that $f = g^c$ if and only if $\mathrm{Lip}(f) \leqslant 1$. Furthermore, if $\mathrm{Lip}(f) \leqslant 1$, then $f^c = -f$.*

Proof. First suppose $f = g^c$ for some g. For $x, y \in \mathcal{X}$,

$$|f(x) - f(y)| = \left| \inf_z [d(x, z) - g(z)] - \inf_z [d(y, z) - g(z)] \right| \leqslant \sup_z |d(x, z) - d(y, z)| \leqslant d(x, y),$$

where the last inequality is the reverse triangle inequality. Thus $\mathrm{Lip}(f) \leqslant 1$.

If $\mathrm{Lip}(f) \leqslant 1$, then $f(x) \leqslant f(y) + d(x, y)$, so $d(x, y) - f(x) \geqslant -f(y)$ for all x and hence $f^c(y) \geqslant -f(y)$. Taking $x = y$ gives $f^c(y) \leqslant -f(y)$. Therefore $f^c = -f$. Applying the same property to $-f$ gives $(-f)^c = f$, so every 1-Lipschitz function is c-concave. □

By Proposition 6.16, a closed dual pair has the form $(f, -f)$ with $\mathrm{Lip}(f) \leqslant 1$. The Kantorovich dual therefore becomes the Kantorovich–Rubinstein formula

$$\mathcal{W}_1(\alpha, \beta) = \max_f \left\{ \int_{\mathcal{X}} f \, \mathrm{d}(\alpha - \beta) \; ; \; \mathrm{Lip}(f) \leqslant 1 \right\} =: \|\alpha - \beta\|_{W_1}. \tag{6.21}$$

This expression depends only on the signed measure $\xi = \alpha - \beta$. On compact $\mathcal{X}$, it therefore extends to every finite signed measure ξ with $\xi(\mathcal{X}) = 0$ by the same supremum and defines the Kantorovich–Rubinstein norm $\|\cdot\|_{W_1}$ [295]. Homogeneity and the triangle inequality follow directly from the supremum, while definiteness follows because Lipschitz functions separate finite Radon measures. On a noncompact pointed metric space, one uses 1-Lipschitz functions normalized at the base point and signed measures with finite first moment.

For a discrete signed measure $\alpha - \beta = \sum_{k=1}^{N} r_k \delta_{z_k}$ on distinct sites, with $\sum_k r_k = 0$, (6.21) becomes the finite-dimensional linear program

$$\mathcal{W}_1(\alpha, \beta) = \max_{(f_k)_k} \left\{ \sum_k f_k r_k \; ; \; \forall\, k, \ell, \; |f_k - f_\ell| \leqslant d(z_k, z_\ell) \right\}. \tag{6.22}$$

This linear program can be solved by generic interior-point or first-order methods. If N support points are involved, however, it still contains $O(N^2)$ Lipschitz constraints, mirroring the $O(nm)$ coupling variables of the original discrete Kantorovich LP; the dual formulation alone does not remove the all-pairs structure. The gain comes on structured metric spaces where the distance is generated locally: it is then enough to impose Lipschitz inequalities on neighboring pairs, because summing along paths recovers the constraints between arbitrary points. The one-dimensional ordered case is the first example; the graph-geodesic case is described later in Proposition 6.22. When $d(x, y) = |x - y|$ on $\mathbb{R}$, ordering the support points $z_1 \leqslant z_2 \leqslant \cdots$ reduces the constraints to neighboring pairs,

$$\mathcal{W}_1(\alpha, \beta) = \max_{(f_k)_k} \left\{ \sum_k f_k r_k \; ; \; \forall\, k, \; |f_{k+1} - f_k| \leqslant z_{k+1} - z_k \right\}.$$

In one dimension this is equivalent to the cumulative formula (2.12).

$\mathcal{W}_1$ on Euclidean spaces. In Euclidean space, the global Lipschitz condition has a local differential form, and its dual variable is a flux. This turns the static coupling problem into a divergence-constrained convex problem.

Let $\alpha, \beta \in \mathcal{M}_+^1(\mathbb{R}^d)$ have finite first moments and set $\xi = \alpha - \beta$. Rademacher's theorem identifies 1-Lipschitz functions with locally Sobolev functions whose Euclidean gradient has norm at most one almost everywhere. Hence

$$\mathcal{W}_1(\alpha, \beta) = \sup_{f \in W^{1,\infty}_{\mathrm{loc}}(\mathbb{R}^d)} \left\{ \int_{\mathbb{R}^d} f \, \mathrm{d}\xi \; ; \; \|\nabla f\|_{L^\infty(\mathbb{R}^d)} \leqslant 1 \right\}. \tag{6.23}$$

The correct flux space is generally not $L^1(\mathbb{R}^d; \mathbb{R}^d)$. For instance, transporting one Dirac mass to another produces a flux concentrated on a segment. The appropriate norm measures both diffuse and singular vector-valued fluxes.

Definition 6.17 (Total variation norm of a vector-valued measure)**.** Let $\mathcal{M}(\mathbb{R}^d; \mathbb{R}^d)$ denote the finite $\mathbb{R}^d$-valued Radon measures. For $s \in \mathcal{M}(\mathbb{R}^d; \mathbb{R}^d)$, define

$$\|s\|_{\mathrm{TV}} := \sup_{\zeta \in C_c(\mathbb{R}^d; \mathbb{R}^d)} \left\{ \int_{\mathbb{R}^d} \langle \zeta, \, \mathrm{d}s \rangle \; ; \; \sup_{x \in \mathbb{R}^d} \|\zeta(x)\|_2 \leqslant 1 \right\}.$$

Remark 6.18 (Absolutely continuous vector measures). If $s = w \, \mathrm{d}x$ with $w \in L^1(\mathbb{R}^d; \mathbb{R}^d)$, then

$$\|s\|_{\mathrm{TV}} = \int_{\mathbb{R}^d} \|w(x)\|_2 \mathrm{d}x.$$

Thus the total variation norm extends the usual L^1 norm of a vector field while also assigning a finite cost to singular fluxes.

For $s \in \mathcal{M}(\mathbb{R}^d; \mathbb{R}^d)$, its divergence is defined distributionally by

$$\langle \operatorname{div} s, \varphi \rangle := - \int_{\mathbb{R}^d} \langle \nabla \varphi, \, \mathrm{d}s \rangle, \qquad \varphi \in C_c^1(\mathbb{R}^d).$$

This yields the Beckmann problem with signed source $\alpha - \beta$.

Definition 6.19 (Beckmann problem)**.** For $\alpha, \beta \in \mathcal{M}_+^1(\mathbb{R}^d)$, define

$$\mathrm{Beck}(\alpha, \beta) := \inf_{s \in \mathcal{M}(\mathbb{R}^d; \mathbb{R}^d)} \{ \|s\|_{\mathrm{TV}} \; ; \; \operatorname{div} s = \alpha - \beta \}.$$

Proposition 6.20 (Euclidean Beckmann formula)**.** *For $\alpha, \beta \in \mathcal{M}_+^1(\mathbb{R}^d)$ with finite first moments,*

$$\mathcal{W}_1(\alpha, \beta) = \mathrm{Beck}(\alpha, \beta). \tag{6.24}$$

Proof. Let s be feasible. For a smooth 1-Lipschitz test function,

$$\int f \, \mathrm{d}(\alpha - \beta) = - \int \langle \nabla f, \, \mathrm{d}s \rangle \leqslant \|s\|_{\mathrm{TV}}.$$

Truncation followed by mollification approximates normalized Lipschitz functions by smooth test functions without increasing the gradient bound asymptotically. Equation (6.23) therefore gives $\mathcal{W}_1(\alpha, \beta) \leqslant \|s\|_{\mathrm{TV}}$.

Conversely, let π be an optimal coupling and define a vector measure by its action on $\zeta \in C_c(\mathbb{R}^d; \mathbb{R}^d)$:

$$\int \langle \zeta(z), \, \mathrm{d}s(z) \rangle := \int_{\mathbb{R}^d \times \mathbb{R}^d} \int_0^1 \langle \zeta((1-t)x + ty), \, y - x \rangle \mathrm{d}t \, \mathrm{d}\pi(x, y).$$

For $\varphi \in C_c^1(\mathbb{R}^d)$, the fundamental theorem of calculus gives

$$\langle \operatorname{div} s, \varphi \rangle = - \int \int_0^1 \langle \nabla \varphi((1-t)x + ty), \, y - x \rangle \mathrm{d}t \, \mathrm{d}\pi(x, y) = \int \varphi \, \mathrm{d}(\alpha - \beta).$$

Moreover, $\|s\|_{\mathrm{TV}} \leqslant \int \|x-y\|_2 \mathrm{d}\pi(x,y) = \mathcal{W}_1(\alpha,\beta)$. Together with the first inequality, this proves equality and attainment. □

This is the Beckmann formulation [31, 469]. The vector measure s describes the oriented amount of mass crossing each region. On open sets disjoint from the source and target mass, $\operatorname{div} s = 0$, which is the local conservation law.

Finite-element or finite-volume discretizations of (6.24) produce nonsmooth convex programs. The same formulation extends to complete Riemannian manifolds by replacing straight segments with minimizing geodesics and using tangent-valued Radon measures together with the Riemannian divergence. If geodesics are nonunique, one chooses a measurable minimizing geodesic for almost every transported pair.

Graph distances and Beckmann flows. Finite graphs give a simple discrete instance where a metric is generated by local moves, so the all-pairs Lipschitz constraints collapse to edge constraints.

Definition 6.21 (Graph geodesic distance)**.** Let $G = (V, E)$ be a connected finite graph with positive edge lengths $(\ell_e)_{e\in E}$. The graph geodesic distance between two vertices is

$$d_G(i,j) = \min_{\gamma : i \rightsquigarrow j} \sum_{e\in\gamma} \ell_e.$$

The minimum is over all paths γ joining i to j.

This graph distance turns $\mathcal{W}_1$ into a finite-dimensional flow problem.

Proposition 6.22 ($\mathcal{W}_1$ and Beckmann flow on a graph)**.** *Let $G = (V, E)$ be a connected finite graph with positive edge lengths $(\ell_e)_{e\in E}$ and graph geodesic distance d_G. For two probability vectors a, b on V, set $r = \mathrm{a} - \mathrm{b}$ and orient each edge $e = (i, j)$. If*

$$(\nabla_G f)_e = f_j - f_i, \qquad \operatorname{div}_G = -\nabla_G^*$$

are the finite-difference gradient and its negative adjoint, then a positive flow on the oriented edge $i \to j$ has positive divergence at i and negative divergence at j. With this convention,

$$\mathcal{W}_{1,G}(\mathrm{a},\mathrm{b}) = \max_f \left\{ \sum_{i\in V} f_i r_i \;;\; |f_i - f_j| \leqslant \ell_e \quad \forall e = (i,j) \right\} = \min_m \left\{ \sum_{e\in E} \ell_e |m_e| \;;\; \operatorname{div}_G m = r \right\}.$$

The vector m_e is an oriented edge flow, and the constraint $\operatorname{div}_G m = r$ is conservation of mass at each vertex.

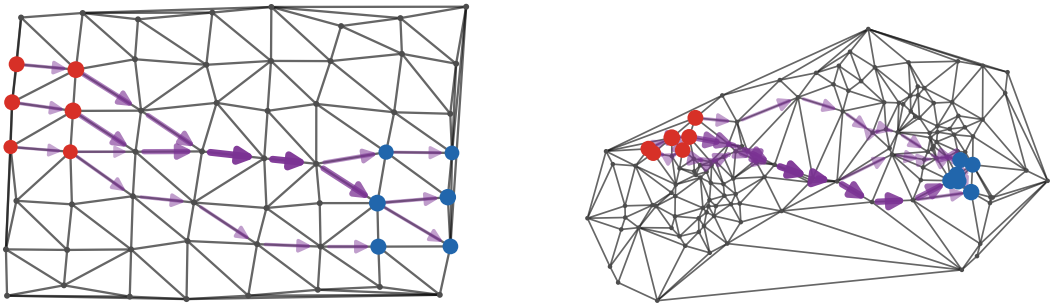
Proof. The edge constraints imply, by summing along every path and then minimizing over paths, that $|f_i - f_j| \leqslant d_G(i,j)$ for all vertices. Conversely, any 1-Lipschitz function for d_G satisfies $|f_i - f_j| \leqslant d_G(i,j) \leqslant \ell_e$ on every edge. The first equality is therefore the Kantorovich–Rubinstein formula on the metric space (V, d_G).

For the second equality, write the graph Beckmann problem and dualize its equality constraint with a potential f:

$$\inf_m \sum_e \ell_e |m_e| + \sup_f \sum_i f_i (r_i - (\operatorname{div}_G m)_i).$$

Using $\operatorname{div}_G = -\nabla_G^*$, the coupling term is $\sum_e m_e (\nabla_G f)_e$. The minimization over each scalar flow m_e is finite exactly when $|(\nabla_G f)_e| \leqslant \ell_e$, and is then equal to zero. The dual problem is therefore precisely the graph Lipschitz dual above. Strong duality holds because this is a finite-dimensional linear program with a nonempty feasible set: connectedness and $\sum_i r_i = 0$ allow the signed surplus to be routed along paths. This proves the graph Beckmann formula. □

Figure 6.7 shows the optimal edge flux on both a quasi-regular and a nonuniform Delaunay graph, emphasizing that the Beckmann variables live only on graph edges.



quasi-regular Delaunay graph denser nonuniform Delaunay graph

Figure 6.7: Graph Beckmann formulation of $\mathcal{W}_1$ on two Delaunay graphs $G = (V, E)$. The left panel uses a quasi-regular vertex set, while the right panel uses a denser nonuniform vertex set sampled from a two-component Gaussian mixture. Red and blue disks encode the positive and negative parts of $r = a - b$, localized on six vertices each. The darker graph edges show the triangulation, while the violet arrows display the optimal signed edge flow m: orientation gives the sign, width is proportional to $\sqrt{|m_e|}$, and the flow satisfies $\operatorname{div}_G m = r$.

Remark 6.23 (Sparse LP and network simplex). Let $N = |V|$ and $M = |E|$. The graph Beckmann problem is a linear program, but it is much smaller than the dense Kantorovich LP on the same vertex set. Indeed, writing $m = m^+ - m^-$ gives

$$\min_{m^+, m^- \geqslant 0} \sum_{e \in E} \ell_e (m_e^+ + m_e^-) \quad \text{subject to} \quad \operatorname{div}_G(m^+ - m^-) = r.$$

This formulation has $2M$ nonnegative variables and $N-1$ independent balance constraints, whereas the standard transport LP between two measures on V has N^2 coupling variables and $2N-1$ independent marginal constraints. For sparse geometric graphs, such as planar Delaunay graphs where typically $M = O(N)$, the graph formulation is therefore linear-size rather than quadratic-size.
The same LP is a minimum-cost transshipment problem. Replace each undirected edge $\{i, j\}$ by the two directed arcs $i \to j$ and $j \to i$, both with cost ℓ_e, and impose the node balances $\sum_j u_{i,j} - \sum_j u_{j,i} = r_i$. This is exactly the setting of the network simplex method: a basis is a spanning tree, a pivot adds one non-tree arc, creates a unique cycle, sends flow along this cycle, and updates the node potentials and reduced costs [51, 409]. A basic implementation needs $O(M)$ work to price all arcs and $O(N)$ work to update the tree at each pivot, hence $O(PM)$ arithmetic operations for P pivots on a sparse graph. The pivot count P depends on the rule and can be large in worst-case simplex analyses, but network-simplex variants and general minimum-cost-flow algorithms give polynomial guarantees in N and M; in practice, this edge-based formulation is often far cheaper than solving the dense N^2-variable transport LP.

This graph formulation is the transshipment version of $\mathcal{W}_1$. It is the natural discrete analogue of (6.24): gradients are edge differences, divergences are incidence-matrix balances, and geodesic distance is the shortest-path length. It can be solved by min-cost flow methods on sparse graphs, and entropic or KL-projection variants lead to flow-Sinkhorn algorithms for graph $\mathcal{W}_1$ [31, 428].

CHAPTER 7

Divergences and Dual Norms

This chapter compares OT with divergence-based and adversarial ways of measuring discrepancy. The central issue is topological: φ-divergences are cheap but strong, while dual norms and GAN objectives can be weak enough to compare singular measures. The discussion connects classical information divergences [147, 4] with modern integral probability metrics and generative modeling [499, 247, 19].

7.1 Dual norms (Integral Probability Metrics)

This section isolates the test-function viewpoint behind weak discrepancies. Dual norms generalize the $\mathcal{W}_1$ test-function principle. They are useful in statistics because they compare distributions by restricting the discriminator class.

Integral probability metrics. Formulation (6.23) is a special case of a dual norm. This viewpoint provides a way to design "weak" discrepancies by testing signed differences of measures against a controlled class of functions.

Definition 7.1 (Dual seminorm and integral probability metric)**.** Let B be a nonempty symmetric convex class of measurable functions that are integrable against the measures under consideration. For a signed measure ξ, define the extended dual seminorm

$$\|\xi\|_B := \sup_f \left\{ \int_{\mathcal{X}} f(x) \mathrm{d}\xi(x) \; ; \; f \in B \right\}. \tag{7.1}$$

It is a norm when it is finite and B separates signed measures. When applied to $\alpha - \beta$ for probability measures, it is called an integral probability metric (IPM); it is a genuine metric precisely when B separates probability measures.

Symmetry makes the supremum equal to $\sup_{f \in B} |\int f \mathrm{d}\xi|$, while convexity makes B a natural unit ball. The choice of B determines both the topology and the statistical behavior of the discrepancy; see [500, 499, 501].

Example 7.2 (Total variation). As recalled in Definition 2.7 and Proposition 2.8, total variation is the dual norm associated with the unit ball of continuous functions

$$B = \{ f \in C(\mathcal{X}) \; ; \; \|f\|_\infty \leqslant 1 \}.$$

Total variation is the canonical nontrivial example of a discrepancy that is both a φ-divergence and a dual norm; see [499].

Example 7.3 ($\mathcal{W}_1$ norm). On zero-mass signed measures with finite first moment, the Kantorovich–Rubinstein norm underlying $\mathcal{W}_1$, as defined in (6.23), is the extended dual seminorm (7.1) associated with

$$B = \{ f \; ; \; \mathrm{Lip}(f) \leqslant 1 \}$$

the set of 1-Lipschitz functions.

Example 7.4 (Flat norm and Dudley metric). If B is uniformly bounded and separates measures, then $\|\cdot\|_B$ is a finite norm on the whole space $\mathcal{M}(\mathcal{X})$ of finite signed measures. This is not the case for $\mathcal{W}_1$: zero total mass, $\xi(\mathcal{X}) = 0$, is necessary, and on an unbounded space one also needs a finite first moment. If the mass is nonzero, $\|\xi\|_B = +\infty$ because constants belong to the Lipschitz ball. This is remedied by imposing a bound on the value of the potential f, which leads for instance to the flat norm,

$$B = \{f \,;\, \mathrm{Lip}(f) \leqslant 1 \quad \text{and} \quad \|f\|_\infty \leqslant 1\}. \tag{7.2}$$

On compact metric spaces, it metrizes weak convergence of finite nonnegative measures, and more generally weak-* convergence on every total-variation-bounded subset of $\mathcal{M}(\mathcal{X})$. The finite-dimensional version is obtained from the usual $\mathcal{W}_1$ dual linear program by adding the box constraints $|f_k| \leqslant 1$. The flat norm is sometimes called the "Kantorovich–Rubinstein" norm [269] and has been used as a fidelity term for inverse problems in imaging [334]. The flat norm is equivalent to the bounded-Lipschitz, or Dudley, metric, whose test class is

$$B = \{f \,;\, \mathrm{Lip}(f) + \|f\|_\infty \leqslant 1\}.$$

On a Euclidean domain, $\mathrm{Lip}(f) = \|\nabla f\|_\infty$ for differentiable f.

Figure 7.1 compares the optimal test functions selected by several integral probability metrics for the same pair of one-dimensional densities.

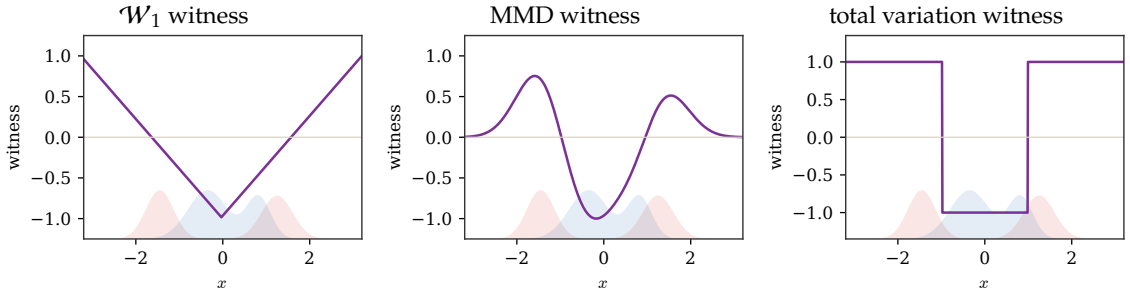


Figure 7.1: Dual witnesses for integral probability metrics. The red and blue curves are two one-dimensional probability densities and the violet curve is a normalized optimal dual witness $f^\star_{\alpha,\beta}$ for the IPM variational problem (7.1). $\mathcal{W}_1$ restricts the slope through Kantorovich–Rubinstein duality (6.23), MMD restricts the RKHS norm as in Proposition 7.10, and total variation can saturate pointwise and therefore reacts sharply to signed density differences.

The following proposition gives a useful compact-space criterion. The dual ball should be rich enough to approximate continuous observables, but compact enough for weak convergence to imply uniform convergence over the discriminator class.

Proposition 7.5 (Metrization by dual norms)**.** *Assume that $\mathcal{X}$ is compact, that $B = -B$, and that the measures considered are probability measures.*

1. *If every function in $C(\mathcal{X})$ can be uniformly approximated by elements of* $\mathrm{Span}(B)$*, then $\|\alpha_n - \alpha\|_B \to 0$ implies $\alpha_n \rightharpoonup \alpha$.*

2. *If $B \subset C(\mathcal{X})$ is compact for $\|\cdot\|_\infty$, then $\alpha_n \rightharpoonup \alpha$ implies $\|\alpha_n - \alpha\|_B \to 0$.*

Proof. For the first implication, $\|\alpha_n - \alpha\|_B \to 0$ and the symmetry of B imply

$$|\langle f,\, \alpha_n - \alpha\rangle| \leqslant \|\alpha_n - \alpha\|_B \quad \text{for} \quad f \in B.$$

If $h = \sum_{j=1}^J c_j f_j \in \mathrm{Span}(B)$, then $|\langle h,\, \alpha_n - \alpha\rangle| \leqslant (\sum_{j=1}^J |c_j|)\|\alpha_n - \alpha\|_B$, so integrals converge for every $h \in \mathrm{Span}(B)$. Let $u \in C(\mathcal{X})$ and choose such an h with $\|u - h\|_\infty \leqslant \eta$. Since α_n and α are probabilities, $|\langle u,\, \alpha_n - \alpha\rangle| \leqslant |\langle h,\, \alpha_n - \alpha\rangle| + 2\eta$. Taking the limsup as $n \to \infty$ and then letting $\eta \to 0$ gives $\langle u,\, \alpha_n\rangle \to \langle u,\, \alpha\rangle$ for all $u \in C(\mathcal{X})$, which is weak convergence.

For the second implication, assume $\alpha_n \rightharpoonup \alpha$ and choose a subsequence $(\alpha_{n_k})_k$ such that $\|\alpha_{n_k} - \alpha\|_B \to \limsup_n \|\alpha_n - \alpha\|_B$. Since B is compact and the map $f \mapsto \langle f,\, \alpha_{n_k} - \alpha\rangle$ is continuous on B, the supremum is attained by some $f_{n_k} \in B$. Extracting a further subsequence if needed, $f_{n_k} \to f$ uniformly for some $f \in B$. Then

$$\langle f_{n_k},\, \alpha_{n_k} - \alpha\rangle = \langle f,\, \alpha_{n_k} - \alpha\rangle + \langle f_{n_k} - f,\, \alpha_{n_k}\rangle - \langle f_{n_k} - f,\, \alpha\rangle.$$

The first term tends to zero by weak convergence and the last two by uniform convergence. Hence every limsup subsequence has limit zero, proving $\|\alpha_n - \alpha\|_B \to 0$. □

Corollary 7.6 (Wasserstein metrizes weak convergence)**.** *On a compact metric space, $\mathcal{W}_p$ metrizes weak convergence on probability measures for every $p \geqslant 1$.*

Proof. For $p = 1$, take $B = \{f \ ; \ \mathrm{Lip}(f) \leqslant 1\}$. The span of B contains all Lipschitz functions, and Lipschitz functions are dense in $C(\mathcal{X})$ on compact metric spaces. This gives the implication $\mathcal{W}_1(\alpha_n, \alpha) \to 0 \Rightarrow \alpha_n \rightharpoonup \alpha$ by Proposition 7.5.

Conversely, constants do not change the pairing with $\alpha_n - \alpha$. Fix $x_0 \in \mathcal{X}$ and normalize potentials by $f(x_0) = 0$. The normalized unit Lipschitz ball is uniformly bounded by $\mathrm{diam}(\mathcal{X})$ and equicontinuous, hence compact in $\|\cdot\|_\infty$ by Arzelà–Ascoli. Proposition 7.5 gives $\mathcal{W}_1(\alpha_n, \alpha) \to 0$. Proposition 4.17 shows that all $\mathcal{W}_p$ distances induce the same topology on a compact space, so the result follows for every $p \geqslant 1$. □

7.2 Dual RKHS Norms and Maximum Mean Discrepancies

Kernel methods turn probability measures into mean elements of a reproducing kernel Hilbert space (RKHS). The resulting Hilbertian dual seminorms are quadratic discrepancies, handled with Euclidean geometry while retaining a weak test-function interpretation. We first recall the positivity assumptions under which the quadratic form on signed measures is nonnegative.

Definition 7.7 (Positive and conditionally definite kernels)**.** A symmetric function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is positive definite if for every $n \geqslant 1$, every $x_1, \dots, x_n \in \mathcal{X}$ and every $r \in \mathbb{R}^n$,

$$\sum_{i,j=1}^{n} r_i r_j k(x_i, x_j) \geqslant 0. \tag{7.3}$$

It is conditionally positive definite of order one if the same inequality is required only for zero-sum vectors, $\langle r, \mathbb{1}_n \rangle = 0$. It is conditionally negative definite of order one if the reverse inequality holds for every such vector, or equivalently if $-k$ is conditionally positive definite.

Here "positive definite" is used in the standard kernel-theory sense of positive semidefinite; strict positivity is an additional property. This distinction matters because a non-strict kernel can induce a degenerate discrepancy.

The conditional notions are the right ones for probability distances, because one applies quadratic forms to signed measures $\xi = \alpha - \beta$ of total mass zero. Adding a term of the form $a(x) + a(y)$ to a kernel does not change $\iint k(x, y)\mathrm{d}\xi(x)\mathrm{d}\xi(y)$ on such measures. Many natural distance kernels are conditionally negative definite, while their negatives are the conditionally positive kernels used to define MMDs.

Example 7.8 (Riesz, energy and Matérn-type kernels). On $\mathbb{R}^d$, translation-invariant kernels are most transparent in Fourier variables. The Riesz family associated with $(-\Delta)^{-s}$ has multiplier $\|\omega\|^{-2s}$ and defines a nonnegative quadratic form on zero-mass measures for which the low-frequency singularity is integrable; this is the kernel counterpart of classical Riesz potentials [43]. The energy distance corresponds to the conditionally positive kernel $k(x, y) = -\|x - y\|$, whose Fourier multiplier is proportional to $\|\omega\|^{-(d+1)}$; for $\xi = \alpha - \beta$, $-\iint \|x - y\|\mathrm{d}\xi(x)\mathrm{d}\xi(y)$ is exactly the customary squared energy distance $2\mathbb{E}\|X - Y\| - \mathbb{E}\|X - X'\| - \mathbb{E}\|Y - Y'\|$; only its equivalent Fourier representation carries a dimension-dependent constant [479, 513].
Shifted kernels replace $(-\Delta)^{-s}$ by $(-\Delta + \lambda \mathrm{Id})^{-s}$ with $\lambda > 0$. Their Fourier multiplier $(\|\omega\|^2 + \lambda)^{-s}$ is bounded at the origin, hence the kernel is positive definite without imposing zero mass. These are Matérn kernels; in closed form they are radial and involve a modified Bessel function [549]. The Laplacian kernel $e^{-\|x-y\|/\sigma}$ is a low-smoothness Matérn example, while the Gaussian kernel $e^{-\|x-y\|^2/(2\sigma^2)}$ is the infinite-smoothness limit after the usual rescaling of the Matérn smoothness parameter.

Definition 7.9 (Kernel seminorm and MMD)**.** Let k be positive definite. More generally, let k be conditionally positive definite and restrict attention to signed measures of total mass zero. For a

signed measure ξ with finite kernel energy, the associated seminorm is

$$\|\xi\|_k^2 := \iint_{X\times X} k(x,y)\mathrm{d}\xi(x)\mathrm{d}\xi(y). \tag{7.4}$$

For two probability measures, the maximum mean discrepancy associated with k is

$$\mathrm{MMD}_k(\alpha,\beta) := \|\alpha-\beta\|_k.$$

It is a genuine distance when the kernel is characteristic, meaning that $\|\alpha-\beta\|_k = 0$ implies $\alpha=\beta$.

These seminorms are usually called maximum mean discrepancies in statistics and machine learning [253, 392], and kernel norms in shape analysis [278]. For a positive-definite kernel, if X, X' are independent with law α, then $\|\alpha\|_k^2 = \mathbb{E}_{X,X'}(k(X,X'))$, whenever this expression is finite. A conditionally positive kernel need not itself be an RKHS kernel; after fixing $x_0 \in \mathcal{X}$, however,

$$\widetilde{k}(x,y) = k(x,y) - k(x,x_0) - k(x_0,y) + k(x_0,x_0)$$

is positive definite and induces the same quadratic form on zero-mass measures.

Proposition 7.10 (Kernel seminorm as an RKHS dual norm)**.** *Let $\mathcal{H}$ be the RKHS with reproducing kernel k, and assume that the kernel mean embedding $m_\xi := \int k(x,\cdot)\mathrm{d}\xi(x)$ is well-defined for the signed measure ξ. Then $\|\xi\|_k = \sup_{\|h\|_{\mathcal{H}}\leqslant 1}\int h(x)\mathrm{d}\xi(x)$, so $\|\cdot\|_k$ is the dual seminorm in the sense of (7.1) associated with the RKHS unit ball.*

Proof. By the reproducing property, $\int h(x)\mathrm{d}\xi(x) = \langle h, \int k(x,\cdot)\mathrm{d}\xi(x)\rangle_{\mathcal{H}} = \langle h, m_\xi\rangle_{\mathcal{H}}$. Cauchy–Schwarz gives $\sup_{\|h\|_{\mathcal{H}}\leqslant 1}\int h\mathrm{d}\xi = \|m_\xi\|_{\mathcal{H}}$. Finally, $\|m_\xi\|_{\mathcal{H}}^2 = \iint k(x,y)\mathrm{d}\xi(x)\mathrm{d}\xi(y)$, which is exactly (7.4). □

Proposition 7.11 (Universal kernels metrize weak convergence)**.** *Assume that $\mathcal{X}$ is compact and that the RKHS generated by the continuous kernel k is dense in $C(\mathcal{X})$ for the uniform norm. Then*

$$\mathrm{MMD}_k(\alpha_n,\alpha)\to 0 \quad\Longleftrightarrow\quad \alpha_n \rightharpoonup \alpha$$

for probability measures on $\mathcal{X}$.

Proof. If $\mathrm{MMD}_k(\alpha_n,\alpha)\to 0$, then for every $g\in\mathcal{H}$,

$$\left|\int g\mathrm{d}(\alpha_n-\alpha)\right| \leqslant \|g\|_{\mathcal{H}}\,\mathrm{MMD}_k(\alpha_n,\alpha)\to 0.$$

For any $h\in C(\mathcal{X})$ and any $\eta>0$, choose $g\in\mathcal{H}$ with $\|h-g\|_\infty\leqslant\eta$. Since α_n and α are probabilities,

$$\left|\int h\,\mathrm{d}(\alpha_n-\alpha)\right| \leqslant 2\eta + \left|\int g\,\mathrm{d}(\alpha_n-\alpha)\right|,$$

and the last term tends to zero. This proves weak convergence. Conversely, if $\alpha_n\rightharpoonup\alpha$, then $\alpha_n\otimes\alpha_n$, $\alpha_n\otimes\alpha$ and $\alpha\otimes\alpha$ converge weakly on the compact product space. Applying this to the continuous bounded function k in the identity

$$\mathrm{MMD}_k(\alpha_n,\alpha)^2 = \iint k\,\mathrm{d}\alpha_n\mathrm{d}\alpha_n - 2\iint k\,\mathrm{d}\alpha_n\mathrm{d}\alpha + \iint k\,\mathrm{d}\alpha\mathrm{d}\alpha$$

gives convergence to zero. □

We refer to [44, 278, 480] for more details on RKHS functional spaces. The density property used above has a standard name.

Definition 7.12 (Universal kernel)**.** A continuous positive-definite kernel k on a compact space $\mathcal{X}$ is universal if finite sums of sections $\sum_{i=1}^n a_i k(x_i,\cdot)$ are dense in $C(\mathcal{X})$ for the uniform norm.

Remark 7.13 (Spectral characterization of universality). For a bounded continuous translation-invariant kernel on $\mathbb{R}^d$, $k(x,y)=k_0(x-y)$, a standard sufficient condition is that its Bochner spectral

measure have support equal to all of $\mathbb{R}^d$; under the usual C_0 assumptions, this characterizes C_0-universality [501, 500].

The preceding discrepancies are widely used as sample-based criteria, both for testing whether two populations agree and for evaluating generative models.

Example 7.14 (Two-sample testing and generative-model evaluation). Given samples $X_1, \dots, X_n \sim \alpha$ and $Y_1, \dots, Y_m \sim \beta$, a two-sample test evaluates a statistic $D(\hat{\alpha}_n, \hat{\beta}_m)$ under the null hypothesis $H_0 : \alpha = \beta$. Wasserstein distances provide geometric choices of D, whereas MMD and the energy distance provide kernel and negative-type alternatives [448, 253, 513]. In generative-model evaluation, the Fréchet Inception Distance embeds the data in a neural feature space, fits a Gaussian to each empirical feature distribution, and computes the squared $\mathcal{W}_2$ distance between these Gaussian laws: the squared distance between their empirical means plus the Bures covariance term of Proposition 2.54 [276]. The Kernel Inception Distance instead estimates a squared MMD, typically with a polynomial kernel, in the same feature space [60]. In both settings, the geometry encoded by D must be distinguished from the finite-sample calibration of $D(\hat{\alpha}_n, \hat{\beta}_m)$; Sections 10.2 and 10.3 analyze the resulting rates, bias and fluctuations.

In the special case where α is a discrete measure, one thus has the simple expression

$$\|\alpha\|_k^2 = \sum_{i=1}^{n} \sum_{i'=1}^{n} a_i a_{i'} k_{i,i'} = \langle ka, a \rangle \quad \text{where} \quad k_{i,i'} := k(x_i, x_{i'}).$$

In particular, when $\alpha = \sum_{i=1}^n a_i \delta_{x_i}$ and $\beta = \sum_{i=1}^n b_i \delta_{x_i}$ are supported on the same set of points, $\|\alpha - \beta\|_k^2 = \langle k(a-b), a-b \rangle$. This is a Euclidean seminorm on the simplex, and it is a genuine norm on differences of histograms exactly when $r^\top k r > 0$ for every nonzero r satisfying $\langle r, \mathbb{1}_n \rangle = 0$. To compute the discrepancy between two discrete measures, one can use

$$\|\alpha - \beta\|_k^2 = \sum_{i,i'} a_i a_{i'} k(x_i, x_{i'}) + \sum_{j,j'} b_j b_{j'} k(y_j, y_{j'}) - 2 \sum_{i,j} a_i b_j k(x_i, y_j). \tag{7.5}$$

7.3 φ-divergences

This section develops divergences based on pointwise density ratios. They are computationally simple and statistically classical, but on nondiscrete spaces they generally induce a topology much stronger than weak convergence and do not see small spatial displacements between mutually singular measures.

Definition by density ratios. We now consider a radically different class of discrepancies. On a common discrete support they cost only $O(n)$ to evaluate, but on a continuous space they generally fail to metrize weak convergence. Bregman divergences provide another construction from convex functionals. They should not be conflated with the density-ratio divergences introduced below; their precise relation is discussed after the main examples.

Definition 7.15 (Entropy function)**.** A function $\varphi : \mathbb{R} \to \mathbb{R} \cup \{+\infty\}$ is an entropy function if it is proper, lower semicontinuous and convex, with $\operatorname{dom} \varphi \subset [0, +\infty)$ and $\operatorname{dom} \varphi \cap (0, +\infty) \neq \emptyset$. Its recession slope is

$$\varphi'_\infty = \lim_{s \to +\infty} \frac{\varphi(s)}{s} \in \mathbb{R} \cup \{+\infty\}.$$

If $\varphi'_\infty = \infty$, then φ grows faster than any linear function and φ is said to be *superlinear*. Any entropy function φ induces a φ-divergence (also known as Ciszár divergence [147, 4] or f-divergence) as follows.

Definition 7.16 (φ-Divergences)**.** Let φ be an entropy function and let $\alpha, \beta \in \mathcal{M}_+(\mathcal{X})$. Write the Lebesgue decomposition of α with respect to β as $\alpha = (d\alpha/d\beta)\beta + \alpha^\perp$, where $\alpha^\perp \perp \beta$. The associated divergence is

$$\mathcal{D}_\varphi(\alpha|\beta) := \int_{\mathcal{X}} \varphi\left(\frac{d\alpha}{d\beta}\right) d\beta + \varphi'_\infty \alpha^\perp(\mathcal{X}), \tag{7.6}$$

with the convention $0 \cdot (+\infty) = 0$. It is extended by $+\infty$ whenever either argument is not a nonnegative measure.

The additional term $\varphi'_\infty \alpha^\perp(\mathcal{X})$ in (7.6) is the recession contribution of the perspective functional. It gives the weak-* lower-semicontinuous extension of the density-ratio integral when singular mass appears. This is essential for entropies with linear growth at infinity, such as the absolute value (7.10) defining the TV norm. If φ has superlinear growth, e.g.the usual entropy (7.9), then $\varphi'_\infty = +\infty$ so that $\mathcal{D}_\varphi(\alpha|\beta) = +\infty$ if α does not have a density with respect to β.

In the discrete setting, assuming

$$\alpha = \sum_i \mathrm{a}_i \delta_{x_i} \quad \text{and} \quad \beta = \sum_i \mathrm{b}_i \delta_{x_i} \tag{7.7}$$

are supported on the same set of n points $(x_i)_{i=1}^n \subset \mathcal{X}$, (7.6) defines a divergence on Σ_n

$$\mathrm{D}_\varphi(\mathrm{a}|\mathrm{b}) = \sum_{i \in \mathrm{Supp}(\mathrm{b})} \varphi\left(\frac{\mathrm{a}_i}{\mathrm{b}_i}\right) \mathrm{b}_i + \varphi'_\infty \sum_{i \notin \mathrm{Supp}(\mathrm{b})} \mathrm{a}_i, \tag{7.8}$$

where $\mathrm{Supp}(\mathrm{b}) \coloneqq \{i \in [\![n]\!] \,;\, \mathrm{b}_i \neq 0\}$.

Proposition 7.17 (Basic properties of φ-divergences)**.** *If φ is an entropy function, then $\mathcal{D}_\varphi$ is jointly 1-homogeneous, convex and weak-* lower semicontinuous in (α, β).*

Proof. Introduce the lower-semicontinuous perspective

$$\forall\, (u,v) \in (\mathbb{R}_+)^2, \quad \psi(u,v) = \begin{cases} v\varphi(u/v) & \text{if} \quad v > 0, \\ u\varphi'_\infty & \text{if} \quad v = 0. \end{cases}$$

The joint 1-homogeneity follows directly from this formula. To see convexity, first take $v_0, v_1 > 0$ and $\tau \in [0,1]$. With $\bar{v} = (1-\tau)v_0 + \tau v_1$, $\theta_0 = (1-\tau)v_0/\bar{v}$, and $\theta_1 = \tau v_1/\bar{v}$, one has $\theta_0 + \theta_1 = 1$ and $((1-\tau)u_0 + \tau u_1)/\bar{v} = \theta_0 u_0/v_0 + \theta_1 u_1/v_1$. Jensen's inequality for φ, followed by multiplication by $\bar{v}$, proves convexity of ψ when both second arguments are positive. Its recession value at $v = 0$ gives the lower-semicontinuous convex extension to the closed quadrant. In the discrete case, $\mathrm{D}_\varphi(\mathrm{a}|\mathrm{b}) = \sum_i \psi(\mathrm{a}_i, \mathrm{b}_i)$, so homogeneity and convexity follow by summation. For general finite measures, representing both arguments with respect to a common dominating measure yields the integral of ψ; the standard lower-semicontinuity theorem for convex integral functionals gives weak-* lower semicontinuity, with the recession value accounting for singular mass. □

The following proposition records when $\mathcal{D}_\varphi$ is nonnegative.

Proposition 7.18 (Non-negativity of φ-divergences)**.** *Assume that φ is normalized by $\varphi(1) = 0$. For probability distributions $(\alpha, \beta) \in \mathcal{M}_+^1(\mathcal{X})$, one has $\mathcal{D}_\varphi(\alpha|\beta) \geqslant 0$. If φ is strictly convex, then one has $\mathcal{D}_\varphi(\alpha|\beta) = 0$ if and only if $\alpha = \beta$. This property extends to arbitrary finite nonnegative measures $(\alpha, \beta) \in \mathcal{M}_+(\mathcal{X})$ if one furthermore imposes that $\varphi \geqslant 0$.*

Proof. Let $m = \alpha + \beta$ and write $a = \mathrm{d}\alpha/\mathrm{d}m$ and $b = \mathrm{d}\beta/\mathrm{d}m$. Then $a + b = 1$ m-almost everywhere and, using the perspective ψ above, $\mathcal{D}_\varphi(\alpha|\beta) = \int \psi(a,b)\mathrm{d}m$. Since $m(\mathcal{X}) = 2$, the measure $m/2$ is a probability. Jensen's inequality, the 1-homogeneity of ψ, and $\varphi(1) = 0$ give

$$\frac{1}{2}\mathcal{D}_\varphi(\alpha|\beta) \geqslant \psi\left(\frac{1}{2}\int a\mathrm{d}m, \frac{1}{2}\int b\mathrm{d}m\right) = \psi(1/2, 1/2) = 0.$$

If φ is strictly convex, the map $s \mapsto \psi(s, 1-s)$ is strictly convex on its effective domain: distinct points on the line $u + v = 1$ cannot lie on the same ray, which is the only degeneracy of a strict perspective. Equality in Jensen therefore forces $a = 1/2$, hence $a = b$ and $\alpha = \beta$. For arbitrary finite nonnegative measures, $\varphi \geqslant 0$ also implies $\varphi'_\infty \geqslant 0$, so every term in (7.6) is nonnegative. □

Classical examples and topology. The following examples calibrate the strength of φ-divergences. KL is sensitive to absolute continuity, while total variation gives the strong topology and therefore behaves very differently from Wasserstein-type weak metrics.

Example 7.19 (Kullback–Leibler divergence). The Kullback–Leibler divergence $\mathrm{KL} := \mathcal{D}_{\varphi_{\mathrm{KL}}}$, also known as the relative entropy, was already introduced in (8.12) and (8.7). It is the divergence associated to the Shannon–Boltzmann entropy function φ_{KL}, given by

$$\varphi_{\mathrm{KL}}(s) = \begin{cases} s\log(s) - s + 1 & \text{for } s > 0, \\ 1 & \text{for } s = 0, \\ +\infty & \text{otherwise.} \end{cases} \tag{7.9}$$

Example 7.20 (Total variation). With the convention used throughout this book, the total variation discrepancy $\mathrm{TV} := \mathcal{D}_{\varphi_{\mathrm{TV}}}$ is the full variation norm, without the factor $1/2$ sometimes used for probability measures. It is associated with

$$\varphi_{\mathrm{TV}}(s) = \begin{cases} |s-1| & \text{for } s \geqslant 0, \\ +\infty & \text{otherwise.} \end{cases} \tag{7.10}$$

It defines a norm on the full space of measures $\mathcal{M}(\mathcal{X})$, with

$$\mathrm{TV}(\alpha|\beta) = \|\alpha - \beta\|_{\mathrm{TV}}, \quad \text{where} \quad \|\alpha\|_{\mathrm{TV}} = |\alpha|(\mathcal{X}) = \int_{\mathcal{X}} \mathrm{d}|\alpha|(x). \tag{7.11}$$

If α has a density ρ_α on $\mathcal{X} = \mathbb{R}^d$, then the TV norm is the L^1 norm on functions, $\|\alpha\|_{\mathrm{TV}} = \int_{\mathcal{X}} |\rho_\alpha(x)|\mathrm{d}x = \|\rho_\alpha\|_{L^1}$. If α is discrete as in (7.7), then the TV norm is the ℓ^1 norm of vectors in $\mathbb{R}^n$, $\|\alpha\|_{\mathrm{TV}} = \sum_i |\mathrm{a}_i| = \|\mathrm{a}\|_{\ell^1}$.

KL and total variation are two very different φ-divergences: the former is smooth and sensitive to density ratios, whereas the latter is a nonsmooth norm. Pinsker's fundamental inequality nevertheless controls the square of the latter by the former [435].

Theorem 7.21 (Pinsker inequality)**.** *Let $\alpha, \beta \in \mathcal{M}_+^1(\mathcal{X})$ be probability measures on a measurable space. With the full-variation convention* (7.11)*,*

$$\mathrm{TV}(\alpha|\beta)^2 = \|\alpha - \beta\|_{\mathrm{TV}}^2 \leqslant 2\,\mathrm{KL}(\alpha|\beta).$$

In particular, on a finite space, for $\mathrm{a}, \mathrm{b} \in \Sigma_n$,

$$\|\mathrm{a} - \mathrm{b}\|_{\ell^1}^2 \leqslant 2\mathrm{KL}(\mathrm{a}|\mathrm{b}).$$

Proof. The claim is immediate if $\mathrm{KL}(\alpha|\beta) = +\infty$, so assume $\alpha \ll \beta$. Let A be a positive Hahn set for the signed measure $\alpha - \beta$, and set $a = \alpha(A)$ and $b = \beta(A)$. Since $(\alpha - \beta)(\mathcal{X}) = 0$,

$$a - b = \frac{1}{2}\|\alpha - \beta\|_{\mathrm{TV}}.$$

Applying Jensen's inequality separately on A and A^c, or equivalently coarse-graining onto the partition (A, A^c), gives

$$\mathrm{KL}(\alpha|\beta) \geqslant a\log\frac{a}{b} + (1-a)\log\frac{1-a}{1-b}.$$

For fixed $b \in (0,1)$, subtract $2(a-b)^2$ from the binary relative entropy on the right. The resulting function of a has value and first derivative zero at $a = b$, while its second derivative is

$$\frac{1}{a} + \frac{1}{1-a} - 4 \geqslant 0.$$

Hence the binary relative entropy is at least $2(a-b)^2 = \frac{1}{2}\|\alpha - \beta\|_{\mathrm{TV}}^2$. The boundary cases follow by lower semicontinuity. For a finite space, the identity $\|\alpha - \beta\|_{\mathrm{TV}} = \|\mathrm{a} - \mathrm{b}\|_{\ell^1}$ gives the second formula. □

Remark 7.22 (Strong vs. weak topology). The total variation norm (7.11) defines the so-called "strong" topology on the space of measures. For probability measures on a compact metric space, one has the

sharp elementary comparison

$$\mathcal{W}_1(\alpha,\beta) \leqslant \frac{\operatorname{diam}(\mathcal{X})}{2}\|\alpha-\beta\|_{\mathrm{TV}}.$$

Indeed, a 1-Lipschitz test function can be shifted by a constant so that its sup norm is at most $\operatorname{diam}(\mathcal{X})/2$. Thus total-variation convergence implies the weak convergence metrized by Wasserstein distances. The converse is false: if $x_n \to x$ with $x_n \neq x$, then $\delta_{x_n} \rightharpoonup \delta_x$ and $\mathcal{W}_1(\delta_{x_n},\delta_x) \to 0$, whereas $\|\delta_{x_n}-\delta_x\|_{\mathrm{TV}} = 2$ for every n. A chief advantage of the weak topology is compactness: when $\mathcal{X}$ is compact, every sequence in $\mathcal{M}_+^1(\mathcal{X})$ has a weakly convergent subsequence. This compactness underlies many existence arguments for optimization over probability measures.

Main families of φ-divergences. Several classical divergences fit in the same template. The power-divergence family

$$\varphi_\gamma(s) = \frac{s^\gamma - \gamma s + \gamma - 1}{\gamma(\gamma-1)} \qquad (\gamma \neq 0,1)$$

interpolates, up to conventional multiplicative normalizations, between Pearson's χ^2 divergence at $\gamma = 2$, Hellinger behavior at $\gamma = 1/2$, KL as $\gamma \to 1$, and reverse KL (or Burg entropy) $\varphi_0(s) = -\log s + s - 1$ as $\gamma \to 0$. The Hellinger divergence is often written separately with $\varphi_H(s) = (\sqrt{s}-1)^2$. If $\alpha = \rho_\alpha \lambda$ and $\beta = \rho_\beta \lambda$ relative to any common dominating measure λ, then

$$\mathfrak{h}(\alpha,\beta) := \mathcal{D}_{\varphi_H}(\alpha|\beta)^{1/2} = \|\sqrt{\rho_\alpha} - \sqrt{\rho_\beta}\|_{L^2(\lambda)}.$$

The Jensen–Shannon distance [203, 410] is the square root of the symmetrized, bounded KL-to-the-mixture divergence

$$\mathrm{JS}(\alpha,\beta)^2 = \frac{1}{2}\,\mathrm{KL}\left(\alpha\middle|\frac{\alpha+\beta}{2}\right) + \frac{1}{2}\,\mathrm{KL}\left(\beta\middle|\frac{\alpha+\beta}{2}\right),$$

and satisfies $0 \leqslant \mathrm{JS}(\alpha,\beta)^2 \leqslant \log 2$. Its exact entropy generator is

$$\varphi_{\mathrm{JS}}(s) = \frac{1}{2}\left[s\log s - (s+1)\log\left(\frac{s+1}{2}\right)\right], \qquad \mathcal{D}_{\varphi_{\mathrm{JS}}}(\alpha|\beta) = \mathrm{JS}(\alpha,\beta)^2.$$

Total variation corresponds to the non-smooth entropy $\varphi_{\mathrm{TV}}(s) = |s-1|$ and is exceptional because it is both a φ-divergence and an integral probability metric.

Figure 7.2 places the principal generators and their induced scalar density-ratio penalties side by side, clarifying how their different growth and boundary behavior affect measure comparison.

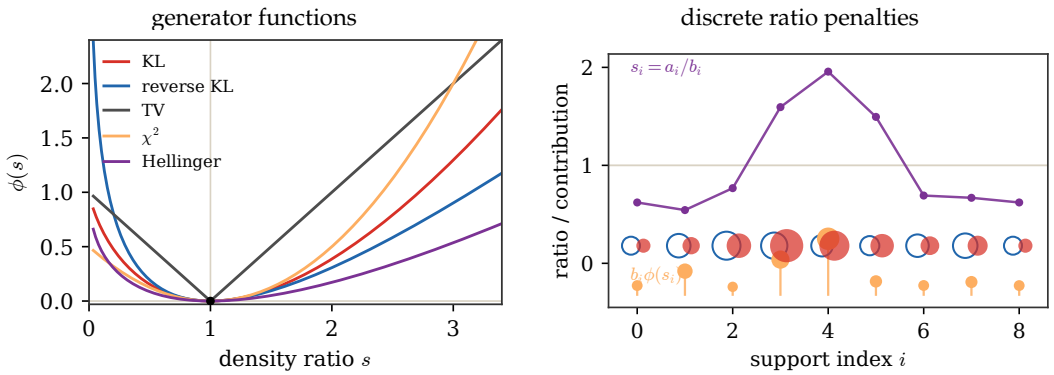


Figure 7.2: φ-divergences through density ratios. The left panel shows normalized generators for common divergences as functions of $s = \mathrm{d}\alpha/\mathrm{d}\beta$; all curves vanish at $s = 1$ up to affine normalization. The right panel shows the discrete formula $D_\varphi(a|b) = \sum_i b_i \varphi(a_i/b_i)$: hollow blue circles encode b_i, filled red circles encode a_i, the violet curve gives the ratios a_i/b_i, and orange lollipops show local KL-type contributions.

Remark 7.23 (φ-divergences versus Bregman divergences). Except for KL-type entropies, φ-divergences should not be confused with Bregman divergences. A φ-divergence compares measures pointwise through the density ratio $\mathrm{d}\alpha/\mathrm{d}\beta$. It is invariant under measurable bijections and, more generally, cannot increase under a measurable coarse-graining or Markov kernel [147]. A Bregman

> divergence is generated by a convex functional on a linear space and measures first-order Taylor error. KL is special because the integral entropy $\alpha \mapsto \int \rho \log \rho$ produces a Bregman divergence whose density-ratio form is also a φ-divergence.

Variational dual formula. The following formula turns a pointwise density-ratio penalty into a dual optimization problem over test functions. It is the analogue, for φ-divergences, of the Kantorovich dual formula for transport costs.

Proposition 7.24 (Variational representation of a φ-divergence)**.** *Assume that $\mathcal{X}$ is compact and let $\alpha, \beta \in \mathcal{M}_+(\mathcal{X})$. Define the Legendre transform of the entropy, already extended by $+\infty$ on negative arguments, by*

$$\varphi^*(s) := \sup_{r \geqslant 0} \{sr - \varphi(r)\}.$$

Then

$$\mathcal{D}_\varphi(\alpha|\beta) = \sup_{f \in C(\mathcal{X})} \left\{ \int_{\mathcal{X}} f \, \mathrm{d}\alpha - \int_{\mathcal{X}} \varphi^*(f) \mathrm{d}\beta \right\}. \tag{7.12}$$

Equivalently, the convex conjugate of $\mathcal{D}_\varphi(\cdot|\beta)$ on $\mathcal{M}(\mathcal{X})$ is

$$\forall f \in C(\mathcal{X}), \quad \mathcal{D}_\varphi^*(f|\beta) = \int_{\mathcal{X}} \varphi^*(f(x)) \mathrm{d}\beta(x). \tag{7.13}$$

Proof. First suppose that $\varphi'_\infty = +\infty$. Finite divergence then forces $\alpha = \rho\beta$ for some $\rho \geqslant 0$, and pointwise convex duality gives

$$\begin{aligned} \mathcal{D}_\varphi^*(f|\beta) &= \sup_{\rho \geqslant 0} \int_{\mathcal{X}} \big(f(x)\rho(x) - \varphi(\rho(x))\big) \mathrm{d}\beta(x) \\ &= \int_{\mathcal{X}} \sup_{r \geqslant 0} \big(f(x)r - \varphi(r)\big) \mathrm{d}\beta(x) = \int_{\mathcal{X}} \varphi^*(f(x)) \mathrm{d}\beta(x). \end{aligned}$$

For finite φ'_∞, the effective domain of φ^* is bounded above by φ'_∞. Hence a singular mass contributes at most $\varphi'_\infty \alpha^\perp(\mathcal{X})$ to the pairing, exactly matching the recession term in (7.6); approximation by continuous test functions recovers this value. The same conjugate formula therefore holds in general. Proposition 7.17 gives convexity and weak-* lower semicontinuity, so Fenchel–Moreau yields (7.12). □

7.4 GANs via Duality

GANs fit naturally into the dual viewpoint: the discriminator is a parameterized potential and the generator moves a reference measure. This section first explains the original divergence-based GAN objective, then contrasts it with integral probability metrics such as MMD and Wasserstein distances.

The goal is to fit a generative parametric model $\alpha_\theta = g_{\theta,\sharp}\zeta$ to empirical data $\beta = \frac{1}{m}\sum_{j=1}^m \delta_{y_j}$, where $\zeta \in \mathcal{M}_+^1(\mathcal{Z})$ is a fixed latent probability measure and $g_\theta : \mathcal{Z} \to \mathcal{X}$ is the generator, often a neural network.

Divergence-based adversarial losses. Any φ-divergence can be written in adversarial form through the dual formula (7.12):

$$\min_\theta \mathcal{D}_\varphi(\alpha_\theta|\beta) = \min_\theta \sup_f \int_{\mathcal{X}} f(x) \mathrm{d}\alpha_\theta(x) - \mathcal{D}_\varphi^*(f|\beta) = \min_\theta \sup_f \int_{\mathcal{X}} f(g_\theta(z)) \mathrm{d}\zeta(z) - \frac{1}{m}\sum_{j=1}^m \varphi^*(f(y_j)).$$

Replacing the unrestricted potential f by a neural network f_ξ gives the restricted saddle problem

$$\min_\theta \max_\xi \int_{\mathcal{Z}} f_\xi(g_\theta(z)) \mathrm{d}\zeta(z) - \frac{1}{m}\sum_{j=1}^m \varphi^*(f_\xi(y_j)).$$

For fixed θ, this restriction gives a lower bound on the exact divergence unless the discriminator class is rich enough to attain the supremum. This distinction is essential for empirical data: if β is discrete and α_θ is non-atomic, a superlinear divergence is $+\infty$, whereas its neural restricted objective can remain finite.

The original vanilla GAN [247] corresponds, up to an additive constant and the usual reparametrization by a discriminator taking values in $(0,1)$, to the unscaled Jensen–Shannon generator $\widehat{\varphi}_{\mathrm{JS}} = 2\varphi_{\mathrm{JS}}$,

$$\widehat{\varphi}_{\mathrm{JS}}(s) = s\log s - (s+1)\log\frac{s+1}{2}, \qquad \widehat{\varphi}^{*}_{\mathrm{JS}}(u) = -\log(2-e^{u}), \quad u < \log 2.$$

Thus $\mathcal{D}_{\widehat{\varphi}_{\mathrm{JS}}} = 2\,\mathrm{JS}^2$, consistently with the normalization above. In practice the min–max problem is solved by alternating stochastic gradient descent/ascent on (θ, ξ). Although the unrestricted maximization is concave in the function f, neural parametrization generally destroys concavity in ξ; the generator problem is likewise non-convex in θ. These facts contribute to instability and mode collapse. Moreover, density-ratio losses are informative when the measures overlap but can saturate when model and data are mutually singular: JS^2 reaches its maximum $\log 2$ on disjoint supports.

Dual norms and integral probability metrics. Instead of a density-ratio divergence, one can minimize a dual norm (7.1), also called an integral probability metric,

$$\min_{\theta} \|\alpha_\theta - \beta\|_B = \min_{\theta} \sup_{f\in B} \int_X f(x)\mathrm{d}(\alpha_\theta - \beta)(x) = \min_{\theta} \sup_{f\in B} \int_{\mathcal{Z}} f(g_\theta(z))\mathrm{d}\zeta - \frac{1}{m}\sum_{j=1}^{m} f(y_j).$$

MMD-GANs take B to be a unit ball in an RKHS [199]; Wasserstein GANs take B to be a Lipschitz ball, following Kantorovich–Rubinstein duality [19, 223]. The advantage is topological: for a continuous kernel on a compact space, the RKHS unit ball is uniformly bounded and equicontinuous, while the normalized Lipschitz ball is compact by Arzelà–Ascoli. The corresponding objectives are therefore continuous for weak convergence. They can compare singular empirical and generated measures through test functions instead of requiring pointwise density ratios. The price is that the discriminator class must be controlled geometrically, by a kernel norm, a Lipschitz constraint or a related regularization.

Example 7.25 (Application to imitation learning). In imitation learning, one can compare the expert occupancy measure ρ_E and the learner occupancy measure ρ_θ on state-action space. OT gives either a primal matching loss $W(\rho_\theta, \rho_E)$, or a dual adversarial reward obtained from a Kantorovich potential. Thus the discriminator in an adversarial imitation method can be interpreted as a learned reward shaping the learner toward the expert distribution, exactly as the GAN discriminator above is a learned potential. Wasserstein adversarial imitation and primal Wasserstein imitation exploit this distribution-matching viewpoint while retaining the geometry of state-action space, for instance through a cost that compares nearby states and actions more mildly than distant ones [553, 174].

Remark 7.26 (Weight clipping is only a proxy). Wasserstein GANs originally used weight clipping, constraining $\|\xi\|_\infty \leqslant 1$ as a proxy for enforcing $f_\xi \in B = \{f \,;\, \mathrm{Lip}(f) \leqslant 1\}$. The parameter box is convex, but its image through a neural-network parametrization is neither the full Lipschitz ball nor generally a convex class of functions, and optimization in the parameters remains non-concave. Clipping is therefore a practical heuristic rather than a faithful implementation of the Kantorovich–Rubinstein constraint.

CHAPTER 8

Entropic Regularization: Sinkhorn Algorithm

Entropic regularization makes optimal transport smooth, strictly convex and scalable. This chapter first explains the discrete KL-regularized problem, derives Sinkhorn's alternating matrix scaling algorithm, and then rewrites the same construction as a relative-entropy projection problem. It then records the general continuous formulation, develops the dual soft-transform picture, explains the path-space Schrödinger problem behind the static coupling formulation, and presents the main convex regularization variants and the debiased Sinkhorn divergence. A final section records a less standard viewpoint: after fixing the potential gauge, the finite-dimensional Sinkhorn equations admit a local holomorphic continuation to complex values of the temperature. The presentation connects the older matrix-scaling literature [490, 492, 491] with modern entropic OT [166, 430].

8.1 Entropic Regularization for Discrete Measures

Entropy turns a possibly non-unique linear program into a unique smooth problem. The price is bias, but the reward is differentiability and fast scaling algorithms.

Entropy penalty. The idea of the entropic regularization of optimal transport is to penalize concentrated couplings by adding the negative of the discrete Shannon–Boltzmann entropy.

Definition 8.1 (Discrete Shannon–Boltzmann entropy)**.** For a nonnegative matrix P, its Shannon–Boltzmann entropy is

$$\mathrm{H}(P) := -\sum_{i,j} P_{i,j} \log(P_{i,j}),$$

with the convention $0 \log(0) = 0$.

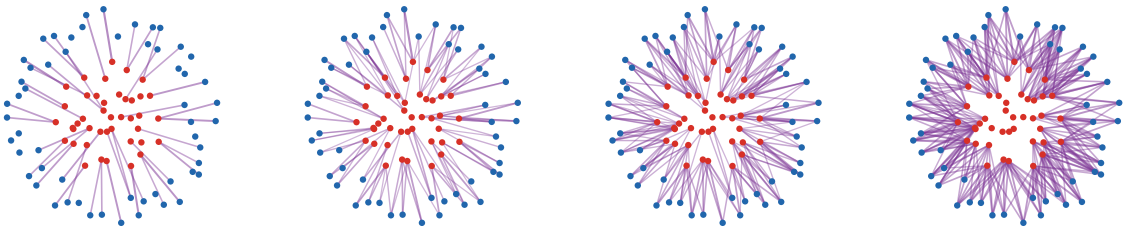
Entropy regularization gives the following discrete transport problem.

Definition 8.2 (Discrete entropic optimal transport)**.** For $\varepsilon > 0$, define

$$\mathrm{L}_C^{\varepsilon}(a, b) := \min_{P \in U(a,b)} \langle P, C \rangle - \varepsilon \mathrm{H}(P). \tag{8.1}$$

This is a strictly convex approximation of the discrete Kantorovich problem (3.2).

The temperature ε controls the compromise between transport cost and entropy. For small ε, the cost dominates and most of the visible mass follows a concentrated geometric matching. Increasing ε rewards more diffuse couplings, so the significant mass leaving each source point is spread over a wider range of targets; in the large-temperature limit, the geometry of the cost is progressively washed out and the plan approaches the independent coupling $a \otimes b$, as proved later in Proposition 8.10. Figure 8.1 makes this transition visible. Although only the largest entries are drawn, every entry of each displayed plan is positive, as proved next.



$\varepsilon = 0.018$ $\varepsilon = 0.045$ $\varepsilon = 0.120$ $\varepsilon = 0.320$

Figure 8.1: Entropically regularized couplings between the canonical red disk and blue annulus point clouds for four fixed regularization strengths. The squared Euclidean cost is normalized by its median and each plan is computed by log-domain Sinkhorn until the marginal residual is below the notebook tolerance. Violet segments display the largest visible entries of the computed plan, with thickness and opacity proportional to transported mass. The plans are strictly positive for every $\varepsilon > 0$, but the visible mass pattern evolves from nearly radial and concentrated to diffuse as ε increases.

Proposition 8.3 (Existence and uniqueness of entropic OT)**.** *Assume that* a, b *are probability histograms and that* C *is finite. For every* $\varepsilon > 0$*, problem* (8.1) *admits a unique minimizer. If all entries of* a *and* b *are positive, then this minimizer is positive on every entry.*

Proof. The transport polytope $\mathrm{U}(\mathrm{a}, \mathrm{b})$ is non-empty and compact, and the objective is continuous on it with the convention $0 \log 0 = 0$, so a minimizer exists. The scalar function $r \mapsto r \log r$ is strictly convex on $[0, +\infty)$. On the relative interior, the Hessian $-\partial^2 \mathrm{H}(\mathrm{P}) = \mathrm{diag}(1/\mathrm{P}_{i,j})$ is positive definite. Hence $-\mathrm{H}$ is strictly convex on every nontrivial feasible segment, and $\langle \mathrm{P}, \mathrm{C} \rangle - \varepsilon \mathrm{H}(\mathrm{P})$ has at most one minimizer.

If $\mathrm{a}_i, \mathrm{b}_j > 0$ and a minimizer had $\mathrm{P}_{i,j} = 0$, then for small $t > 0$ the perturbation $\mathrm{P}_t = (1 - t)\mathrm{P} + t\, \mathrm{a} \otimes \mathrm{b}$ remains feasible. The directional derivative of the entropic part at $t = 0$ is $-\infty$ because the derivative of $r \log r$ at 0 is $-\infty$ along a positive direction. Thus the objective decreases for small t, contradicting optimality. Therefore the minimizer is strictly positive. □

Smoothing effect. By Proposition 8.3, problem (8.1) has a unique optimal solution. Beyond providing uniqueness, this smoothing makes $\mathrm{L}_{\mathrm{C}}^{\varepsilon}(\mathrm{a}, \mathrm{b})$ a smooth function of a, b and C whenever these variables remain in the relative interior of their domains. In finite dimension, this follows from the smooth dual problem, which is strictly concave after fixing its additive gauge, and the envelope theorem. The entropy acts as a barrier for the positivity constraint: Proposition 8.3 shows that the solution is strictly positive on the support of $\mathrm{a} \otimes \mathrm{b}$. At the opposite, high-temperature end, Proposition 8.10 shows that $\mathrm{P}_{\varepsilon} \to \mathrm{a} \otimes \mathrm{b}$ as $\varepsilon \to +\infty$.

Figure 8.2 visualizes this entropic path on a transport-polytope slice and compares it with entropy applied to the slack variables of a generic linear program.

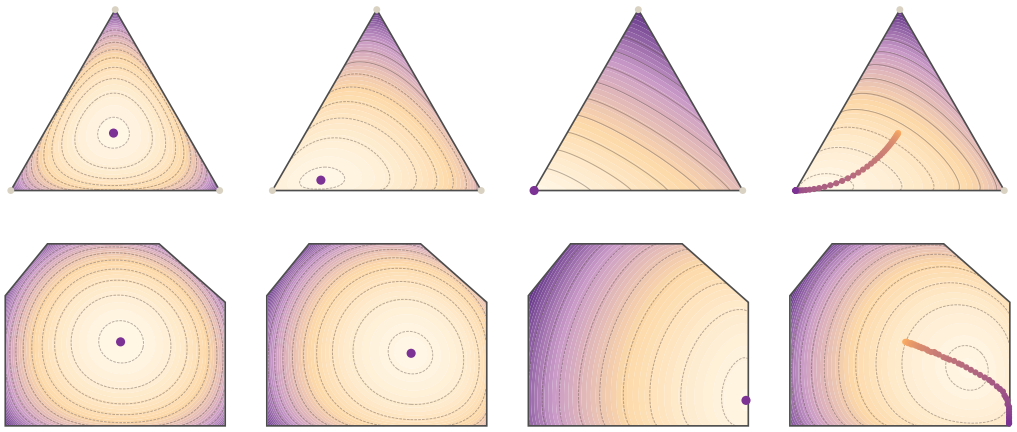


Figure 8.2: Entropic regularization and slack barriers. The first row shows the penalized objective $\langle c, p\rangle + \varepsilon \sum_i p_i(\log p_i - 1)$ on the three-state probability simplex $\{p \in \mathbb{R}^3_+ : \sum_{i=1}^3 p_i = 1\}$, embedded in $\mathbb{R}^3$ and visualized as a triangle; color and level sets represent the regularized functional itself, not only the linear part. The second row shows the analogous entropy-on-slacks objective $\ell^\top z + \varepsilon H(b - Az)$ on a two-dimensional polyhedron $Az \leqslant b$, with $H(s) = \sum_i s_i(\log s_i - 1)$. Large ε selects an interior reference point, while small ε moves the minimizer toward a low-cost face.

Remark 8.4 (Entropy barriers versus generic LP barriers). For a generic linear program $\min_z \ell^\top z$ subject to $Az \leqslant b$, one can introduce positive slacks $s = b - Az$ and use an entropy-on-slacks penalty $H(s) = \sum_i s_i(\log s_i - 1)$ as a smooth interior regularization. This is a useful analogy for Figure 8.2, but it is not the standard interior-point barrier for linear programming. The canonical barrier on the positive orthant is the Burg, or reverse-KL, logarithmic barrier $-\sum_i \log s_i$; it is self-concordant and therefore fits the Newton theory of interior-point methods [399]. The price is that a generic Newton step solves a dense linear system, leading to cubic per-iteration scaling in the relevant number of variables or constraints. Optimal transport is special: the entropy is placed on the entries of P, while the constraints are only the row and column marginals. This separable structure turns the associated Bregman projections into diagonal rescalings, hence into the Sinkhorn matrix-vector iterations developed next.

8.2 Sinkhorn's Algorithm

Sinkhorn's algorithm is alternating normalization of rows and columns. This section derives the scaling form of the optimizer and explains why each iteration only needs matrix-vector products.

The underlying matrix-scaling iteration has a long history, including iterative proportional fitting and the work of Sinkhorn and Knopp [490, 492, 491]. Its modern role in OT was transformed by Cuturi's entropic formulation [166]: the algorithm became a practical large-scale tool for machine learning, and also changed the way OT is viewed in ML, from a mostly geometric distance to a differentiable computational primitive.

The following proposition shows that the solution of (8.1) has a specific form. The parameterization is essentially dual: a coupling has nm entries and $n + m - 1$ independent marginal constraints, whereas the two scaling vectors use $n + m$ variables modulo one multiplicative gauge.

Proposition 8.5 (Scaling form of entropic OT)**.** P *is the unique solution to* (8.1) *if and only if there exists* $(\mathrm{u}, \mathrm{v}) \in \mathbb{R}^n_+ \times \mathbb{R}^m_+$ *such that*

$$\forall\, (i,j) \in [\![n]\!] \times [\![m]\!], \quad \mathrm{P}_{i,j} = \mathrm{u}_i \mathrm{K}_{i,j} \mathrm{v}_j \quad \textit{where} \quad \mathrm{K}_{i,j} := e^{-\frac{\mathrm{C}_{i,j}}{\varepsilon}}, \tag{8.2}$$

and $\mathrm{P} \in \mathrm{U}(\mathrm{a}, \mathrm{b})$.

Proof. Without loss of generality, we assume $\mathrm{a}_i, \mathrm{b}_j > 0$ (otherwise, rows or columns with zero mass are fixed to zero and can be removed). By Proposition 8.3, the minimizer is strictly positive on the remaining support.

We can thus ignore the positivity constraint when introducing two dual variables $f \in \mathbb{R}^n, g \in \mathbb{R}^m$ for each marginal constraint so that the Lagrangian of (8.1) reads

$$\mathcal{E}(P, f, g) = \langle P, C\rangle + \varepsilon \sum_{i,j} P_{i,j} \log(P_{i,j}) + \langle f, a - P\mathbb{1}_m\rangle + \langle g, b - P^\top \mathbb{1}_n\rangle.$$

The first-order conditions are

$$\frac{\partial \mathcal{E}(P, f, g)}{\partial P_{i,j}} = C_{i,j} + \varepsilon(\log(P_{i,j}) + 1) - f_i - g_j = 0.$$

Thus $P_{i,j} = e^{(f_i+g_j-C_{i,j})/\varepsilon-1}$, which has the stated form with the positive scalings $u_i := e^{f_i/\varepsilon-1}$ and $v_j := e^{g_j/\varepsilon}$. □

The factorization in (8.2) can be written compactly as $P = \operatorname{diag}(u) K \operatorname{diag}(v)$. u, v must therefore satisfy the nonlinear mass-conservation equations

$$\operatorname{diag}(u) K \operatorname{diag}(v)\mathbb{1}_m = a, \quad \text{and} \quad \operatorname{diag}(v) K^\top \operatorname{diag}(u)\mathbb{1}_n = b, \tag{8.3}$$

Using entrywise multiplication, these equations reduce to

$$u \odot (Kv) = a \quad \text{and} \quad v \odot (K^\top u) = b \tag{8.4}$$

where $\odot$ denotes entrywise multiplication. This is the matrix-scaling problem; see [398] and the references therein.

Diagonal normalization of a positive matrix has a long history, from iterative proportional fitting in statistics and economics [317, 558, 228] to modern matrix-balancing algorithms [288, 149]. When $n = m$ and a, b are uniform, this is diagonal scaling toward bistochasticity. Proposition 8.3 guarantees a unique scaled matrix P, although the two scaling vectors retain one multiplicative gauge degree of freedom. If some entries of K vanish, equivalently if C can take the value $+\infty$, existence and uniqueness require additional support conditions; here we keep $K > 0$.

Sinkhorn's algorithm solves the two equations alternately: it first updates u to enforce the source marginal and then updates v to enforce the target marginal,

$$u^{(\ell+1)} := \frac{a}{Kv^{(\ell)}} \quad \text{and} \quad v^{(\ell+1)} := \frac{b}{K^\top u^{(\ell+1)}}, \tag{8.5}$$

initialized with any positive vector, for instance $v^{(0)} = \mathbb{1}_m$. Division is entrywise. Different initializations may select different representatives of the same scaling class, because $(\lambda u, v/\lambda)$ yields the same coupling for every $\lambda > 0$. The alternating normalization can be read directly on the coupling matrix: a row update enforces the source marginal and generally perturbs the target marginal, while the next column update does the converse.

Algorithm 8.1: Sinkhorn scaling

Input: Positive weights a, b, cost matrix C, regularization $\varepsilon > 0$, tolerance tol, iteration budget $L \geqslant 1$.
Output: Entropic coupling P.
Initialize: Set $K_{i,j} = e^{-C_{i,j}/\varepsilon}$, $v^{(0)} = \mathbb{1}_m$, $r^{(0)} = +\infty$, and $\ell = 0$.
While $r^{(\ell)} >$ tol and $\ell < L$ **do**:
  Set $\ell \leftarrow \ell + 1$.
  $u^{(\ell)} = \frac{a}{Kv^{(\ell-1)}}$.
  $v^{(\ell)} = \frac{b}{K^\top u^{(\ell)}}$.
  $P^{(\ell)} = \operatorname{diag}(u^{(\ell)}) K \operatorname{diag}(v^{(\ell)})$.
  Set $r^{(\ell)} = \max\{\|P^{(\ell)}\mathbb{1}_m - a\|_1, \|(P^{(\ell)})^\top \mathbb{1}_n - b\|_1\}$.
Return $P^{(\ell)}$.

Figure 8.3 exposes the alternating feasibility mechanism on a small matrix: each row or column normalization enforces one marginal exactly while generally perturbing the other.

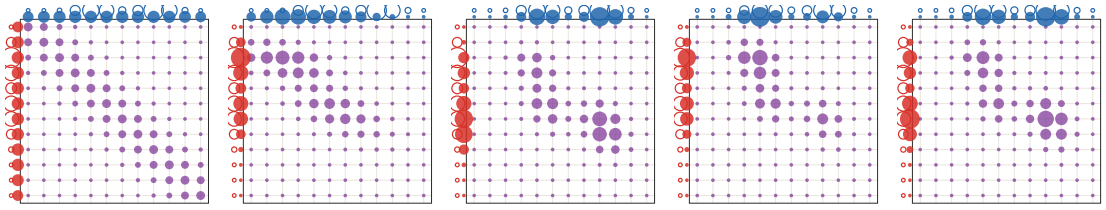


Figure 8.3: Marginal constraints during Sinkhorn scaling on a twelve-bin one-dimensional problem. Violet circles represent the current coupling matrix, framed by the thin black box; the red source marginal is displayed on the left and the blue target marginal below the matrix. Hollow side circles show the prescribed marginals, while filled circles show the current marginals. Row normalizations align the red marginal and leave a blue defect; column normalizations align the blue marginal and leave a red defect.

Figure 8.4 shows the same alternating projection mechanism on a dense one-dimensional discretization, where the marginal defects appear as continuous side curves.

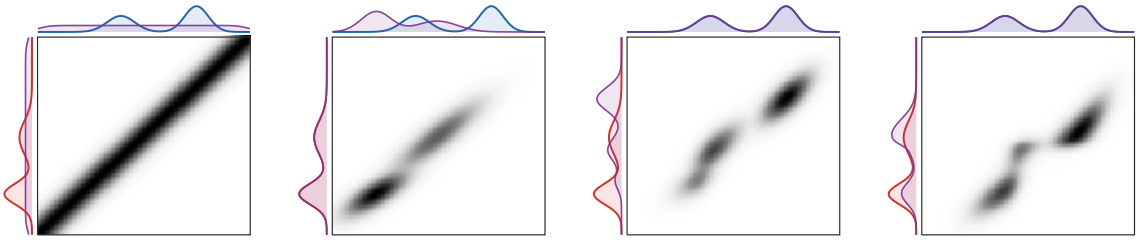


Figure 8.4: Dense Sinkhorn scaling for one-dimensional Gaussian-mixture marginals. The grayscale matrix is the current coupling, with a thin box delimiting only the matrix and not the side marginal plots. The red and blue side curves are the prescribed source and target marginals, while the violet side curves are the current row and column sums. A row scaling makes the violet curve coincide with the red source marginal, a column scaling makes it coincide with the blue target marginal, and the alternation rapidly stabilizes both sides.

After convergence, the regularization strength controls how much of the Gibbs kernel remains visible in the optimal plan. Small ε produces a concentrated transport band, while larger ε spreads the same marginals into a smoother coupling.

Figure 8.5 compares these converged plans at four temperatures while keeping both marginals fixed.

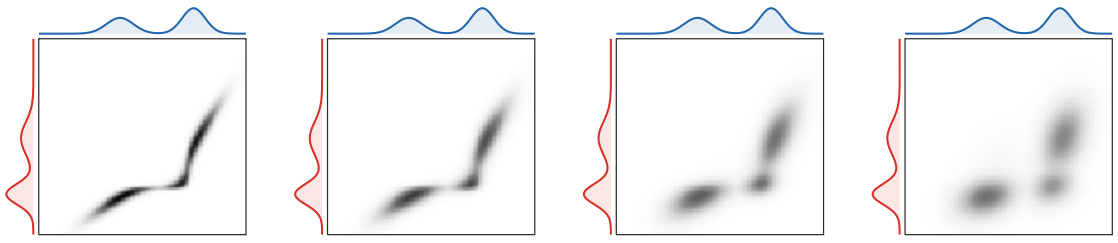


Figure 8.5: Final Sinkhorn couplings for the same one-dimensional Gaussian-mixture marginals and four regularization strengths. Each boxed matrix is the converged solution of the KL-regularized problem and uses a common grayscale normalization; the side curves display the fixed source and target marginals. Decreasing ε sharpens the plan toward an optimal-transport graph, whereas increasing ε keeps more of the diffuse product-measure structure.

Chapter 9 gives the formal convergence analysis. Before that, Figure 8.6 tracks the same Sinkhorn scaling in dual variables. The logarithmic row and column scalings, normalized to fix their additive gauge, rapidly settle to stable Kantorovich–Sinkhorn potentials.

Complexity bounds for Sinkhorn and comparisons with accelerated first-order methods are discussed in [7, 197, 309]. A chief computational advantage of Sinkhorn's algorithm, besides its simplicity, is that the only expensive operation is multiplication by the Gibbs kernel. With a dense $n \times m$ kernel, N_{it} iterations

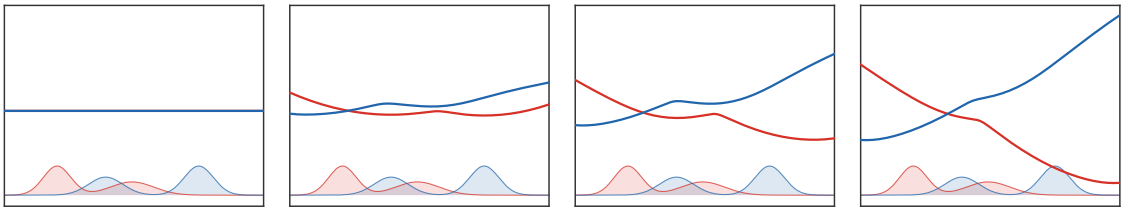


Figure 8.6: KL-normalized Sinkhorn dual potentials along Sinkhorn scaling. The setting is the same one-dimensional Gaussian-mixture problem as Figure 8.8, with fixed regularization strength $\varepsilon = 0.045$. The bottom silhouettes show both the source histogram a in red and the target histogram b in blue, while the red and blue curves are the logarithmic scaling potentials. All panels share the same axes, making stabilization visible.

therefore cost $O(N_{\text{it}}nm)$. Chapter 9 gives a more precise convergence statement: for fixed ε, the entropic dual gap has an $O(1/\ell)$ bound with a constant proportional to $1/\varepsilon$, while Hilbert-metric arguments give linear convergence when the Gibbs kernel is uniformly positive. Approximating unregularized OT also requires balancing optimization error against entropic bias. In finite dimension, choosing ε of order δ and solving the regularized problem to accuracy of order δ leads to an iteration bound of order $1/\delta^2$, up to logarithmic and cost-range factors [7, 197].

Figure 8.7 complements the complexity discussion by plotting the marginal defect across half-steps and showing how smaller temperatures slow the observed linear regime.

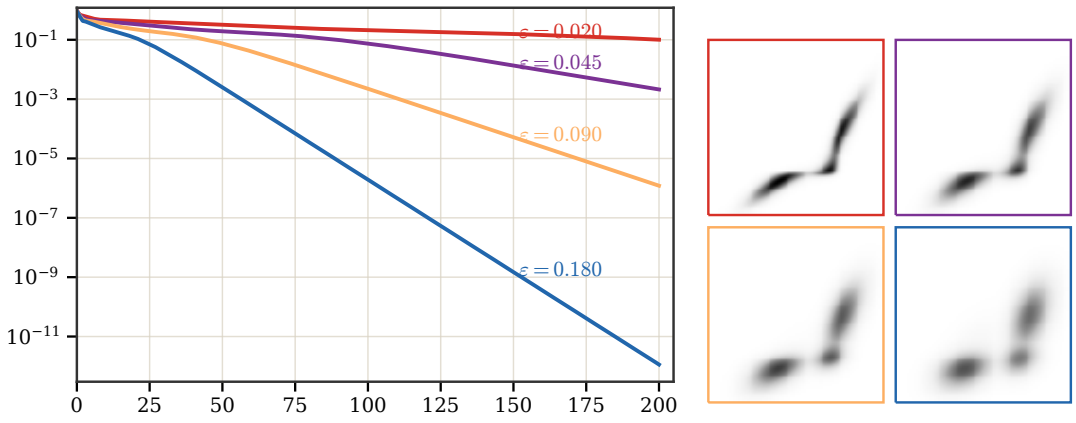


Figure 8.7: Marginal violation along Sinkhorn half-steps. The four curves use the same one-dimensional Gaussian-mixture problem and are paired with the corresponding limiting plans. The plotted error is $\frac{1}{2}(\|\mathrm{P}^{(\ell)}\mathbb{1} - \mathrm{a}\|_1 + \|(\mathrm{P}^{(\ell)})^\top\mathbb{1} - \mathrm{b}\|_1)$ on a logarithmic scale. The colored boxes on the right use the same colors as the convergence curves. For fixed positive ε, the curves enter a linear regime; smaller ε gives a sharper transport geometry but a more peaked Gibbs kernel and slower scaling.

In many applications, however, one does not need a highly accurate optimizer for the OT subproblem. Downstream performance is often tied more to the geometric bias of the discrepancy than to exact optimality, so a moderate number of iterations can suffice. This should be contrasted with generic interior-point methods, which use logarithmic barriers and require solving Newton systems. A dense transportation problem has nm primal variables, so a naive dense factorization can cost $O((nm)^3)$ per Newton step, or $O(n^6)$ when $n = m$. The overall complexity also depends on the formulation, sparsity, conditioning and iteration bound; the per-step comparison, rather than a universal total bound, is the relevant point here.

The second crucial aspect of Sinkhorn is that matrix-vector multiplication streams extremely well on GPU. Even better, if one is interested in computing many OT problems with a fixed cost matrix C, one can replace many matrix-vector multiplications with matrix-matrix multiplications, so that the computational gain can be substantial.

Remark 8.6 (Separable Gaussian kernels on grids). When the samples lie on a Cartesian grid and $c(x,y) = \|x-y\|^2$, the Gibbs kernel is Gaussian and factorizes along coordinates. If the grid has q points per axis in dimension d, so that $N = q^d$ grid points are used, then

$$K(x,y) = \exp\left(-\frac{\|x-y\|^2}{\varepsilon}\right) = \prod_{\ell=1}^{d} \exp\left(-\frac{(x_\ell - y_\ell)^2}{\varepsilon}\right).$$

Multiplication by K can therefore be applied by successively multiplying along each coordinate direction, equivalently by applying one-dimensional Gaussian kernel operators along the axes. On a periodic or sufficiently padded uniform grid these are literal discrete convolutions. A direct dense one-dimensional multiplication costs $O(q^2)$ on each of the q^{d-1} coordinate lines, and this is repeated for d axes. Hence one Sinkhorn half-step costs

$$O(d\, q^{d+1}) = O(d\, N^{1+1/d}) \tag{8.6}$$

instead of $O(N^2)$. With FFT-based or truncated Gaussian convolutions, the same separability can be pushed further, but the simple tensor-product estimate already explains why grid-based Sinkhorn can scale much better than a generic dense coupling.

8.3 Reformulation Using Relative Entropy

The KL formulation identifies Sinkhorn as a projection method. It also prepares the continuous and unbalanced settings, where a reference measure is essential.

Relative entropy. A convenient tool to reformulate and "normalize" this discrete entropy is relative entropy. It is the finite-dimensional divergence that turns entropy regularization into a projection problem and admits a direct measure-theoretic extension.

Definition 8.7 (Discrete relative entropy)**.** For nonnegative matrices P, Q of the same size, the generalized relative entropy, or Kullback–Leibler divergence, is

$$\mathrm{KL}(\mathrm{P}|\mathrm{Q}) \coloneqq \sum_{i,j} \mathrm{P}_{i,j} \log\left(\frac{\mathrm{P}_{i,j}}{\mathrm{Q}_{i,j}}\right) - \mathrm{P}_{i,j} + \mathrm{Q}_{i,j}. \tag{8.7}$$

The convention is $0\log(0) = 0$, and $\mathrm{KL}(\mathrm{P}|\mathrm{Q}) = +\infty$ if there exists (i,j) such that $\mathrm{Q}_{i,j} = 0$ but $\mathrm{P}_{i,j} \neq 0$.

For the specific case of comparing probability distributions, where P and Q have the same total mass, this further simplifies to

$$\mathrm{KL}(\mathrm{P}|\mathrm{Q}) = \sum_{i,j} \mathrm{P}_{i,j} \log\left(\frac{\mathrm{P}_{i,j}}{\mathrm{Q}_{i,j}}\right).$$

For the reference matrix $\mathrm{Q} = \mathbb{1}_{n\times m}$, one has

$$-\mathrm{KL}(\mathrm{P}|\mathbb{1}_{n\times m}) = \mathrm{H}(\mathrm{P}) + \sum_{i,j} \mathrm{P}_{i,j} - n\,m.$$

On fixed-mass couplings the last two terms are constant, so KL regularization with reference $\mathbb{1}_{n\times m}$ is equivalent to subtracting Shannon–Boltzmann entropy. KL is a particular instance of both a φ-divergence (as defined in Section 7.3) and a Bregman divergence; up to standard affine rescalings, it is the canonical overlap between these two families. This special property is at the heart of the fact that this regularization leads to elegant algorithms and a tractable mathematical analysis.

Proposition 8.8 (Non-negativity and definiteness of relative entropy)**.** *For all* $\mathrm{P}, \mathrm{Q} \in \mathbb{R}_+^{n\times m}$*, one has* $\mathrm{KL}(\mathrm{P}|\mathrm{Q}) \geqslant 0$*, with equality if and only if* $\mathrm{P} = \mathrm{Q}$*.*

Proof. Write $\varphi(s) = s\log s - s + 1$. Convexity gives $\varphi(s) \geqslant \varphi(1) + \varphi'(1)(s-1) = 0$, and strict convexity gives equality only at $s = 1$. Hence

$$\mathrm{KL}(\mathrm{P}|\mathrm{Q}) = \sum_{(i,j):\,\mathrm{Q}_{i,j}>0} \mathrm{Q}_{i,j}\varphi(\mathrm{P}_{i,j}/\mathrm{Q}_{i,j}) \geqslant 0,$$

whenever $P \ll Q$ entrywise; otherwise the divergence is $+\infty$. Equality forces $P_{i,j} = Q_{i,j}$ wherever $Q_{i,j} > 0$, while entries with $Q_{i,j} = 0$ vanish in both matrices. Thus equality is equivalent to $P = Q$. □

When P and Q are probability matrices and $P \ll Q$, the same divergence can also be written

$$\mathrm{KL}(P|Q) = \sum_{i,j} \psi(P_{i,j}/Q_{i,j})Q_{i,j}, \qquad \psi(s) = s \log s.$$

More generally, for any convex ψ such that $\psi(1) = 0$, Jensen's inequality gives

$$\sum_{i,j} \psi(P_{i,j}/Q_{i,j})Q_{i,j} \geqslant \psi\left(\sum_{i,j} P_{i,j}\right) = \psi(1) = 0.$$

KL reformulation of regularized OT. Choosing the tensor product $a \otimes b = (a_i b_j)_{i,j}$ as reference measure leads to the normalized problem

$$\min_{P \in U(a,b)} \langle P, C\rangle + \varepsilon \mathrm{KL}(P|a \otimes b). \tag{8.8}$$

For every $P \in U(a, b)$, the two objectives satisfy

$$\langle P, C\rangle + \varepsilon \mathrm{KL}(P|a \otimes b) = \langle P, C\rangle - \varepsilon H(P) + \varepsilon\big(H(a) + H(b)\big).$$

Thus (8.8) has exactly the same minimizer as the original entropic problem (8.1), while its optimal value is $L_C^\varepsilon(a, b) + \varepsilon\big(H(a) + H(b)\big)$. This normalization will matter again for unbalanced OT, where changing the reference measure is no longer merely a harmless notational choice.

More generally, for balanced OT with fixed positive marginals, changing the tensor-product reference only adds a constant to the objective, as shown in the following proposition.

Proposition 8.9 (Reference measure shift for KL)**.** *After removing zero-mass rows and columns, assume that* $a, a' \in \Sigma_n$ *and* $b, b' \in \Sigma_m$ *have positive entries. For every* $P \in U(a, b)$*, one has*

$$\mathrm{KL}(P|a \otimes b) = \mathrm{KL}(P|a' \otimes b') - \mathrm{KL}(a|a') - \mathrm{KL}(b|b').$$

Consequently, for fixed positive marginals, changing the positive tensor-product reference measure only adds a constant on the transport polytope. In particular, (8.8) *and* (8.1) *have the same unique minimizer.*

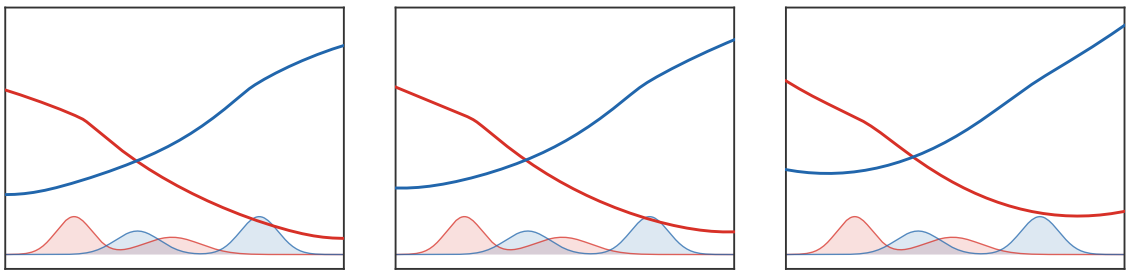
Proof. Expanding the logarithm and using the marginal constraints gives

$$\begin{aligned}\mathrm{KL}(P|a \otimes b) &= \mathrm{KL}(P|a' \otimes b') + \sum_i a_i \log \frac{a'_i}{a_i} + \sum_j b_j \log \frac{b'_j}{b_j} \\ &= \mathrm{KL}(P|a' \otimes b') - \mathrm{KL}(a|a') - \mathrm{KL}(b|b').\end{aligned}$$

□

The tensor-product reference is nevertheless useful when supports vary, because it makes explicit which entries are allowed to vanish. It is also the normalization that passes cleanly to the continuous formulation below, where the reference measure is $\alpha \otimes \beta$ rather than an ambient Lebesgue measure.

Figure 8.8 shows how the corresponding KL-normalized dual potentials deform with temperature, from nearly hard Kantorovich potentials to smoother log-sum-exp profiles.



$\varepsilon = 0.010$ $\varepsilon = 0.045$ $\varepsilon = 0.20$

Figure 8.8: KL-normalized Sinkhorn dual potentials for the same one-dimensional Gaussian-mixture histograms. The bottom silhouettes show both the source histogram a in red and the target histogram b in blue. The red and blue curves are the logarithmic scalings $f_i^\varepsilon = \varepsilon \log u_i^\varepsilon$ and $g_j^\varepsilon = \varepsilon \log v_j^\varepsilon$, with gauge $\langle f^\varepsilon, a\rangle = 0$, computed from $P_{i,j} = u_i\, a_i b_j e^{-C_{i,j}/\varepsilon}\, v_j$. The squared Euclidean cost is normalized by its median positive entry, and all panels use the same axes. For $\varepsilon = 0.010$ the curves are already close to the unregularized one-dimensional Kantorovich potentials; increasing ε turns this hard c-transform geometry into smoother log-sum-exp potentials.

The KL-normalized formulation also makes the two limiting regimes of the regularization parameter transparent.

Proposition 8.10 (Convergence with ε)**.** *Assume, after removing zero-mass rows and columns, that* a *and* b *are positive and that* C *is finite. The unique solution* P_ε *of* (8.1) *converges to the optimal solution with maximal entropy within the set of all optimal solutions of the Kantorovich problem, namely*

$$P_\varepsilon \xrightarrow{\varepsilon\to 0} \operatorname*{argmin}_P \ \{-H(P)\ ;\ P \in U(a,b), \langle P, C\rangle = L_C(a,b)\} \tag{8.9}$$

so that in particular

$$L_C^\varepsilon(a,b) \xrightarrow{\varepsilon\to 0} L_C(a,b).$$

Moreover,

$$P_\varepsilon \xrightarrow{\varepsilon\to\infty} a \otimes b. \tag{8.10}$$

Proof. **Case** $\varepsilon \to 0$**.** We consider a sequence $(\varepsilon_\ell)_\ell$ such that $\varepsilon_\ell \to 0$ and $\varepsilon_\ell > 0$. We denote P_ℓ the solution of (8.1) for $\varepsilon = \varepsilon_\ell$. Since $U(a,b)$ is bounded, we can extract a sequence (that we do not relabel for the sake of simplicity) such that $P_\ell \to P^\star$. Since $U(a,b)$ is closed, $P^\star \in U(a,b)$. We consider any P such that $\langle C, P\rangle = L_C(a,b)$. Using the equivalent KL-normalized formulation (8.8), optimality of P and P_ℓ for their respective optimization problems (for $\varepsilon = 0$ and $\varepsilon = \varepsilon_\ell$) gives

$$0 \leqslant \langle C, P_\ell\rangle - \langle C, P\rangle \leqslant \varepsilon_\ell(\mathrm{KL}(P|a\otimes b) - \mathrm{KL}(P_\ell|a\otimes b)). \tag{8.11}$$

Since KL is continuous, taking the limit $\ell \to +\infty$ in this expression shows that $\langle C, P^\star\rangle = \langle C, P\rangle$ so that $P^\star$ is a feasible point of (8.9). Furthermore, dividing by ε_ℓ in (8.11) and taking the limit shows that $\mathrm{KL}(P^\star|a\otimes b) \leqslant \mathrm{KL}(P|a\otimes b)$, which shows that $P^\star$ is a solution of (8.9). Since the solution $P_0^\star$ to this program is unique by strict convexity of $\mathrm{KL}(\cdot|a\otimes b)$ on the optimal face, one has $P^\star = P_0^\star$, and the whole sequence is converging.

Case $\varepsilon \to +\infty$**.** Subtracting $\min_{i,j} C_{i,j}$ from the cost changes every feasible objective by the same constant, so it does not change the minimizer. We can therefore assume $C \geqslant 0$. Evaluating the energy at $a\otimes b$ (which belongs to the constraint set $U(a,b)$), one has

$$\langle C, P_\varepsilon\rangle + \varepsilon \mathrm{KL}(P_\varepsilon|a\otimes b) \leqslant \langle C, a\otimes b\rangle + \varepsilon \times 0$$

and since $\langle C, P_\varepsilon\rangle \geqslant 0$, this leads to

$$\mathrm{KL}(P_\varepsilon|a\otimes b) \leqslant \varepsilon^{-1}\langle C, a\otimes b\rangle \leqslant \frac{\|C\|_\infty}{\varepsilon}$$

so that $\mathrm{KL}(P_\varepsilon|a\otimes b) \to 0$ and thus $P_\varepsilon \to a\otimes b$ since KL is a valid divergence. □

8.4 General Formulation

The continuous formulation replaces matrices by measures and discrete KL by relative entropy. This section records the measure-theoretic problem, explains how the temperature ε connects exact transport to the independent product coupling, and states the two asymptotic regimes that are useful later: a large-temperature expansion around independence and a small-temperature expansion around quadratic optimal transport. The path-space interpretation developed in Section 8.6 explains how the same static endpoint problem is obtained from an optimization problem on stochastic paths.

Measure formulation. The only structural change from the discrete problem is that matrix entries are replaced by densities with respect to a reference product measure. This keeps the marginal constraints unchanged while turning the entropy into a relative entropy.

One can consider arbitrary measures by replacing the discrete entropy with relative entropy with respect to the product measure $\mathrm{d}\alpha \otimes \mathrm{d}\beta(x,y) := \mathrm{d}\alpha(x)\mathrm{d}\beta(y)$. The measure-theoretic definition is the exact analogue of Definition 8.7, with absolute continuity replacing the entrywise support condition.

Definition 8.11 (Relative entropy of measures)**.** For finite nonnegative measures π and ξ on $\mathcal{X}\times\mathcal{Y}$, the relative entropy is

$$\mathrm{KL}(\pi|\xi) := \int_{\mathcal{X}\times\mathcal{Y}} \log\left(\frac{\mathrm{d}\pi}{\mathrm{d}\xi}\right)\mathrm{d}\pi + \xi(\mathcal{X}\times\mathcal{Y}) - \pi(\mathcal{X}\times\mathcal{Y}). \tag{8.12}$$

By convention, $\mathrm{KL}(\pi|\xi) = +\infty$ if π is not absolutely continuous with respect to ξ.

This relative entropy yields the continuous counterpart of Definition 8.2.

Definition 8.12 (Continuous entropic optimal transport)**.** For $\varepsilon > 0$, define

$$\mathcal{L}_c^{\varepsilon}(\alpha,\beta) := \min_{\pi\in\mathcal{U}(\alpha,\beta)} \int_{\mathcal{X}\times\mathcal{Y}} c(x,y)\mathrm{d}\pi(x,y) + \varepsilon\,\mathrm{KL}(\pi|\alpha\otimes\beta). \tag{8.13}$$

For fixed balanced marginals, the specific product reference in the entropy only matters up to additive constants, exactly as in Proposition 8.9, provided the alternative reference marginals are mutually absolutely continuous with α and β. Its support and absolute-continuity structure are nevertheless essential: they determine which couplings have finite entropy. This distinction becomes substantive in the unbalanced setting, where the marginal constraints no longer freeze the additive terms.

Probabilistic interpretation. The KL penalty has a simple statistical meaning: it measures how much dependence the coupling creates between its two marginals.

Definition 8.13 (Mutual information)**.** If $(X,Y)\sim\pi$ have marginals $X\sim\alpha$ and $Y\sim\beta$, the mutual information of the pair is

$$\mathcal{I}(X,Y) := \mathrm{KL}(\pi|\alpha\otimes\beta).$$

It is nonnegative and vanishes if and only if X and Y are independent.

With this terminology, the entropic problem (8.13) is equivalent to

$$\inf_{X\sim\alpha,\,Y\sim\beta} \mathbb{E}\big(c(X,Y)\big) + \varepsilon\mathcal{I}(X,Y).$$

Large ε therefore favors nearly independent endpoints, while small ε weakens the dependence penalty and recovers an optimal Monge–Kantorovich coupling in the limit. When the unregularized quadratic problem has a unique Brenier map, the limiting coupling is deterministic.

Sinkhorn for general measures. The multiplicative scaling structure of (8.2) extends from matrices to positive functions. Define the Gibbs kernel and its two integral operators by

$$\begin{aligned} k_\varepsilon(x,y) &:= \exp\left(-\frac{c(x,y)}{\varepsilon}\right), \\ (\mathcal{K}_\varepsilon v)(x) &:= \int_{\mathcal{Y}} k_\varepsilon(x,y)v(y)\mathrm{d}\beta(y), \quad (\mathcal{K}_\varepsilon^* u)(y) := \int_{\mathcal{X}} k_\varepsilon(x,y)u(x)\mathrm{d}\alpha(x). \end{aligned} \tag{8.14}$$

The star is justified by Fubini's identity $\int_{\mathcal{X}} u\,\mathcal{K}_\varepsilon v \mathrm{d}\alpha = \int_{\mathcal{Y}} v\,\mathcal{K}_\varepsilon^* u \mathrm{d}\beta$. For compact marginal supports and continuous c, Propositions 8.20 and 8.22 provide optimal dual potentials $(f_\varepsilon, g_\varepsilon)$. Set $u_\varepsilon = e^{f_\varepsilon/\varepsilon}$ and $v_\varepsilon = e^{g_\varepsilon/\varepsilon}$. The continuous density law (8.24) then becomes

$$\frac{\mathrm{d}\pi_\varepsilon}{\mathrm{d}(\alpha\otimes\beta)}(x,y) = u_\varepsilon(x)k_\varepsilon(x,y)v_\varepsilon(y), \tag{8.15}$$

Because the reference measure $\alpha\otimes\beta$ already contains the prescribed marginals, their target densities with respect to α and β are both one. The marginal constraints are therefore equivalent to

$$u_\varepsilon\,\mathcal{K}_\varepsilon v_\varepsilon = 1 \quad \alpha\text{-a.e.}, \qquad v_\varepsilon\,\mathcal{K}_\varepsilon^* u_\varepsilon = 1 \quad \beta\text{-a.e.}$$

Starting, for instance, from $v^{(0)} = 1$, continuous Sinkhorn alternately enforces these two identities:

$$u^{(\ell+1)} = \frac{1}{\mathcal{K}_\varepsilon v^{(\ell)}}, \qquad v^{(\ell+1)} = \frac{1}{\mathcal{K}_\varepsilon^* u^{(\ell+1)}}. \tag{8.16}$$

Equivalently, the updates read pointwise

$$u^{(\ell+1)}(x) = \frac{1}{\displaystyle\int_{\mathcal{Y}} k_\varepsilon(x,y)v^{(\ell)}(y)\mathrm{d}\beta(y)}, \qquad v^{(\ell+1)}(y) = \frac{1}{\displaystyle\int_{\mathcal{X}} k_\varepsilon(x,y)u^{(\ell+1)}(x)\mathrm{d}\alpha(x)}.$$

Given $v^{(\ell)}$, the first update produces an intermediate coupling with $\mathcal{X}$-marginal α; the second produces one with $\mathcal{Y}$-marginal β, while generally perturbing the first marginal again. As in the discrete case, (u,v) is defined only up to the gauge $(u,v)\mapsto(\lambda u, v/\lambda)$.

There is one useful situation in which no iteration is required: choose the target by applying the normalized Gibbs kernel itself to the source.

Proposition 8.14 (Closed-form Gibbs coupling). *Let β_0 be a sigma-finite reference measure on $\mathcal{Y}$ and suppose that*

$$Z_\varepsilon(x) := \int_{\mathcal{Y}} k_\varepsilon(x,y)\mathrm{d}\beta_0(y) \in (0,+\infty) \qquad \text{for } \alpha\text{-a.e. } x.$$

Define the normalized Gibbs transition and its output density by

$$p_\varepsilon(x,y) := \frac{k_\varepsilon(x,y)}{Z_\varepsilon(x)}, \qquad q_\varepsilon(y) := \int_{\mathcal{X}} p_\varepsilon(x,y)\mathrm{d}\alpha(x), \qquad \mathrm{d}\beta_\varepsilon = q_\varepsilon\,\mathrm{d}\beta_0. \tag{8.17}$$

Assume that $\log Z_\varepsilon \in L^1(\alpha)$, $\log q_\varepsilon \in L^1(\beta_\varepsilon)$, and the plan below has finite entropic objective. Then the unique solution of (8.13) between α and β_ε is

$$\mathrm{d}\pi_\varepsilon(x,y) = p_\varepsilon(x,y)\mathrm{d}\alpha(x)\mathrm{d}\beta_0(y). \tag{8.18}$$

Relative to $\alpha\otimes\beta_\varepsilon$, its Sinkhorn scalings are explicitly

$$\frac{\mathrm{d}\pi_\varepsilon}{\mathrm{d}(\alpha\otimes\beta_\varepsilon)}(x,y) = \frac{k_\varepsilon(x,y)}{Z_\varepsilon(x)q_\varepsilon(y)} = u_\varepsilon(x)k_\varepsilon(x,y)v_\varepsilon(y), \qquad u_\varepsilon = \frac{1}{Z_\varepsilon}, \quad v_\varepsilon = \frac{1}{q_\varepsilon}.$$

Proof. By construction, $p_\varepsilon(x,\cdot)\mathrm{d}\beta_0$ is a probability measure. Fubini's theorem therefore shows that (8.18) has first marginal α and second marginal β_ε.

Let $\gamma\in\mathcal{U}(\alpha,\beta_\varepsilon)$ have finite objective. Since $-\varepsilon\log k_\varepsilon = c$, the displayed density of π_ε gives

$$\varepsilon\,\mathrm{KL}(\gamma|\pi_\varepsilon) = \int c\,\mathrm{d}\gamma + \varepsilon\,\mathrm{KL}(\gamma|\alpha\otimes\beta_\varepsilon) + \varepsilon\int \log Z_\varepsilon\,\mathrm{d}\alpha + \varepsilon\int \log q_\varepsilon\,\mathrm{d}\beta_\varepsilon.$$

The last two terms depend only on the prescribed marginals. Thus minimizing the entropic objective is equivalent to minimizing $\mathrm{KL}(\gamma|\pi_\varepsilon)$, whose unique minimizer is $\gamma = \pi_\varepsilon$. □

If $k_\varepsilon(x,\cdot)$ is already normalized with respect to β_0, then $Z_\varepsilon = 1$ and (8.17) reduces to the particularly simple formula

$$\frac{\mathrm{d}\beta_\varepsilon}{\mathrm{d}\beta_0}(y) = \int_{\mathcal{X}} k_\varepsilon(x,y)\mathrm{d}\alpha(x).$$

For example, let $\mathcal{X} = \mathcal{Y} = \mathbb{R}^d$, let β_0 be Lebesgue measure, and take $c(x,y) = \|x-y\|^2$. Then $Z_\varepsilon = (\pi\varepsilon)^{d/2}$ and

$$p_\varepsilon(x,y) = (\pi\varepsilon)^{-d/2} e^{-\|x-y\|^2/\varepsilon}, \qquad \beta_\varepsilon = \alpha * \mathcal{N}\Big(0, \frac{\varepsilon}{2}\mathrm{Id}\Big).$$

Equivalently, (8.18) is the law of $(X, X + \sqrt{\varepsilon/2}\, G)$ for independent $X \sim \alpha$ and $G \sim \mathcal{N}(0, \mathrm{Id})$. Time-indexed Gaussian blurrings of this type are the forward noising mechanism behind diffusion models, with an additional deterministic rescaling for variance-preserving Ornstein–Uhlenbeck schedules; see the diffusion-model discussion in Section 16.1.

For discrete measures, setting $\mathrm{u}_i = \mathrm{a}_i u(x_i)$ and $\mathrm{v}_j = \mathrm{b}_j v(y_j)$ gives $\mathrm{P}_{i,j} = \mathrm{a}_i \mathrm{b}_j u(x_i)\mathrm{K}_{i,j} v(y_j) = \mathrm{u}_i \mathrm{K}_{i,j} \mathrm{v}_j$, so (8.16) reduces exactly to the matrix iteration (8.5). Its logarithmic interpretation as continuous dual block ascent is derived in the dual Sinkhorn paragraph. Continuous convergence is revisited through the generalized Fortet-type monotonicity argument of Section 9.2; the finite-dimensional linear rate is studied through Hilbert's metric in Section 9.4.

Convergence with ε. The continuous problem has the same qualitative temperature limits as Proposition 8.10, but one must be more careful about the zero-temperature selection. In finite dimension an exact optimal plan has finite entropy whenever its support is contained in the support of $\mathrm{a} \otimes \mathrm{b}$. For smooth quadratic transport, by contrast, the optimal Monge plan is typically singular with respect to $\alpha \otimes \beta$, so the robust statement is weak convergence of minimizers rather than entropy selection on the exact optimal face. The following statement is the usual Γ-convergence mechanism for entropic OT [336, 108]; the density hypothesis below is a convenient way to isolate the only approximation point needed in the proof.

Proposition 8.15 (Convergence with ε for measures)**.** *Let $\mathcal{X}, \mathcal{Y}$ be compact metric spaces, let $c \in \mathcal{C}(\mathcal{X}\times\mathcal{Y})$, and let $\alpha \in \mathcal{P}(\mathcal{X})$, $\beta \in \mathcal{P}(\mathcal{Y})$. Assume that finite-entropy couplings are dense in $\mathcal{U}(\alpha,\beta)$ for weak convergence with convergence of the cost integral. For $\varepsilon > 0$, let π_ε be a minimizer of (8.13). Then*

$$\mathcal{L}_c^\varepsilon(\alpha,\beta) \longrightarrow \mathcal{L}_c(\alpha,\beta) \quad \text{as} \quad \varepsilon \downarrow 0,$$

and every weak cluster point of $(\pi_\varepsilon)_\varepsilon$ is an optimal plan for the unregularized Kantorovich problem. If the latter optimal plan is unique, then π_ε converges weakly to it. In particular, for $\mathcal{X} = \mathcal{Y} \subset \mathbb{R}^d$, $c(x,y) = \|x-y\|^2$ and α absolutely continuous, $\pi_\varepsilon \rightharpoonup (\mathrm{Id}, T)_\sharp \alpha$, where T is the Brenier map.

As $\varepsilon \to +\infty$,

$$\pi_\varepsilon \longrightarrow \alpha\otimes\beta \quad \text{in total variation}, \qquad \mathcal{L}_c^\varepsilon(\alpha,\beta) \longrightarrow \int_{\mathcal{X}\times\mathcal{Y}} c\,\mathrm{d}\alpha\mathrm{d}\beta.$$

Proof. Existence follows from compactness of $\mathcal{U}(\alpha,\beta)$ and lower semicontinuity of the relative entropy. For the limit $\varepsilon \downarrow 0$, take any sequence $\varepsilon_k \downarrow 0$ and a weak cluster point $\pi^\star$ of π_{ε_k}. Since $\mathrm{KL}(\pi|\alpha\otimes\beta) \geqslant 0$ on probability couplings,

$$\liminf_k \mathcal{L}_c^{\varepsilon_k}(\alpha,\beta) \geqslant \liminf_k \int c\,\mathrm{d}\pi_{\varepsilon_k} = \int c\,\mathrm{d}\pi^\star \geqslant \mathcal{L}_c(\alpha,\beta).$$

Conversely, by the density assumption, for every $\delta > 0$ there exists a coupling π^δ with finite relative entropy and $\int c\,\mathrm{d}\pi^\delta \leqslant \mathcal{L}_c(\alpha,\beta) + \delta$. Testing (8.13) with π^δ gives

$$\limsup_k \mathcal{L}_c^{\varepsilon_k}(\alpha,\beta) \leqslant \int c\,\mathrm{d}\pi^\delta.$$

Letting $\delta \downarrow 0$ proves convergence of the values and forces $\int c\,\mathrm{d}\pi^\star = \mathcal{L}_c(\alpha,\beta)$. Uniqueness of the unregularized optimizer then identifies all cluster points. Brenier's theorem gives the stated quadratic consequence when α is absolutely continuous.

For the limit $\varepsilon \to +\infty$, subtract a constant from c and assume $c \geqslant 0$. Since $\alpha \otimes \beta$ is admissible and has zero relative entropy with respect to itself,

$$\int c \, \mathrm{d}\pi_\varepsilon + \varepsilon \, \mathrm{KL}(\pi_\varepsilon | \alpha \otimes \beta) \leqslant \int c \, \mathrm{d}\alpha \mathrm{d}\beta.$$

Hence $\mathrm{KL}(\pi_\varepsilon | \alpha \otimes \beta) \leqslant \varepsilon^{-1} \int c \, \mathrm{d}\alpha \mathrm{d}\beta$. Pinsker's inequality from Theorem 7.21 gives total-variation convergence to $\alpha \otimes \beta$, and the convergence of the values follows by continuity of c. □

Large-temperature expansion. When ε is large, the entropy dominates and the optimal plan is a small perturbation of the product coupling. The useful quantity is the component of the cost that cannot be absorbed into row and column potentials.

Proposition 8.16 (Large-temperature expansion)**.** *Let $r = \alpha \otimes \beta$ and assume $c \in L^\infty(r)$. Define the product averages*

$$\bar{c}_{\mathcal{X}}(x) \coloneqq \int_{\mathcal{Y}} c(x,y) \, \mathrm{d}\beta(y), \qquad \bar{c}_{\mathcal{Y}}(y) \coloneqq \int_{\mathcal{X}} c(x,y) \, \mathrm{d}\alpha(x), \qquad \bar{c} \coloneqq \int_{\mathcal{X}\times\mathcal{Y}} c \, \mathrm{d}r,$$

and the centered interaction

$$c_0(x,y) \coloneqq c(x,y) - \bar{c}_{\mathcal{X}}(x) - \bar{c}_{\mathcal{Y}}(y) + \bar{c}.$$

Assume that the large-temperature branch $p_\varepsilon = \mathrm{d}\pi_\varepsilon / \mathrm{d}r$, $f_\varepsilon, g_\varepsilon$ admits an expansion to third order in ε^{-1} near 0 in $L^\infty(r)$, with gauge $\int g_\varepsilon \, \mathrm{d}\beta = 0$. Then

$$p_\varepsilon(x,y) = 1 - \frac{c_0(x,y)}{\varepsilon} + O(\varepsilon^{-2}) \quad \text{in } L^2(r),$$

and

$$\mathcal{L}_c^\varepsilon(\alpha,\beta) = \bar{c} - \frac{1}{2\varepsilon} \int_{\mathcal{X}\times\mathcal{Y}} c_0(x,y)^2 \, \mathrm{d}r(x,y) + \frac{1}{6\varepsilon^2} \int_{\mathcal{X}\times\mathcal{Y}} c_0(x,y)^3 \, \mathrm{d}r(x,y) + O(\varepsilon^{-3}).$$

Moreover, with

$$A(x) \coloneqq \int_{\mathcal{Y}} c_0(x,y)^2 \, \mathrm{d}\beta(y), \qquad B(y) \coloneqq \int_{\mathcal{X}} c_0(x,y)^2 \, \mathrm{d}\alpha(x), \qquad \sigma^2 \coloneqq \int_{\mathcal{X}\times\mathcal{Y}} c_0^2 \, \mathrm{d}r,$$

one has

$$f_\varepsilon(x) = \bar{c}_{\mathcal{X}}(x) - \frac{A(x)}{2\varepsilon} + O(\varepsilon^{-2}),$$

and

$$g_\varepsilon(y) = \bar{c}_{\mathcal{Y}}(y) - \bar{c} + \frac{\sigma^2 - B(y)}{2\varepsilon} + O(\varepsilon^{-2})$$

in the same gauge.

Proof. The optimality equation can be written

$$p_\varepsilon(x,y) = \exp\left(\frac{f_\varepsilon(x) + g_\varepsilon(y) - c(x,y)}{\varepsilon} \right),$$

with the constraints $\int p_\varepsilon(x,y) \, \mathrm{d}\beta(y) = 1$ and $\int p_\varepsilon(x,y) \, \mathrm{d}\alpha(x) = 1$. Write $p_\varepsilon = 1 + \varepsilon^{-1} q + O(\varepsilon^{-2})$. The linearized marginal constraints impose

$$\int q(x,y) \, \mathrm{d}\beta(y) = 0, \qquad \int q(x,y) \, \mathrm{d}\alpha(x) = 0.$$

Expanding the objective gives

$$\int c \, p_\varepsilon \, \mathrm{d}r + \varepsilon \int p_\varepsilon \log(p_\varepsilon) \, \mathrm{d}r = \bar{c} + \frac{1}{\varepsilon} \left(\int c \, q \, \mathrm{d}r + \frac{1}{2} \int q^2 \, \mathrm{d}r \right) + O(\varepsilon^{-2}).$$

Thus q is the minimizer of the quadratic expression over zero-row and zero-column perturbations. Orthogonal projection in $L^2(r)$ gives $q = -c_0$, because c_0 has zero conditional means. This proves the first-order density and value expansions.

To identify the next displayed terms, write $p_\varepsilon = 1 + \varepsilon^{-1} q + \varepsilon^{-2} s + O(\varepsilon^{-3})$ and $f_\varepsilon = f_0 + \varepsilon^{-1} f_1 + O(\varepsilon^{-2})$, $g_\varepsilon = g_0 + \varepsilon^{-1} g_1 + O(\varepsilon^{-2})$, where $f_0 = \bar{c}_\mathcal{X}$ and $g_0 = \bar{c}_\mathcal{Y} - \bar{c}$. Since

$$p_\varepsilon = \exp\left(-\frac{c_0}{\varepsilon} + \frac{f_1(x) + g_1(y)}{\varepsilon^2} + O(\varepsilon^{-3})\right),$$

one has $s = f_1 + g_1 + c_0^2/2$. The second-order marginal constraints and the gauge $\int g_1 \, \mathrm{d}\beta = 0$ give $f_1 = -A/2$ and $g_1 = (\sigma^2 - B)/2$. Finally,

$$p_\varepsilon \log p_\varepsilon = \varepsilon^{-1} q + \varepsilon^{-2}\left(s + \frac{1}{2} q^2\right) + \varepsilon^{-3}\left(qs - \frac{1}{6} q^3\right) + O(\varepsilon^{-4}),$$

and the zero-conditional-mean property of c_0 cancels all additive terms in s. The order-ε^{-2} contribution to the value is therefore $\frac{1}{2}\int c_0^3 \, \mathrm{d}r - \frac{1}{3}\int c_0^3 \, \mathrm{d}r = \frac{1}{6}\int c_0^3 \, \mathrm{d}r$. □

Small-temperature expansion for smooth densities. At the opposite end, entropic transport is a viscous perturbation of quadratic optimal transport. The expansion contains an $\varepsilon \log \varepsilon$ term coming from the Gaussian normalization of Brownian bridges, then an endpoint entropy correction, and finally a Fisher-information term integrated along the McCann interpolation. The following statement translates the small-noise expansion of the Schrödinger problem [153] to the present convention, where the static cost is $\|x - y\|^2 + \varepsilon \, \mathrm{KL}(\cdot | \alpha \otimes \beta)$. The same second-order term is the reason why Sinkhorn divergences cancel the leading entropic bias [142].

Proposition 8.17 (Small-temperature quadratic expansion)**.** *Let $\alpha = \rho_\alpha \, \mathrm{d}x$ and $\beta = \rho_\beta \, \mathrm{d}x$ be probability measures on $\mathbb{R}^d$ with bounded compactly supported densities. Let $\alpha_t = \rho_t \, \mathrm{d}x$ be their quadratic displacement interpolation and assume*

$$\mathcal{I}_{\mathrm{geo}}(\alpha, \beta) := \int_0^1 \int_{\mathbb{R}^d} \|\nabla \log \rho_t(x)\|^2 \rho_t(x) \, \mathrm{d}x \, \mathrm{d}t < +\infty.$$

For $c(x, y) = \|x - y\|^2$, define

$$\mathrm{H}(\alpha) := \int_{\mathbb{R}^d} \rho_\alpha \log \rho_\alpha \, \mathrm{d}x, \qquad \mathrm{H}(\beta) := \int_{\mathbb{R}^d} \rho_\beta \log \rho_\beta \, \mathrm{d}x.$$

Then, as $\varepsilon \downarrow 0$,

$$\mathcal{L}^\varepsilon_{\|\cdot\|^2}(\alpha, \beta) = \mathcal{W}_2^2(\alpha, \beta) - \frac{d\varepsilon}{2} \log(\pi\varepsilon) - \frac{\varepsilon}{2}\left(\mathrm{H}(\alpha) + \mathrm{H}(\beta)\right) + \frac{\varepsilon^2}{16} \mathcal{I}_{\mathrm{geo}}(\alpha, \beta) + o(\varepsilon^2).$$

If, in addition, the endpoint densities are smooth and positive on their supports and the quadratic optimal map is a smooth non-degenerate diffeomorphism, let $f_\varepsilon, g_\varepsilon$ be optimal Sinkhorn potentials and f_0, g_0 Kantorovich potentials, normalized by $\int g_\varepsilon \, \mathrm{d}\beta = \int g_0 \, \mathrm{d}\beta = 0$. Then $f_\varepsilon \to f_0$ and $g_\varepsilon \to g_0$ locally uniformly on the interiors of the supports [406]. Moreover, the scalar part of the first potential is fixed by

$$\int f_\varepsilon \, \mathrm{d}\alpha = \mathcal{W}_2^2(\alpha, \beta) - \frac{d\varepsilon}{2} \log(\pi\varepsilon) - \frac{\varepsilon}{2}\big(\mathrm{H}(\alpha) + \mathrm{H}(\beta)\big) + \frac{\varepsilon^2}{16} \mathcal{I}_{\mathrm{geo}}(\alpha, \beta) + o(\varepsilon^2).$$

The spatial order-ε corrections follow from the Laplace prefactors in the two soft c-transform equations and depend on the endpoint densities and the Hessian of the quadratic contact phase.

Proof. For the Brownian reference with heat kernel variance T, the dynamic Schrödinger formulation gives

$$T\mathscr{C}_T(\alpha, \beta) = \frac{1}{2} \mathcal{W}_2^2(\alpha, \beta) + \frac{T}{2}\left(\mathrm{H}(\alpha) + \mathrm{H}(\beta)\right) + \frac{T^2}{8} \mathcal{I}_{\mathrm{geo}}(\alpha, \beta) + o(T^2),$$

under the stated regularity assumptions [153]. Here $\mathscr{C}_T$ uses entropy relative to the sigma-finite Brownian endpoint measure:

$$\mathscr{C}_T(\alpha, \beta) = \inf_{\pi \in \mathcal{U}(\alpha, \beta)} \mathscr{H}(\pi | p_T(x, y) \, \mathrm{d}x \mathrm{d}y), \qquad p_T(x, y) = (2\pi T)^{-d/2} e^{-\|x - y\|^2/(2T)},$$

where $\mathscr{H}(\pi|\xi) := \int \log(\mathrm{d}\pi/\mathrm{d}\xi)\mathrm{d}\pi$ for sigma-finite ξ. When ξ is a probability measure, this agrees with Definition 8.11; unlike the generalized KL of that definition, it remains meaningful when the Brownian reference has infinite total mass. Expanding the heat kernel, multiplying by $2T$, and setting $\varepsilon = 2T$ gives

$$2T\mathscr{C}_T(\alpha,\beta) = \mathcal{L}^{\varepsilon}_{\|\cdot\|^2}(\alpha,\beta) + 2T\,(\mathrm{H}(\alpha)+\mathrm{H}(\beta)) + dT\log(2\pi T).$$

Equivalently,

$$\mathcal{L}^{\varepsilon}_{\|\cdot\|^2}(\alpha,\beta) = 2T\mathscr{C}_T(\alpha,\beta) - 2T\,(\mathrm{H}(\alpha)+\mathrm{H}(\beta)) - dT\log(2\pi T),$$

which is exactly the displayed expansion because $2T = \varepsilon$. At a dual optimum the exponential penalty integrates to zero, so the gauge $\int g_\varepsilon \mathrm{d}\beta = 0$ gives $\int f_\varepsilon \mathrm{d}\alpha = \mathcal{L}^{\varepsilon}_{\|\cdot\|^2}(\alpha,\beta)$. This proves the displayed scalar expansion. Local convergence and the spatial corrections follow from Laplace's method applied to the soft c-transform equations; changing the additive gauge redistributes scalar terms between the two potentials. □

8.5 Dual of Sinkhorn

The dual point of view replaces couplings by potentials and soft c-transforms. It is the right formulation for stabilized implementations and differentiation.

Discrete dual. The following proposition details the dual problem associated with the KL-normalized formulation (8.8). This formulation has the same minimizer as (8.1); its optimal value is shifted by the constant $\varepsilon\mathrm{H}(\mathrm{a}) + \varepsilon\mathrm{H}(\mathrm{b})$.

Proposition 8.18 (Dual of entropic OT). *The optimal value of* (8.8) *is*

$$\min_{\mathrm{P}\in\mathrm{U}(\mathrm{a},\mathrm{b})} \langle \mathrm{P},\mathrm{C}\rangle + \varepsilon\mathrm{KL}(\mathrm{P}|\mathrm{a}\otimes\mathrm{b}) = \max_{\mathrm{f}\in\mathbb{R}^n,\mathrm{g}\in\mathbb{R}^m} \langle \mathrm{f},\mathrm{a}\rangle + \langle \mathrm{g},\mathrm{b}\rangle - \varepsilon\sum_{i,j}\exp\left(\frac{\mathrm{f}_i+\mathrm{g}_j-\mathrm{C}_{i,j}}{\varepsilon}\right)\mathrm{a}_i\mathrm{b}_j + \varepsilon. \tag{8.19}$$

The optimal (f,g) *are linked to scalings* (u,v) *appearing in* (8.2) *through*

$$\mathrm{u}_i = \mathrm{a}_i e^{\mathrm{f}_i/\varepsilon} \quad \textit{and} \quad \mathrm{v}_j = \mathrm{b}_j e^{\mathrm{g}_j/\varepsilon}. \tag{8.20}$$

Proof. We introduce Lagrange multipliers and consider

$$\min_{\mathrm{P}\geqslant 0}\max_{\mathrm{f},\mathrm{g}} \langle \mathrm{C},\mathrm{P}\rangle + \varepsilon\mathrm{KL}(\mathrm{P}|\mathrm{a}\otimes\mathrm{b}) + \langle \mathrm{a}-\mathrm{P}\mathbb{1},\mathrm{f}\rangle + \langle \mathrm{b}-\mathrm{P}^\top\mathbb{1},\mathrm{g}\rangle.$$

Finite-dimensional convex duality allows us to exchange the minimum over P with the maximum over (f,g), giving

$$\max_{\mathrm{f},\mathrm{g}} \langle \mathrm{f},\mathrm{a}\rangle + \langle \mathrm{g},\mathrm{b}\rangle + \varepsilon\min_{\mathrm{P}\geqslant 0}\left(\mathrm{KL}(\mathrm{P}|\mathrm{a}\otimes\mathrm{b}) - \langle\frac{\mathrm{f}\oplus\mathrm{g}-\mathrm{C}}{\varepsilon},\mathrm{P}\rangle\right) = \langle \mathrm{f},\mathrm{a}\rangle + \langle \mathrm{g},\mathrm{b}\rangle - \varepsilon\mathrm{KL}^*\left(\frac{\mathrm{f}\oplus\mathrm{g}-\mathrm{C}}{\varepsilon}|\mathrm{a}\otimes\mathrm{b}\right).$$

One concludes by using (7.13) for $\varphi(r) = r\log(r) - r + 1$

$$\mathrm{KL}^*(H|\mathrm{a}\otimes\mathrm{b}) = \sum_{i,j}\varphi^*(H_{i,j})\mathrm{a}_i\mathrm{b}_j.$$

Indeed, the scalar maximization $\varphi^*(s) = \sup_{r\geqslant 0}\{rs - r\log r + r - 1\}$ has first-order condition $s - \log r = 0$, hence $r = e^s$ and $\varphi^*(s) = e^s - 1$. □

Discrete soft c-transforms. Since the dual problem (8.19) is smooth and concave, one can perform alternating block maximization. For a fixed g, maximizing with respect to f leads to the following equation after zeroing the derivative with respect to f:

$$\mathrm{a}_i - e^{\frac{\mathrm{f}_i}{\varepsilon}}\mathrm{a}_i\sum_j \exp\left(\frac{\mathrm{g}_j-\mathrm{C}_{i,j}}{\varepsilon}\right)\mathrm{b}_j = 0$$

which leads to the explicit solution

$$\mathrm{f}_i = -\varepsilon\log\sum_j \exp\left(\frac{\mathrm{g}_j-\mathrm{C}_{i,j}}{\varepsilon}\right)\mathrm{b}_j.$$

The log-sum-exp form is best read as a smoothed minimum. This gives the entropic analogue of the hard c-transform.

Definition 8.19 (Soft-min and discrete soft c-transform)**.** For $h \in \mathbb{R}^m$ and weights $\mathrm{b} \in \Sigma_m$, the weighted soft-min at temperature $\varepsilon > 0$ is

$$\min_{\mathrm{b}}^{\varepsilon}(h) := -\varepsilon \log \sum_j e^{-h_j/\varepsilon} \mathrm{b}_j.$$

It converges to $\min_{j:\, b_j>0} h_j$ as $\varepsilon \to 0$; for positive weights this is $\min_j h_j$. Given a cost matrix C, the discrete soft c-transforms are

$$\mathrm{f}_i = \min_{\mathrm{b}}^{\varepsilon}(\mathrm{C}_{i,\cdot} - \mathrm{g}), \qquad \mathrm{g}_j = \min_{\mathrm{a}}^{\varepsilon}(\mathrm{C}_{\cdot,j} - \mathrm{f}). \tag{8.21}$$

Exponentiating these iterations recovers exactly the Sinkhorn algorithm. These iterations, however, become unstable for small ε. To apply the algorithm in this regime, one needs to stabilize it using the celebrated log-sum-exp trick. This follows from noticing that, similarly to the minimum operator, one has

$$\min_{\mathrm{b}}^{\varepsilon}(h - \mathrm{cst}) = \min_{\mathrm{b}}^{\varepsilon}(h) - \mathrm{cst}$$

and to replace the computation of $\min_{\mathrm{b}}^{\varepsilon}(h)$ by its stabilized version (equal when using infinite precision computation) $\min_{\mathrm{b}}^{\varepsilon}(h - \min(h)) + \min(h)$.

Algorithm 8.2: Log-domain Sinkhorn by soft transforms

Input: Positive weights a, b, cost matrix C, regularization $\varepsilon > 0$, tolerance tol, iteration budget $L \geqslant 1$.
Output: Entropic coupling P computed from stabilized potentials.
Initialize: Set $\mathrm{g}^{(0)} = 0$, $\eta^{(0)} = +\infty$, and $\ell = 0$.
While $\eta^{(\ell)} >$ tol and $\ell < L$ **do**:
  Set $\ell \leftarrow \ell + 1$.
  For $i = 1, \dots, n$ **do**:
    Set $M_i = \max_j\{\mathrm{g}_j^{(\ell-1)} - \mathrm{C}_{i,j}\}$.
    Set $\mathrm{f}_i^{(\ell)} = -M_i - \varepsilon \log \sum_j \mathrm{b}_j \exp((\mathrm{g}_j^{(\ell-1)} - \mathrm{C}_{i,j} - M_i)/\varepsilon)$.
  For $j = 1, \dots, m$ **do**:
    Set $N_j = \max_i\{\mathrm{f}_i^{(\ell)} - \mathrm{C}_{i,j}\}$.
    Set $\mathrm{g}_j^{(\ell)} = -N_j - \varepsilon \log \sum_i \mathrm{a}_i \exp((\mathrm{f}_i^{(\ell)} - \mathrm{C}_{i,j} - N_j)/\varepsilon)$.
  Set $\mathrm{P}_{i,j}^{(\ell)} = \mathrm{a}_i \mathrm{b}_j \exp((\mathrm{f}_i^{(\ell)} + \mathrm{g}_j^{(\ell)} - \mathrm{C}_{i,j})/\varepsilon)$.
  Set $\eta^{(\ell)} = \max\{\|\mathrm{P}^{(\ell)}\mathbb{1}_m - \mathrm{a}\|_1, \|(\mathrm{P}^{(\ell)})^\top \mathbb{1}_n - \mathrm{b}\|_1\}$.
Return $\mathrm{P}^{(\ell)}$.

Figure 8.9 visualizes the corresponding soft minimum: decreasing ε sharpens the smooth best response toward the hard c-transform envelope.

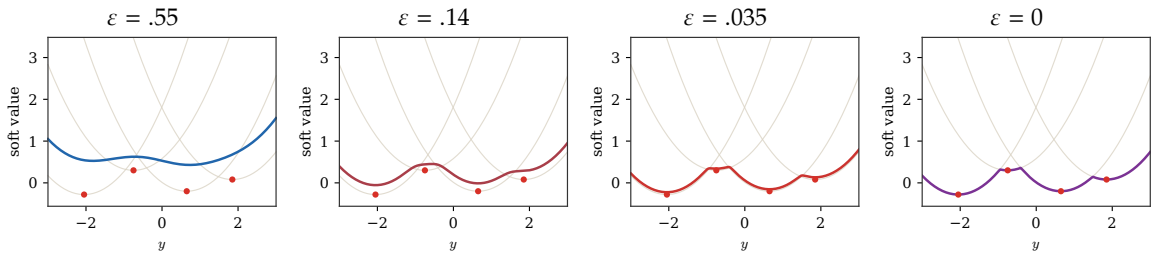


Figure 8.9: Soft c-transforms for decreasing temperatures. The hard transform is the lower envelope of shifted cost functions, while a positive ε replaces the pointwise minimum by a log-sum-exp soft minimum. As ε decreases, the soft transform sharpens and approaches the non-smooth hard envelope.

Continuous dual and soft-transforms. The continuous dual formula is not a discretization heuristic: it follows from the same entropy conjugacy as its matrix counterpart, with the two marginal constraints represented by continuous potentials.

Proposition 8.20 (Continuous entropic duality). *Let $\mathcal{X}$ and $\mathcal{Y}$ be compact metric spaces, let $\alpha \in \mathcal{P}(\mathcal{X})$ and $\beta \in \mathcal{P}(\mathcal{Y})$, and let $c \in C(\mathcal{X} \times \mathcal{Y})$. For every $\varepsilon > 0$, the KL-regularized problem (8.13) satisfies*

$$\mathcal{L}_c^{\varepsilon}(\alpha, \beta) = \sup_{f \in C(\mathcal{X}),\, g \in C(\mathcal{Y})} \mathcal{D}_{\varepsilon}(f, g), \tag{8.22}$$

where the concave dual objective is

$$\mathcal{D}_{\varepsilon}(f, g) := \int_{\mathcal{X}} f(x) \mathrm{d}\alpha(x) + \int_{\mathcal{Y}} g(y) \mathrm{d}\beta(y) - \varepsilon \int_{\mathcal{X}\times\mathcal{Y}} \left(e^{\frac{f(x)+g(y)-c(x,y)}{\varepsilon}} - 1 \right) \mathrm{d}\alpha(x) \mathrm{d}\beta(y). \tag{8.23}$$

If $\pi^\star$ and $(f^\star, g^\star)$ are primal and dual optimizers, then

$$\frac{\mathrm{d}\pi^\star}{\mathrm{d}(\alpha \otimes \beta)}(x, y) = \exp\left(\frac{f^\star(x) + g^\star(y) - c(x, y)}{\varepsilon} \right) \qquad (\alpha \otimes \beta)\text{-a.e.} \tag{8.24}$$

Proof. Write $\xi = \alpha \otimes \beta$ and $\varphi(r) = r \log r - r + 1$ for $r \geqslant 0$. Since the primal objective is infinite unless $\pi \ll \xi$, put $r = \mathrm{d}\pi / \mathrm{d}\xi$. The marginal constraints are equivalent to

$$\int_{\mathcal{Y}} r(x, y) \mathrm{d}\beta(y) = 1 \quad \alpha\text{-a.e.}, \qquad \int_{\mathcal{X}} r(x, y) \mathrm{d}\alpha(x) = 1 \quad \beta\text{-a.e.}$$

Introduce potentials f and g for these two constraints. The Lagrangian is

$$\int_{\mathcal{X}} f \mathrm{d}\alpha + \int_{\mathcal{Y}} g \mathrm{d}\beta + \int_{\mathcal{X}\times\mathcal{Y}} \left[\varepsilon \varphi(r(x, y)) + \big(c(x, y) - f(x) - g(y)\big) r(x, y) \right] \mathrm{d}\xi(x, y).$$

The scalar conjugate of φ is $\varphi^*(s) = e^s - 1$. Therefore pointwise minimization over $r \geqslant 0$ gives

$$\inf_{r \geqslant 0} \{ \varepsilon \varphi(r) + (c - f - g) r \} = -\varepsilon \varphi^*\left(\frac{f + g - c}{\varepsilon} \right) = -\varepsilon \left(e^{(f+g-c)/\varepsilon} - 1 \right).$$

Integrating this identity yields $\mathcal{D}_{\varepsilon}(f, g)$ and proves weak duality. Strong duality follows from Fenchel–Rockafellar entropy duality in the exponential Orlicz pair [336, 108]: the feasible density $r \equiv 1$, corresponding to the product coupling ξ, is strictly positive and supplies the qualification point, while the boundedness of c makes its objective finite. Because c is continuous on the compact product, the dual supremum can be restricted to continuous potentials; equivalently, alternating soft c-transforms regularize bounded potentials into continuous ones without decreasing the dual value. This proves (8.22).

Finally, equality in the pointwise Fenchel inequality forces $r^\star(x, y)$ to be the unique minimizer of the scalar problem above. Its first-order condition is

$$\varepsilon \log r^\star(x, y) + c(x, y) - f^\star(x) - g^\star(y) = 0.$$

Exponentiation gives (8.24). □

This is the smooth counterpart of the hard feasibility constraint $f \oplus g \leqslant c$ from the Kantorovich dual: violations are penalized exponentially and disappear in the limit $\varepsilon \to 0$.

The corresponding soft c-transforms are the exact block maximizers of this dual objective. They are the continuous log-integral counterparts of Definition 8.19.

Definition 8.21 (Continuous soft c-transforms). For $f \in C(\mathcal{X})$ and $g \in C(\mathcal{Y})$, define

$$f^{c,\varepsilon}(y) := -\varepsilon \log \left(\int_{\mathcal{X}} e^{\frac{f(x)-c(x,y)}{\varepsilon}} \mathrm{d}\alpha(x) \right), \qquad y \in \mathcal{Y}, \tag{8.25}$$

$$g^{\bar{c},\varepsilon}(x) := -\varepsilon \log \left(\int_{\mathcal{Y}} e^{\frac{g(y)-c(x,y)}{\varepsilon}} \mathrm{d}\beta(y) \right), \qquad x \in \mathcal{X}. \tag{8.26}$$

In the case of discrete measures, these formulas reduce to (8.21). The same calculus also gives the entropic semi-discrete problem: when one measure is discrete, one alternates between a finite-dimensional potential and an integral soft transform, often estimated stochastically.

Proposition 8.22 (Existence and uniqueness of entropic dual potentials)**.** *Assume that* $\mathcal{X} = \mathrm{supp}(\alpha)$ *and* $\mathcal{Y} = \mathrm{supp}(\beta)$ *are compact and that* c *is continuous. The dual problem* (8.22) *has solutions, and its set of solutions is*

$$(f^\star + \lambda, g^\star - \lambda), \qquad \lambda \in \mathbb{R}.$$

Proof. Normalize potentials by imposing $\int f \mathrm{d}\alpha = 0$. Replacing any pair (f, g) by the corresponding soft c-transforms does not decrease the dual objective, because each soft transform is the exact maximizer in one block variable. For transformed potentials, the oscillations are bounded by the oscillation of the cost: $\|f\|_V + \|g\|_V \leqslant 2(\sup c - \inf c)$. Moreover the modulus of continuity of the soft transforms is controlled by that of c; for instance

$$\left|g^{\bar{c},\varepsilon}(x) - g^{\bar{c},\varepsilon}(x')\right| \leqslant \sup_y |c(x,y) - c(x',y)|\,.$$

After normalization, maximizing sequences are therefore uniformly bounded and equicontinuous. Arzelà–Ascoli gives a uniformly converging subsequence, and continuity of (8.23) gives a maximizer.

For uniqueness, use strict convexity of $H \mapsto \int e^{H/\varepsilon}\mathrm{d}(\alpha\otimes\beta)$ on the image of $(f,g) \mapsto H = f \oplus g - c$, modulo constants. If two optimal pairs exist, their midpoint is also optimal; strict convexity forces the two functions $f \oplus g$ to agree $\alpha\otimes\beta$-almost everywhere. Full support and continuity then imply equality everywhere on $\mathcal{X}\times\mathcal{Y}$, so $f - f'$ is constant and $g - g'$ is the opposite constant. Without replacing the ambient spaces by the marginal supports, uniqueness has this meaning only on those supports; values outside them do not enter the dual objective. □

> **Remark 8.23 (Convexity properties of soft transforms).** The log-sum-exp part behaves like a smoothed maximum and preserves convexity. Since the soft transform takes the negative of this quantity after inserting the cost, it preserves the usual c-concavity structure. In particular, for the bilinear cost $c(x,y) = -\langle x, y\rangle$, the transform $f^{c,\varepsilon}$ is concave for any f. Therefore, for the quadratic cost $c(x,y) = \|x-y\|^2/2$, the optimal potentials have the form $f^\star(x) = \|x\|^2/2 - \varphi^\star(x)$ and $g^\star(y) = \|y\|^2/2 - \psi^\star(y)$, where $\varphi^\star$ and $\psi^\star$ are convex.

Dual Sinkhorn for general measures. The soft transforms are not only regularized analogues of hard c-transforms: they are the exact block-maximization steps of the continuous dual objective (8.23). Indeed, for fixed g and a perturbation $h \in C(\mathcal{X})$,

$$\frac{\mathrm{d}}{\mathrm{d}s}\mathcal{D}_\varepsilon(f + sh, g)\Big|_{s=0} = \int_{\mathcal{X}} h(x)\left[1 - e^{f(x)/\varepsilon}\int_{\mathcal{Y}} e^{(g(y)-c(x,y))/\varepsilon}\mathrm{d}\beta(y)\right]\mathrm{d}\alpha(x).$$

Consequently, maximizing first over f and then over g gives the alternating updates

$$f^{(\ell+1)} = (g^{(\ell)})^{\bar{c},\varepsilon}, \qquad g^{(\ell+1)} = (f^{(\ell+1)})^{c,\varepsilon}. \tag{8.27}$$

The decorations record the domains: the $\bar{c}$-transform (8.26) sends a potential on $\mathcal{Y}$ to one on $\mathcal{X}$, while the c-transform (8.25) sends a potential on $\mathcal{X}$ to one on $\mathcal{Y}$.

This dual iteration is exactly the logarithmic form of the continuous scaling iteration (8.16). More precisely, setting

$$u^{(\ell)} = e^{f^{(\ell)}/\varepsilon}, \qquad v^{(\ell)} = e^{g^{(\ell)}/\varepsilon}$$

and using the kernel operators (8.14), exponentiation of (8.27) yields

$$u^{(\ell+1)} = \frac{1}{\mathcal{K}_\varepsilon v^{(\ell)}}, \qquad v^{(\ell+1)} = \frac{1}{\mathcal{K}_\varepsilon^* u^{(\ell+1)}}.$$

Thus the coupling density reconstructed from the current potentials is $u^{(\ell)}(x)k_\varepsilon(x,y)v^{(\ell)}(y)$ with respect to $\alpha\otimes\beta$, exactly as in the scaling law (8.15). At a fixed point, its two marginals are α and β; equivalently, the potentials jointly maximize the continuous dual.

Neural dual solvers. The convex-potential structure of the preceding remark suggests a sample-based alternative to evaluating soft transforms on a grid or on all pairs of samples. For the bilinear cost $c(x,y) = -\langle x, y\rangle$, the signs of the dual potentials are convex: writing $\Phi = -f$ and $\Psi = -g$, the zero-temperature constraint $f(x) + g(y) \leqslant -\langle x, y\rangle$ becomes the Fenchel-type inequality

$$\Phi(x) + \Psi(y) \geqslant \langle x, y\rangle.$$

For the quadratic cost this is the same statement after subtracting the quadratic terms, as in Remark 8.23. One can therefore maximize the continuous dual (8.23) over parameterized convex potentials, estimating the integrals by stochastic samples.

A useful parameterization is given by input-convex neural networks (ICNNs) [13, 365]. The construction mirrors elementary closure rules for convex functions: nonnegative linear combinations preserve convexity, composition with a convex nondecreasing scalar nonlinearity preserves convexity, and the ReLU $r \mapsto \max(r, 0)$ is both convex and nondecreasing. Thus a feed-forward network with nonnegative hidden-to-hidden weights and affine skip connections from the input defines a convex function of its input. This gives a flexible cone of convex trial potentials, although the finite-dimensional optimization over the network weights is not a convex optimization problem. Universal-approximation statements must also be read with this distinction in mind: max-affine functions are dense among continuous convex functions on compact convex sets, and ICNN-type architectures are designed to inherit this approximation principle. General ReLU universal approximation results, such as the width $d+1$ theorem on compact subsets of $\mathbb{R}^d$ [268], provide useful background but do not by themselves enforce convexity. In practice, neural dual solvers trade exact Sinkhorn scaling for amortized stochastic optimization of the dual potentials.

> **Remark 8.24 (Gaussian marginals).** For $c(x,y) = \|x-y\|^2$ and Gaussian marginals, the soft transforms preserve quadratic functions, because products and convolutions of Gaussian functions remain Gaussian. Hence optimal entropic potentials are quadratic and the optimal entropic coupling is Gaussian. Section 9.6 makes this finite-dimensional closure explicit.

8.6 Path-Space Schrödinger Problem

Schrödinger's reciprocal problem is naturally posed on paths rather than on endpoint pairs. A noisy reference dynamics defines a probability law on trajectories; after optimizing out the conditional law of a path given its endpoints, only the endpoint coupling remains, and the Sinkhorn problem appears.

Unregularized path-space transport. Throughout this section, both endpoint measures live on the same state space $\mathcal{X}$; this is the setting relevant to dynamic transport and Brownian bridges. The static coupling formulation above also makes sense between two different spaces $\mathcal{X}$ and $\mathcal{Y}$. The path-space formulation optimizes directly over laws of trajectories.

Definition 8.25 (Unregularized path-space transport)**.** Let $\Omega = C([0,1];\mathcal{X})$, let $e_t(\omega) = \omega_t$, and let $\mathcal{A} : \Omega \to [0,+\infty]$ be a path action. Define

$$\mathrm{PT}_{\mathcal{A}}(\alpha,\beta) := \inf_{M \in \mathcal{P}(\Omega)} \left\{ \int_\Omega \mathcal{A}(\omega)\,\mathrm{d}M(\omega) \;;\; (e_0)_\sharp M = \alpha,\ (e_1)_\sharp M = \beta \right\}. \tag{8.28}$$

A feasible M is a law of random trajectories whose initial and final laws are α and β. For the quadratic Wasserstein geometry on $\mathbb{R}^d$, one takes

$$\mathcal{A}(\omega) = \begin{cases} \displaystyle\int_0^1 \|\dot\omega_t\|^2\,\mathrm{d}t, & \text{if } \omega \text{ is absolutely continuous,} \\ +\infty, & \text{otherwise.} \end{cases}$$

This is the Lagrangian version of the Benamou–Brenier formulation recalled in the path-space formulation of Section 14.2.

The endpoint cost induced by the action is

$$c_{\mathcal{A}}(x,y) := \inf_{\omega \in \Omega} \{\mathcal{A}(\omega) \;;\; e_0(\omega) = x,\ e_1(\omega) = y\}. \tag{8.29}$$

For the quadratic action above, the minimizing path is the straight segment $\omega_t = (1-t)x + ty$, and $c_{\mathcal{A}}(x,y) = \|x-y\|^2$.

Proposition 8.26 (Endpoint reduction of path-space transport)**.** *Assume that minimizing paths in* (8.29) *can be selected measurably, or more generally that the infimum can be approximated by measurable selections. Then* (8.28) *has the same value as the Kantorovich problem*

$$\inf_{\pi \in \mathcal{U}(\alpha,\beta)} \int_{\mathcal{X}\times\mathcal{X}} c_{\mathcal{A}}(x,y)\,\mathrm{d}\pi(x,y).$$

Moreover, if $\pi^\star$ is an optimal endpoint coupling and $\omega^{x,y}$ is an optimal path from x to y, then

$$M^\star = \int_{\mathcal{X}\times\mathcal{X}} \delta_{\omega^{x,y}}\, \mathrm{d}\pi^\star(x,y)$$

is an optimal path law.

Proof. Let M be any feasible path law and set $\pi = (e_0, e_1)_\sharp M$. Then $\pi \in \mathcal{U}(\alpha,\beta)$ and, by the definition of $c_{\mathcal{A}}$,

$$\int_\Omega \mathcal{A}(\omega)\,\mathrm{d}M(\omega) \geqslant \int_{\mathcal{X}\times\mathcal{X}} c_{\mathcal{A}}(x,y)\,\mathrm{d}\pi(x,y).$$

This proves that the path-space value is at least the Kantorovich value. Conversely, given a coupling π and a measurable selection $(x,y)\mapsto \omega^{x,y}$ with action arbitrarily close to $c_{\mathcal{A}}(x,y)$, the mixture of Dirac path laws $M = \int \delta_{\omega^{x,y}}\,\mathrm{d}\pi(x,y)$ has endpoints α and β and action equal, up to the selected approximation error, to $\int c_{\mathcal{A}}\,\mathrm{d}\pi$. Optimizing over π and letting the approximation error vanish proves equality. If exact minimizing paths are selected for an optimal $\pi^\star$, the displayed $M^\star$ is optimal. □

Thus the classical coupling problem can be read as a path problem where the endpoints are chosen first and the connecting path is then selected with minimal action. The Schrödinger problem changes exactly this last step: between two endpoints, it keeps the random fluctuations of a reference dynamics instead of collapsing onto a deterministic least-action path.

Entropic path-space problem. Let $\mathcal{R}^\varepsilon \in \mathcal{P}(\Omega)$ be a reference path law, for instance a Brownian or Langevin dynamics at noise level ε. The corresponding entropy projection defines the dynamic Schrödinger problem.

Definition 8.27 (Schrödinger bridge)**.**

$$\mathrm{SB}_\varepsilon(\alpha,\beta) := \inf_{M\in\mathcal{P}(\Omega)} \left\{ \varepsilon\,\mathrm{KL}(M|\mathcal{R}^\varepsilon)\ ;\ (e_0)_\sharp M = \alpha,\ (e_1)_\sharp M = \beta \right\}. \tag{8.30}$$

It asks for the most likely path law, relative to the prior dynamics $\mathcal{R}^\varepsilon$, among all path laws matching the observed endpoint marginals. This viewpoint goes back to Schrödinger's reciprocal problem [481] and is surveyed in modern OT language in [336, 337]; stochastic-control formulations are developed in [131].

Stochastic-control interpretation. For Brownian references, the entropy projection has a direct control meaning. To fix constants, suppose that $\mathcal{R}^\varepsilon$ is the law of

$$\mathrm{d}X_t = \sqrt{\varepsilon}\,\mathrm{d}B_t, \qquad X_0 \sim \alpha,$$

so that the generator is $(\varepsilon/2)\Delta$. Under the standard finite-entropy assumptions, a path law M with the same initial marginal and $M \ll \mathcal{R}^\varepsilon$ can be represented, in the weak sense, by an adapted drift u_t of finite energy,

$$\mathrm{d}X_t = u_t\,\mathrm{d}t + \sqrt{\varepsilon}\,\mathrm{d}B_t, \qquad X_0 \sim \alpha.$$

Conversely, such a drift defines an absolutely continuous change of law when the usual Girsanov integrability conditions hold. Girsanov's formula then gives

$$\mathrm{KL}(M|\mathcal{R}^\varepsilon) = \frac{1}{2\varepsilon}\mathbb{E}_M \int_0^1 \|u_t\|^2\,\mathrm{d}t, \qquad \varepsilon\,\mathrm{KL}(M|\mathcal{R}^\varepsilon) = \frac{1}{2}\mathbb{E}_M \int_0^1 \|u_t\|^2\,\mathrm{d}t.$$

If the reference initial law is not α, an additional term $\varepsilon\,\mathrm{KL}(\alpha|\mathcal{R}_0^\varepsilon)$ appears, but it is fixed by the endpoint constraint. Hence, up to this fixed endpoint cost, the Schrödinger bridge can be read as

$$\inf_u \left\{ \frac{1}{2}\mathbb{E}\int_0^1 \|u_t\|^2\,\mathrm{d}t\ ;\ \mathrm{d}X_t = u_t\,\mathrm{d}t + \sqrt{\varepsilon}\,\mathrm{d}B_t, \quad X_0\sim\alpha,\ X_1\sim\beta \right\},$$

namely as the least energetic change of drift that steers the Brownian prior from α to β. The optimizer may be chosen Markovian. If h_t denotes the positive space-time harmonic Schrödinger factor associated with the terminal constraint, then the optimal feedback drift has the Doob-transform form

$$u_t^\star(x) = \varepsilon\nabla \log h_t(x).$$

Equivalently, with the value potential $\varphi_t = \varepsilon \log h_t$, one has $u_t^\star = \nabla\varphi_t$ and φ_t solves the viscous Hamilton–Jacobi equation $\partial_t\varphi_t + \frac{1}{2}\|\nabla\varphi_t\|^2 + \frac{\varepsilon}{2}\Delta\varphi_t = 0$. This control viewpoint and the endpoint-coupling reduction below are two faces of the same object: the controlled diffusion describes the whole path law, whereas the static Sinkhorn coupling records only its two endpoints.

Viscous Benamou–Brenier formulations. The Schrödinger interpolation also admits dynamic optimal-control descriptions, which are viscous analogues of the Benamou–Brenier formula. In this paragraph, v_t denotes the forward drift called u_t above, while u_t denotes the associated current velocity. Write $\sigma > 0$ for the diffusion intensity; thus the uncontrolled noise is $\sqrt{\sigma}\,\mathrm{d}B_t$ and its generator is $(\sigma/2)\Delta$. Formally, for smooth positive densities the controlled Fokker–Planck equation is

$$\partial_t\rho_t + \operatorname{div}(\rho_t v_t) = \frac{\sigma}{2}\Delta\rho_t,$$

one minimizes, among curves joining the prescribed endpoint densities, the kinetic action

$$\int_0^1 \int \frac{1}{2}\|v_t(x)\|^2\rho_t(x)\mathrm{d}x\mathrm{d}t.$$

Equivalently, one can absorb the diffusion into the velocity by writing

$$u_t = v_t - \frac{\sigma}{2}\nabla\log\rho_t, \qquad \partial_t\rho_t + \operatorname{div}(\rho_t u_t) = 0.$$

Expanding $v_t = u_t + \frac{\sigma}{2}\nabla\log\rho_t$ and using the continuity equation for (ρ_t, u_t) gives

$$\begin{aligned}\int_0^1 \int \frac{1}{2}\|v_t\|^2\rho_t\,\mathrm{d}x\,\mathrm{d}t = \int_0^1 \int \left(\frac{1}{2}\|u_t\|^2 + \frac{\sigma^2}{8}\|\nabla\log\rho_t\|^2\right)\rho_t\,\mathrm{d}x\,\mathrm{d}t \\ + \frac{\sigma}{2}\left[\int \rho_1\log\rho_1\,\mathrm{d}x - \int \rho_0\log\rho_0\,\mathrm{d}x\right].\end{aligned}$$

Since the entropy term depends only on the prescribed endpoints, the same minimizers are obtained from the modified Benamou–Brenier action

$$\int_0^1 \int \left(\frac{1}{2}\|u_t(x)\|^2 + \frac{\sigma^2}{8}\|\nabla\log\rho_t(x)\|^2\right)\rho_t(x)\mathrm{d}x\mathrm{d}t.$$

Thus the Schrödinger bridge is a least-action interpolation with both transport kinetic energy and a Fisher-information penalty. If one instead writes the viscous equation with diffusion coefficient $\sigma\Delta\rho_t$, the same formula is obtained after replacing σ above by 2σ, so the Fisher coefficient becomes $\sigma^2/2$.

The reduction to endpoint couplings follows from disintegration. We write

$$\mathcal{R}^\varepsilon(\mathrm{d}\omega) = \int \mathcal{R}^{\varepsilon,x,y}(\mathrm{d}\omega)\,\mathcal{R}^\varepsilon_{0,1}(\mathrm{d}x,\mathrm{d}y), \qquad \mathcal{R}^\varepsilon_{0,1} := (e_0,e_1)_\sharp\mathcal{R}^\varepsilon,$$

where $\mathcal{R}^{\varepsilon,x,y}$ is the reference bridge conditioned on endpoints (x,y). Similarly, for any feasible M, write

$$M(\mathrm{d}\omega) = \int M^{x,y}(\mathrm{d}\omega)\,\pi(\mathrm{d}x,\mathrm{d}y), \qquad \pi := (e_0,e_1)_\sharp M.$$

Proposition 8.28 (Endpoint reduction of the Schrödinger problem)**.** *Assume that the regular conditional laws above exist and that the relative-entropy chain rule applies, with value $+\infty$ when absolute continuity fails. Then*

$$\mathrm{SB}_\varepsilon(\alpha,\beta) = \inf_{\pi\in\mathcal{U}(\alpha,\beta)} \varepsilon\,\mathrm{KL}(\pi|\mathcal{R}^\varepsilon_{0,1}). \tag{8.31}$$

For a fixed endpoint coupling π with finite $\mathrm{KL}(\pi|\mathcal{R}^\varepsilon_{0,1})$, the minimizing path law is the mixture of reference bridges

$$M^\pi = \int \mathcal{R}^{\varepsilon,x,y}\,\mathrm{d}\pi(x,y). \tag{8.32}$$

Consequently, if $\pi^\star$ solves (8.31), then $M^{\pi^\star}$ solves the path-space problem (8.30).

Proof. If M has finite relative entropy with respect to $\mathcal{R}^\varepsilon$, then $\pi = (e_0, e_1)_\sharp M$ is necessarily absolutely continuous with respect to $\mathcal{R}^\varepsilon_{0,1}$. The chain rule for relative entropy gives

$$\mathrm{KL}(M|\mathcal{R}^\varepsilon) = \mathrm{KL}(\pi|\mathcal{R}^\varepsilon_{0,1}) + \int_{\mathcal{X}\times\mathcal{X}} \mathrm{KL}(M^{x,y}|\mathcal{R}^{\varepsilon,x,y})\,\mathrm{d}\pi(x,y). \tag{8.33}$$

The second term is nonnegative and vanishes exactly when $M^{x,y} = \mathcal{R}^{\varepsilon,x,y}$ for π-almost every (x,y). Thus, once an endpoint coupling π with finite $\mathrm{KL}(\pi|\mathcal{R}^\varepsilon_{0,1})$ is fixed, the best path law is (8.32), and the remaining minimization is precisely (8.31). If no such endpoint coupling exists, both sides are $+\infty$. □

This proposition makes the connection between the dynamic and static views precise. The static Schrödinger problem stores only the endpoints of the optimal random trajectories; the full bridge is recovered by filling each transported endpoint pair with the corresponding reference bridge. In the zero-noise limit, these bridges concentrate on least-action paths and one recovers the unregularized path-space transport problem, hence the Monge–Kantorovich problem [336].

Brownian bridges and Sinkhorn couplings. For $\mathcal{X} = \mathbb{R}^d$, take $\mathcal{R}^\varepsilon$ to be a Brownian reference dynamics. With the convention $\mathrm{d}X_t = \sqrt{\varepsilon}\,\mathrm{d}B_t$ used above, its unit-time transition density is proportional to $e^{-\|x-y\|^2/(2\varepsilon)}$, so multiplying the endpoint entropy by ε produces the static cost $\|x-y\|^2/2$. Equivalently, after renaming the static temperature by a factor of two, the heat-kernel density has the form

$$p_\varepsilon(x,y) \propto \exp\left(-\frac{\|x-y\|^2}{\varepsilon}\right).$$

More generally, suppose that the endpoint prior can be written, up to a normalization constant independent of π, as

$$\mathcal{R}^\varepsilon_{0,1}(\mathrm{d}x,\mathrm{d}y) \propto \exp\left(-\frac{c(x,y)}{\varepsilon}\right)\alpha(\mathrm{d}x)\beta(\mathrm{d}y).$$

This includes the usual heat-kernel reference after rewriting it with respect to $\alpha\otimes\beta$, whenever the one-time endpoint densities are fixed and mutually absolutely continuous. The additional one-body density factors only add constants under the marginal constraints. Then, for every $\pi \in \mathcal{U}(\alpha,\beta)$,

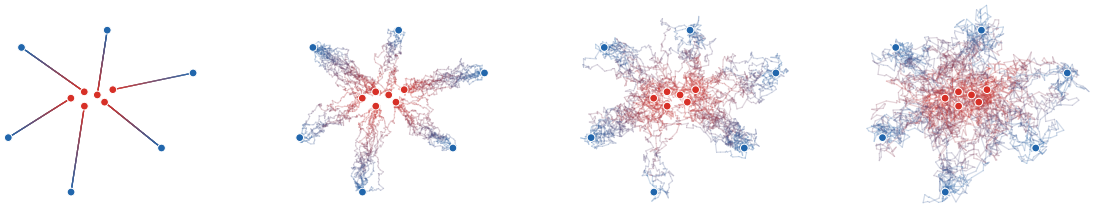
$$\varepsilon\,\mathrm{KL}(\pi|\mathcal{R}^\varepsilon_{0,1}) = \int c(x,y)\,\mathrm{d}\pi(x,y) + \varepsilon\,\mathrm{KL}(\pi|\alpha\otimes\beta) + \text{constant},$$

where the constant does not depend on π. Hence (8.31) is exactly the continuous Sinkhorn problem (8.13), up to an additive constant in the value. The optimal static coupling $\pi^\star_\varepsilon$ is the endpoint law of the most likely controlled noisy dynamics, while the associated path law is

$$M^\star_\varepsilon = \int \mathcal{R}^{\varepsilon,x,y}\,\mathrm{d}\pi^\star_\varepsilon(x,y).$$

In words, Sinkhorn computes which endpoints should be paired; the path-space Schrödinger bridge then connects each paired endpoint by a Brownian bridge rather than by a deterministic straight line.

Figure 8.10 illustrates this endpoint-to-path lifting on a small discrete example. The red atoms form a compact source cloud and the blue atoms surround them. The first panel uses the unregularized OT coupling and zero bridge noise; the next panels increase ε, which simultaneously softens the endpoint coupling and amplifies the Brownian fluctuations between paired endpoints.



$\varepsilon = 0$ $\varepsilon = 0.035$ $\varepsilon = 0.10$ $\varepsilon = 0.28$

Figure 8.10: Endpoint couplings lifted to Brownian bridges. Each panel transports six equally weighted red atoms concentrated near the center to six blue atoms distributed around them. For the displayed value of ε, the endpoint coupling is either the exact quadratic OT plan ($\varepsilon = 0$) or the entropic Sinkhorn plan ($\varepsilon > 0$). A total of 60 paths is sampled by a multinomial allocation with probabilities $\pi_{i,j}$, and each selected pair (x_i, y_j) is filled by a Brownian bridge with covariance proportional to $\varepsilon t(1-t)$. Colors encode time from red to blue.

8.7 Marginal-Dependent Problems

The balanced Sinkhorn problem fixes both marginals exactly. Many nearby models instead optimize the transported marginals, but only through penalties or constraints applied separately to each marginal. The useful point, emphasized by the generalized scaling algorithms of Chizat, Peyré, Schmitzer and Vialard [140], is that entropic OT remains a diagonal-scaling problem whenever these marginal terms admit simple KL-proximal maps.

Let $\mathcal{F}$ and $\mathcal{G}$ be proper convex lower semicontinuous functionals on finite nonnegative measures on $\mathcal{X}$ and $\mathcal{Y}$. The unregularized marginal-dependent transport problem is

$$\inf_{\pi \in \mathcal{M}_+(\mathcal{X}\times\mathcal{Y})} \int c(x,y)\mathrm{d}\pi(x,y) + \mathcal{F}(\pi_1) + \mathcal{G}(\pi_2), \tag{8.34}$$

where $\pi_1 = (\mathrm{p}_\mathcal{X})_\sharp \pi$ and $\pi_2 = (\mathrm{p}_\mathcal{Y})_\sharp \pi$ are the two marginals of π. Entropic regularization turns this problem into the scaling-friendly form below.

Definition 8.29 (Marginal-dependent entropic transport)**.** For reference measures α, β, proper convex marginal functionals $\mathcal{F}, \mathcal{G}$, and $\varepsilon > 0$, define

$$\inf_{\pi \in \mathcal{M}_+(\mathcal{X}\times\mathcal{Y})} \int c(x,y)\mathrm{d}\pi(x,y) + \mathcal{F}(\pi_1) + \mathcal{G}(\pi_2) + \varepsilon\,\mathrm{KL}(\pi|\alpha\otimes\beta). \tag{8.35}$$

For discrete reference weights a, b and convex functions $\mathsf{F} : \mathbb{R}^n_+ \to \mathbb{R}\cup\{+\infty\}$, $\mathsf{G} : \mathbb{R}^m_+ \to \mathbb{R}\cup\{+\infty\}$, its matrix form is

$$\inf_{\mathsf{P}\in\mathbb{R}^{n\times m}_+} \langle \mathsf{C}, \mathsf{P}\rangle + \mathsf{F}(\mathsf{P}\mathbb{1}_m) + \mathsf{G}(\mathsf{P}^\top \mathbb{1}_n) + \varepsilon \mathrm{KL}(\mathsf{P}|\mathsf{a}\otimes\mathsf{b}). \tag{8.36}$$

In this subsection, when the total mass of π is not fixed, the KL term is understood in the generalized sense associated with $\varphi(s) = s\log s - s + 1$. Thus, if $\lambda = \alpha\otimes\beta$ and $\pi = \rho\lambda + \pi^\perp$ is the Lebesgue decomposition of π with respect to λ, then

$$\mathrm{KL}(\pi|\lambda) = \int \big(\rho\log\rho - \rho + 1\big)\mathrm{d}\lambda$$

with value $+\infty$ when $\pi^\perp \neq 0$. On probability couplings this coincides with the usual relative entropy.

Several standard constructions fit this template. Balanced OT is recovered by taking $\mathcal{F} = \iota_{\{\alpha\}}$ and $\mathcal{G} = \iota_{\{\beta\}}$. Unbalanced OT replaces these hard indicators by marginal divergences, as developed later in Section 11.1. An entropic JKO step fixes the first marginal to the previous iterate and puts the energy on the second marginal, for instance $\mathcal{F} = \iota_{\{\alpha_t\}}$ and $\mathcal{G} = E$, with cost $c/(2\tau)$; this is the static counterpart of the minimizing-movement schemes of Chapter 15. Barycenters are the multi-coupling extension: several couplings share one unknown marginal and are treated by the generalized Sinkhorn updates of Section 12.1.

In finite dimension, the second part of Definition 8.29 gives the corresponding matrix problem, with reference weights satisfying $a_i, b_j > 0$. Equivalently, if $K_{i,j} = a_i b_j e^{-C_{i,j}/\varepsilon}$, the terms involving C can be absorbed into the Gibbs reference and the problem is, up to the additive constant $\varepsilon \sum_{i,j}(a_i b_j - K_{i,j})$,

$$\inf_{P \geqslant 0} F(P\mathbb{1}_m) + G(P^\top \mathbb{1}_n) + \varepsilon \mathrm{KL}(P|K).$$

Proposition 8.30 (Dual and scaling for marginal penalties)**.** *Assume $a_i, b_j > 0$, $\varepsilon > 0$, and a Fenchel qualification condition, for instance the existence of a matrix $P > 0$ such that $P\mathbb{1}_m$ and $P^\top \mathbb{1}_n$ belong to the relative interiors of* $\mathrm{dom}(F)$ *and* $\mathrm{dom}(G)$*. The Fenchel dual of* (8.36) *is*

$$\sup_{f \in \mathbb{R}^n, g \in \mathbb{R}^m} -F^*(-f) - G^*(-g) - \varepsilon \sum_{i,j} a_i b_j \left[\exp\left(\frac{f_i + g_j - C_{i,j}}{\varepsilon} \right) - 1 \right]. \tag{8.37}$$

Equivalently, up to the additive constant $\varepsilon \sum_{i,j} a_i b_j$, the last term is

$$-\varepsilon \sum_{i,j} K_{i,j} \exp\left(\frac{f_i + g_j}{\varepsilon} \right).$$

If $u = e^{f/\varepsilon}$ and $v = e^{g/\varepsilon}$, exact block ascent in the two dual variables is the generalized Sinkhorn cycle

$$\begin{aligned} r &\leftarrow \mathrm{prox}^{\mathrm{KL}}_{F/\varepsilon}(Kv), & u &\leftarrow r \oslash (Kv), \\ s &\leftarrow \mathrm{prox}^{\mathrm{KL}}_{G/\varepsilon}(K^\top u), & v &\leftarrow s \oslash (K^\top u), \\ P &= \mathrm{diag}(u) K \,\mathrm{diag}(v), \end{aligned} \tag{8.38}$$

where divisions are entrywise and

$$\mathrm{prox}^{\mathrm{KL}}_{h}(z) := \operatorname*{argmin}_{r \in \mathbb{R}^d_+} h(r) + \mathrm{KL}(r|z). \tag{8.39}$$

In particular, $\mathrm{prox}^{\mathrm{KL}}_{F/\varepsilon}(z)$ *is the minimizer of* $F(r) + \varepsilon \mathrm{KL}(r|z)$.

Proof. Introduce independent variables r and s for the two marginals and use dual variables f, g for the constraints $r = P\mathbb{1}_m$ and $s = P^\top \mathbb{1}_n$. The Lagrangian contains

$$F(r) + \langle f, r \rangle + G(s) + \langle g, s \rangle + \langle C, P \rangle + \varepsilon \mathrm{KL}(P|a \otimes b) - \langle f \oplus g, P \rangle.$$

Minimizing over r and s gives $-F^*(-f)$ and $-G^*(-g)$. For each scalar entry, the convex conjugate of $p \mapsto C_{i,j} p + \varepsilon(p \log(p/(a_i b_j)) - p + a_i b_j)$ is $q \mapsto \varepsilon a_i b_j (e^{(q - C_{i,j})/\varepsilon} - 1)$. Minimizing over P with $q = f_i + g_j$ gives (8.37).

For the scaling form, fix $v > 0$, set $\widetilde{K} = K \,\mathrm{diag}(v)$ and $z = \widetilde{K}\mathbb{1}_m = Kv$. Updating the row scaling means looking for $P = \mathrm{diag}(u)\widetilde{K}$, so its row marginal is $r = P\mathbb{1}_m = u \odot z$. The row-wise chain rule gives

$$\mathrm{KL}(\mathrm{diag}(u)\widetilde{K}|\widetilde{K}) = \sum_i \big(r_i \log(r_i/z_i) - r_i + z_i \big) = \mathrm{KL}(r|z).$$

Thus exact optimization of this block is equivalent to minimizing $F(r) + \varepsilon \mathrm{KL}(r|z)$ over $r \geqslant 0$, hence $r = \mathrm{prox}^{\mathrm{KL}}_{F/\varepsilon}(z)$ and $u = r \oslash z$. The column update is identical with $w = K^\top u$. □

When $F = \iota_{\{a\}}$ and $G = \iota_{\{b\}}$, the first two dual terms are $\langle f, a \rangle + \langle g, b \rangle$, and one recovers the usual entropic dual (8.23). The classical Sinkhorn update is the special case in which the KL proximal maps return the prescribed marginals a and b.

The usefulness of this formulation is that many KL-proximal maps are explicit.

- **Hard marginal constraint.** If $F = \iota_{\{a\}}$, then $\mathrm{prox}^{\mathrm{KL}}_{F/\varepsilon}(z) = a$.

Algorithm 8.3: Generalized Sinkhorn for marginal penalties

Input: Reference weights a, b, cost C, convex marginal penalties F, G, regularization $\varepsilon > 0$, tolerance tol, iteration budget $L \geqslant 1$.
Output: Coupling P and optimized marginals $r = \mathrm{P}\mathbb{1}_m$, $s = \mathrm{P}^\top \mathbb{1}_n$.
Set $\mathrm{K}_{i,j} = a_i b_j e^{-\mathrm{C}_{i,j}/\varepsilon}$.
Initialize: $\mathrm{u} = \mathbb{1}_n$, $\mathrm{v} = \mathbb{1}_m$, $\eta = +\infty$, and $\ell = 0$.
While $\eta >$ tol and $\ell < L$ **do**:
  Set $\ell \leftarrow \ell + 1$.
  Store $\mathrm{u}_{\mathrm{old}} = \mathrm{u}$ and $\mathrm{v}_{\mathrm{old}} = \mathrm{v}$.
  Set $z = \mathrm{K}\mathrm{v}$.
  Compute $r = \mathrm{prox}^{\mathrm{KL}}_{\mathsf{F}/\varepsilon}(z)$.
  Set $\mathrm{u} = r \oslash z$.
  Set $w = \mathrm{K}^\top \mathrm{u}$.
  Compute $s = \mathrm{prox}^{\mathrm{KL}}_{\mathsf{G}/\varepsilon}(w)$.
  Set $\mathrm{v} = s \oslash w$.
  Set $\eta = \max\{\|\mathrm{u} - \mathrm{u}_{\mathrm{old}}\|_\infty, \|\mathrm{v} - \mathrm{v}_{\mathrm{old}}\|_\infty\}$.
Return $\mathrm{P} = \mathrm{diag}(\mathrm{u})\mathrm{K}\,\mathrm{diag}(\mathrm{v})$, $r = \mathrm{P}\mathbb{1}_m$, and $s = \mathrm{P}^\top \mathbb{1}_n$.

- **KL marginal relaxation.** If $\mathsf{F}(r) = \tau \mathrm{KL}(r|\mathrm{a})$ with $\tau > 0$, then, coordinatewise,
$$\mathrm{prox}^{\mathrm{KL}}_{\mathsf{F}/\varepsilon}(z) = z^{\varepsilon/(\tau+\varepsilon)} \odot \mathrm{a}^{\tau/(\tau+\varepsilon)},$$
and the scaling update becomes $\mathrm{u} \leftarrow (\mathrm{a} \oslash z)^{\tau/(\tau+\varepsilon)}$, the damped scaling of unbalanced Sinkhorn.
- **Pointwise bounds.** If $\mathsf{F} = \iota_{\{\ell \leqslant r \leqslant u\}}$, then the proximal map is the coordinatewise clipping $\mathrm{prox}^{\mathrm{KL}}_{\mathsf{F}/\varepsilon}(z)_i = \min\{u_i, \max\{\ell_i, z_i\}\}$.
- **Total-variation marginal relaxation.** If $\mathsf{F}(r) = \tau\|r - \mathrm{a}\|_1$ and $\lambda = \tau/\varepsilon$, then the proximal map is coordinatewise
$$\mathrm{prox}^{\mathrm{KL}}_{\mathsf{F}/\varepsilon}(z)_i = \begin{cases} z_i e^{\lambda}, & z_i < a_i e^{-\lambda}, \\ a_i, & a_i e^{-\lambda} \leqslant z_i \leqslant a_i e^{\lambda}, \\ z_i e^{-\lambda}, & z_i > a_i e^{\lambda}. \end{cases}$$
- **Fixed total mass.** If $\mathsf{F} = \iota_{\{\langle r, \mathbb{1}\rangle = m\}}$, then $\mathrm{prox}^{\mathrm{KL}}_{\mathsf{F}/\varepsilon}(z) = mz/\langle z, \mathbb{1}\rangle$.

These examples explain why generalized Sinkhorn algorithms remain practical: the expensive operation is still multiplication by K or $\mathrm{K}^\top$, while the model-specific part is a low-dimensional KL-proximal update on a marginal.

8.8 Heat Kernels and Hopf–Cole Transforms

The Gaussian kernel used by Sinkhorn is also the Euclidean heat kernel. This viewpoint clarifies when entropic OT admits fast grid and surface implementations, and it places soft minima and Hopf–Cole transforms in the same heat-kernel calculus.

Geodesics in heat. On $\mathbb{R}^d$, the heat kernel for $\partial_t u = \Delta u$ is
$$h_t(x, y) = (4\pi t)^{-d/2} \exp\left(-\frac{\|x - y\|^2}{4t}\right).$$
For the quadratic cost $c(x, y) = \|x - y\|^2$, the Gibbs kernel used by Sinkhorn satisfies
$$\mathrm{K}_\varepsilon(x, y) := e^{-\|x-y\|^2/\varepsilon} = (\pi\varepsilon)^{d/2} h_{\varepsilon/4}(x, y). \tag{8.40}$$
The scalar factor is absorbed by the Sinkhorn scalings and therefore does not change the optimal coupling. On a Riemannian manifold or surface M, let $L = -\Delta_M$ be the positive Laplace–Beltrami operator and replace the dense Gibbs kernel by the intrinsic heat operator
$$\mathsf{H}_\varepsilon := e^{-(\varepsilon/4)L},$$

whose integral kernel is $h_{\varepsilon/4}(x, y)$. For histograms on the same discretized domain, the scaling steps then read

$$u^{(\ell+1)} = a \oslash \big(\mathsf{H}_\varepsilon v^{(\ell)}\big), \qquad v^{(\ell+1)} = b \oslash \big(\mathsf{H}_\varepsilon^\top u^{(\ell+1)}\big), \tag{8.41}$$

where H_ε denotes the discretized integral operator, including quadrature weights, and $\mathsf{H}_\varepsilon^\top$ its discrete adjoint. In a symmetric mass-normalized discretization, the two operators coincide. Thus heat diffusion fits Sinkhorn at its only expensive operation: multiplication by the Gibbs kernel. Equivalently, it uses the effective cost

$$c_\varepsilon(x, y) := -\varepsilon \log h_{\varepsilon/4}(x, y).$$

Varadhan's formula gives

$$c_\varepsilon(x, y) \longrightarrow d_M(x, y)^2 \qquad (\varepsilon \downarrow 0),$$

so convolutional Sinkhorn recovers the squared geodesic ground cost at small temperature without first computing all pairwise geodesic distances [530, 495]. The same asymptotic relation is the analytical basis of geodesics-in-heat distance estimation [163].

The heat operator can itself be applied through shifted-Laplacian solves:

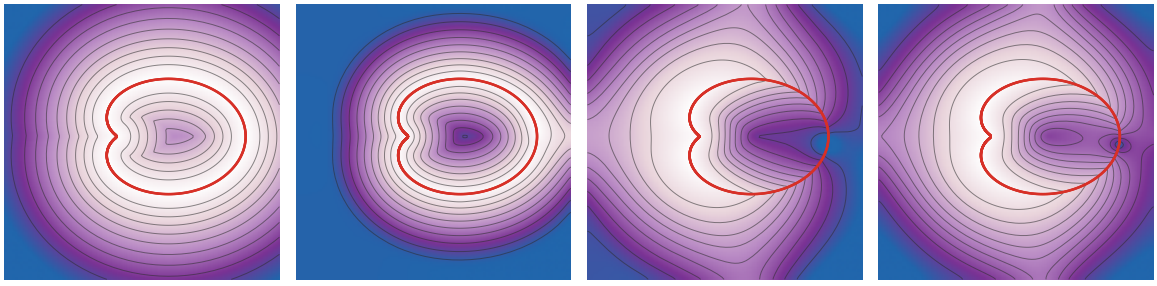
$$\mathsf{H}_\varepsilon = \lim_{q\to\infty} \left(\mathrm{Id} + \frac{\varepsilon}{4q} L\right)^{-q}.$$

On a triangulated surface or image grid, replace L by a sparse positive semidefinite Laplacian L_h and set

$$A_{\varepsilon,q} := \mathrm{Id} + \frac{\varepsilon}{4q} L_h.$$

For fixed (ε, q), one computes a sparse Cholesky factorization $A_{\varepsilon,q} = RR^\top$ once. Each application of H_ε is then approximated by q successive solves with $A_{\varepsilon,q}$, and each solve uses only the two triangular substitutions associated with R. Since (8.41) repeatedly applies the same kernel, this factorization is reused throughout all Sinkhorn iterations [495]. This avoids both a dense kernel matrix and an all-pairs geodesic-distance computation. A temperature-scaling strategy requires one factorization for each value of ε; on a finite-element mesh, the analogous shifted system is written with the mass and stiffness matrices. The small-temperature limit remains delicate because the diffusion length, of order $\sqrt{\varepsilon}$, must be resolved by the mesh: if ε vanishes faster than the squared grid spacing, metrication artifacts and inconsistent discretizations dominate the intended geodesic limit.

Figure 8.11 compares the exact distance to a non-convex source curve with heat-kernel and shifted-Laplacian approximations at several smoothing scales.



exact distance small ε medium ε larger ε

Figure 8.11: Geodesics-in-heat approximation of the distance to a dense red non-convex source curve on a two-dimensional grid. The first panel shows the exact Euclidean distance to the curve. The other panels diffuse a narrow band by the one-step approximation $(\mathrm{Id} + \varepsilon L_h/4)^{-1}$ with Neumann boundary conditions, then recover a distance by the normalized-gradient Poisson solve. Each panel is robustly normalized, and contour values are selected from quantiles of the displayed field. Increasing the Sinkhorn temperature ε suppresses unresolved grid-scale artifacts but progressively rounds the non-convex level-set geometry.

Soft Hopf–Lax and Hopf–Cole. The Hopf–Lax formula is the Hamilton–Jacobi incarnation of the hard c-transform. We use the normalized quadratic cost

$$c(x, y) = \frac{1}{2}\|x - y\|^2,$$

so that the Hopf–Lax operator applied to an initial datum h is precisely

$$(-h)^{\bar{c}}(x) = \inf_y \left\{ h(y) + \frac{1}{2}\|x-y\|^2 \right\},$$

where the sign only reflects the convention of Definition 5.8. This is the usual Hopf–Lax formula for the Hamiltonian $\|p\|^2/2$ [210, 539]; other quadratic scalings amount to multiplying the cost by a constant. In the present entropic setting, the parameter of interest is instead the temperature ε.

The soft version replaces the infimum by a log-sum-exp soft minimum. In the notation of Definition 8.21, and using Lebesgue measure on $\mathbb{R}^d$,

$$(-h)^{\bar{c},\varepsilon}(x) = -\varepsilon \log \int \exp\left(-\frac{h(y) + \|x-y\|^2/2}{\varepsilon} \right) \mathrm{d}y. \tag{8.42}$$

This is a soft c-transform of the function $-h$, and Laplace's principle gives $(-h)^{\bar{c},\varepsilon} \to (-h)^{\bar{c}}$ as $\varepsilon \to 0$ under the usual compactness assumptions on near-minimizers. The same formula is also a heat-kernel formula. If

$$G_\varepsilon(z) := (2\pi\varepsilon)^{-d/2} \exp\left(-\frac{\|z\|^2}{2\varepsilon} \right)$$

is the normalized Gaussian kernel with covariance $\varepsilon \mathrm{Id}$, then (8.42) is exactly

$$(-h)^{\bar{c},\varepsilon}(x) = -\varepsilon \log \left(G_\varepsilon * e^{-h/\varepsilon} \right)(x) - \frac{\varepsilon d}{2} \log(2\pi\varepsilon). \tag{8.43}$$

Thus a soft quadratic c-transform is a Gaussian convolution followed by a logarithm, up to an explicit additive constant independent of x. This is the bridge between soft-minimum operators, heat kernels and entropic transport potentials.

Proposition 8.31 (Soft quadratic c-transform and Legendre approximation). *Let $f : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ be such that the integrals below are finite, and introduce the quadratic shift*

$$\mathsf{S}f(y) := f(y) - \frac{1}{2}\|y\|^2.$$

For $\varepsilon > 0$, define the soft conjugate by applying the soft $\bar{c}$-transform to the shifted function:

$$f^{*,\varepsilon}(p) := \frac{1}{2}\|p\|^2 - \big(-\mathsf{S}f\big)^{\bar{c},\varepsilon}(p). \tag{8.44}$$

Then

$$f^{*,\varepsilon}(p) = \varepsilon \log \int_{\mathbb{R}^d} \exp\left(\frac{\langle p, y\rangle - f(y)}{\varepsilon} \right) \mathrm{d}y, \tag{8.45}$$

and equivalently

$$f^{*,\varepsilon}(p) = \frac{1}{2}\|p\|^2 + \varepsilon \log \left(G_\varepsilon * e^{-\mathsf{S}f/\varepsilon} \right)(p) + \frac{\varepsilon d}{2} \log(2\pi\varepsilon). \tag{8.46}$$

If, for instance, f is proper, lower semicontinuous and superlinear, then $f^{*,\varepsilon}(p) \to f^*(p)$ for every p as $\varepsilon \to 0$.

Proof. The hard identity follows by completing the square:

$$\inf_y \left\{ \mathsf{S}f(y) + \frac{1}{2}\|p-y\|^2 \right\} = \frac{1}{2}\|p\|^2 - \sup_y \{\langle p, y\rangle - f(y)\} = \frac{1}{2}\|p\|^2 - f^*(p).$$

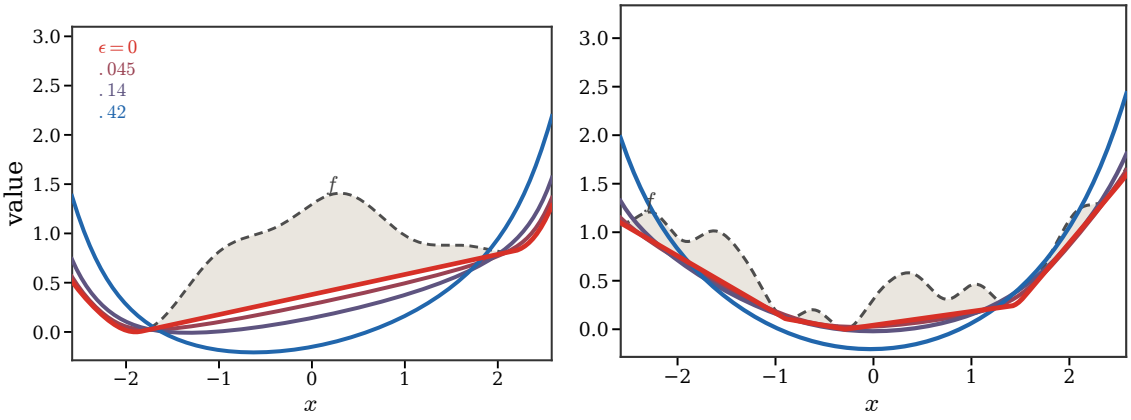
This explains the shift: it turns the Legendre–Fenchel transform into a quadratic $\bar{c}$-transform. Replacing the hard transform by its soft version gives (8.44). Expanding the square in (8.42) with the function $\mathsf{S}f$ gives

$$\exp\left(-\frac{\mathsf{S}f(y) + \|p-y\|^2/2}{\varepsilon} \right) = \exp\left(-\frac{\|p\|^2}{2\varepsilon} \right) \exp\left(\frac{\langle p, y\rangle - f(y)}{\varepsilon} \right),$$

which proves (8.45). Formula (8.46) is the Gaussian-convolution identity (8.43) applied to $\mathsf{S}f$. Under the stated superlinearity assumption, the functions $y \mapsto \langle p, y\rangle - f(y)$ have compact upper level sets, so Laplace's principle applied to (8.45) gives the convergence to $f^*(p)$. □

Remark 8.32 (Fast soft Legendre–Fenchel transforms). Proposition 8.31 gives a computational recipe for a smoothed Legendre–Fenchel transform. On a periodic, or sufficiently padded, grid, the term $G_\varepsilon * e^{-\mathsf{S}f/\varepsilon}$ in (8.46) is a Gaussian convolution. It can therefore be evaluated in $O(N \log N)$ operations for N grid samples using an FFT. This is not an exact hard discrete Legendre transform, but a regularized approximation whose zero-temperature limit recovers the hard transform by Laplace's principle. Exact discrete conjugation and lower-envelope algorithms exploit convex-analytic and computational-geometry structure instead; see Lucet's survey of computational convex analysis and the distance-transform algorithm of Felzenszwalb and Huttenlocher [357, 212].
The convolutional route is nevertheless delicate when ε is small. In fixed precision, the factors $e^{-f/\varepsilon}$ or $e^{-\mathsf{S}f/\varepsilon}$ can underflow or overflow, and the logarithm can amplify small relative convolution errors. Practical implementations therefore use shifts, log-domain evaluations, or stabilized FFT convolutions. In dimension d, one can also use the same separability of the Gaussian kernel as in grid Sinkhorn: direct one-dimensional passes give the tensor-product cost (8.6), namely $O(d\, N^{1+1/d})$, while FFT convolutions provide an additional acceleration when boundary conditions and conditioning permit it.

Figure 8.12 illustrates the biconjugation viewpoint directly. The red curve is the hard lower convex envelope f^{**}, while the blue curves are smoother finite-temperature biconjugates. The two examples show that the same mechanism convexifies both a simple double-well profile and a quadratic oscillatory landscape.



smooth double well quadratic oscillatory profile

Figure 8.12: Soft Legendre biconjugates as approximations of lower convex envelopes. In each panel, the dashed gray curve is a non-convex function f, the red curve is the hard biconjugate f^{**}, and the purple-to-blue curves show $(f^{*,\varepsilon})^{*,\varepsilon}$ for increasing ε. The pale gray area marks the part of f lying above its convex envelope, which is progressively filled in by the soft convexification.

The nonlinear PDEs are linearized by the Hopf–Cole transform. With the same temperature normalization, the change of variables $u_s = e^{-\varphi_s/\varepsilon}$ converts the viscous Hamilton–Jacobi equation

$$\partial_s \varphi_s + \frac{1}{2}\|\nabla \varphi_s\|^2 = \frac{\varepsilon}{2}\Delta \varphi_s,$$

into the linear heat equation $\partial_s u_s = (\varepsilon/2)\Delta u_s$. Conversely, starting from a positive heat solution u_s and setting $\varphi_s = -\varepsilon \log u_s$ gives a Hamilton–Jacobi solution, while $v_s = \nabla \varphi_s = -\varepsilon \nabla \log u_s$ solves the gradient viscous Burgers equation

$$\partial_s v_s + (v_s \cdot \nabla) v_s = \frac{\varepsilon}{2}\Delta v_s.$$

Thus nonlinear Hamilton–Jacobi/Burgers dynamics can be computed through a Laplacian-generated heat flow followed by a logarithmic change of variables. In one dimension this is the classical Cole–Hopf transform for viscous Burgers; in higher dimension this scalar reduction applies to irrotational velocity fields. Figure 8.13 starts from a Gaussian velocity bump, whose inviscid evolution would form a shock on its decreasing flank, and shows how the viscosity parameter $\varepsilon/2$ regularizes this steepening. Together with Figure 8.12, it gives the two complementary viewpoints behind soft-minimum formulas, Gaussian convolution and entropic transport potentials.

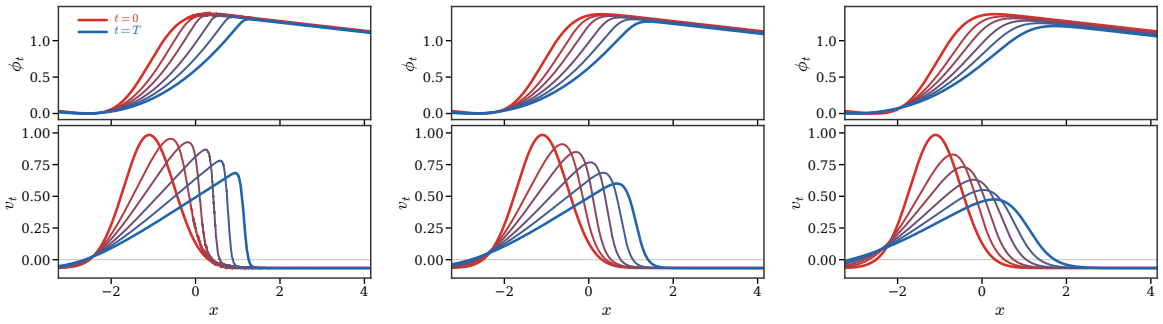


Figure 8.13: Hopf–Cole numerics for viscous Hamilton–Jacobi and Burgers dynamics in one dimension. Each panel fixes ε and evolves the same initial Gaussian-shaped Burgers velocity. The upper curves show the potential φ_t obtained from the heat solution by $\varphi_t = -\varepsilon \log u_t$, while the lower curves show the Burgers velocity $v_t = \partial_x \varphi_t = -\varepsilon\, \partial_x \log u_t$. Colors encode time from red to blue; small ε lets the Gaussian bump develop a sharp shock-like front, while larger ε gives stronger heat smoothing and viscous damping.

8.9 Other Convex Regularizers

KL regularization is the case that leads to multiplicative Sinkhorn scalings. Replacing the KL divergence by another density-ratio penalty keeps the same transport constraints but changes the scalar law linking the optimal density to the dual potentials.

φ-divergence regularization. The primal and dual problems retain the same transport constraints, but the exponential Gibbs law is replaced by the derivative of a general convex conjugate.

Let φ be an entropy function in the sense of Definition 7.15, and recall the φ-divergence $\mathcal{D}_\varphi$ from Definition 7.16. For $\varepsilon > 0$, define the φ-regularized transport value.

Definition 8.33 (φ-regularized optimal transport)**.**

$$\mathcal{L}^{\varepsilon}_{c,\varphi}(\alpha,\beta) \coloneqq \min_{\pi \in \mathcal{U}(\alpha,\beta)} \int_{\mathcal{X}\times\mathcal{Y}} c(x,y)\mathrm{d}\pi(x,y) + \varepsilon \mathcal{D}_\varphi(\pi|\alpha\otimes\beta). \tag{8.47}$$

Proposition 8.34 (Dual and density law for φ-regularized OT)**.** *Assume that the Fenchel–Rockafellar qualification condition holds; for example, let the spaces be compact, let c be continuous, and assume that the product coupling belongs to the continuity domain of the divergence term. Then*

$$\mathcal{L}^{\varepsilon}_{c,\varphi}(\alpha,\beta) = \sup_{f\in C(\mathcal{X}),\, g\in C(\mathcal{Y})} \int f\mathrm{d}\alpha + \int g\mathrm{d}\beta - \varepsilon \int \varphi^*\left(\frac{f(x)+g(y)-c(x,y)}{\varepsilon}\right)\mathrm{d}\alpha(x)\mathrm{d}\beta(y). \tag{8.48}$$

If an optimal plan is absolutely continuous, with density $r^\star = \frac{\mathrm{d}\pi^\star}{\mathrm{d}(\alpha\otimes\beta)}$, and $(f^\star, g^\star)$ are optimal potentials, then

$$\frac{f^\star(x)+g^\star(y)-c(x,y)}{\varepsilon} \in \partial\varphi(r^\star(x,y)) \qquad \alpha\otimes\beta\text{-a.e.} \tag{8.49}$$

In the smooth interior this reads

$$r^\star(x,y) = (\varphi')^{-1}\left(\frac{f^\star(x)+g^\star(y)-c(x,y)}{\varepsilon}\right).$$

Proof. Introduce dual variables (f, g) for the two marginal constraints. For fixed (f, g), the minimization over π gives

$$\int f\mathrm{d}\alpha + \int g\mathrm{d}\beta + \inf_{\pi\in\mathcal{M}_+(\mathcal{X}\times\mathcal{Y})}\left\{\int \big(c-(f\oplus g)\big)\mathrm{d}\pi + \varepsilon\mathcal{D}_\varphi(\pi|\alpha\otimes\beta)\right\}.$$

Using the Legendre formula (7.13) for the convex functional $\mathcal{D}_\varphi(\cdot|\alpha\otimes\beta)$, the infimum equals

$$-\varepsilon\int \varphi^*\left(\frac{f(x)+g(y)-c(x,y)}{\varepsilon}\right)\mathrm{d}\alpha(x)\mathrm{d}\beta(y),$$

which gives (8.48). Equality in the Fenchel inequality is equivalent to the subgradient inclusion

$$\frac{f^\star \oplus g^\star - c}{\varepsilon} \in \partial\varphi(r^\star),$$

and inversion of φ' gives the density law when φ is differentiable and the optimizer is in the interior of its domain. □

For the KL entropy $\varphi(r) = r\log r - r + 1$, one has $\varphi^*(s) = e^s - 1$, and (8.48) reduces exactly to the continuous Sinkhorn dual (8.23). Other choices replace the exponential density law by another scalar transfer function:

$$\begin{aligned}
&\varphi(r) = r\log r - r + 1 \quad &&\Rightarrow \quad r^\star = e^s,\\
&\varphi(r) = r - \log r - 1 \quad &&\Rightarrow \quad r^\star = (1-s)^{-1} \quad (s<1),\\
&\varphi(r) = \tfrac{1}{2}(r-1)^2 \quad &&\Rightarrow \quad r^\star = (1+s)_+,
\end{aligned} \qquad s := \frac{f^\star \oplus g^\star - c}{\varepsilon}.$$

Bregman-divergence regularization. The previous construction regularizes OT by a density-ratio divergence. This differs from using a Bregman divergence generated by a convex functional on the space of measures.

Definition 8.35 (Measure Bregman divergence)**.** If Φ is a differentiable convex functional on a convex class of nonnegative measures and ξ is a reference measure in its domain, the measure Bregman divergence generated by Φ is

$$B_\Phi(\pi|\xi) := \Phi(\pi) - \Phi(\xi) - \int \delta\Phi(\xi)\mathrm{d}(\pi - \xi), \tag{8.50}$$

where $\delta\Phi(\xi)$ is the first variation and the formula is understood whenever the right-hand side is well-defined.

In finite dimension this reduces to Definition 9.1. The KL divergence is special because it is simultaneously a density-ratio divergence and a Bregman divergence.

This distinction leads to a second regularized transport problem with the same marginal constraints.

Definition 8.36 (Bregman-regularized optimal transport)**.** Fix $\xi = \alpha \otimes \beta$. For $\varepsilon > 0$, define

$$\mathcal{L}^\varepsilon_{c,\Phi}(\alpha,\beta) := \inf_{\pi\in\mathcal{U}(\alpha,\beta)} \left\{ \int c\,\mathrm{d}\pi + \varepsilon B_\Phi(\pi|\xi) \right\}.$$

Proposition 8.37 (Dual and density law for Bregman-regularized OT)**.** *Fix the marginals α,β and set $\xi := \alpha\otimes\beta$. Let Φ be a convex Gâteaux-differentiable functional on nonnegative measures on $\mathcal{X}\times\mathcal{Y}$. Its convex conjugate is, for a continuous test function u,*

$$\Phi^*(u) := \sup_{\pi\geqslant 0} \left\{ \int u\,\mathrm{d}\pi - \Phi(\pi) \right\}.$$

Assume that Fenchel duality is exact for this constrained problem, as happens in finite-dimensional discretizations and, more generally, under standard compactness and lower-semicontinuity hypotheses. Then

$$\mathcal{L}^\varepsilon_{c,\Phi}(\alpha,\beta) = \sup_{f,g} \int f\mathrm{d}\alpha + \int g\mathrm{d}\beta - \varepsilon\left[\Phi^*\left(\delta\Phi(\xi) + \frac{f\oplus g - c}{\varepsilon}\right) - \Phi^*(\delta\Phi(\xi))\right]. \tag{8.51}$$

If $(f^\star, g^\star)$ and $\pi^\star$ are optimal and the solution is interior, then

$$\delta\Phi(\pi^\star) = \delta\Phi(\xi) + \frac{f^\star\oplus g^\star - c}{\varepsilon}. \tag{8.52}$$

Proof. Using (8.50), the Bregman-regularized objective can be written, up to a constant independent of π, as

$$\varepsilon\Phi(\pi) + \int \big(c - \varepsilon\delta\Phi(\xi)\big)\mathrm{d}\pi - \varepsilon\Phi(\xi) + \varepsilon\int \delta\Phi(\xi)\mathrm{d}\xi.$$

Introduce dual potentials (f, g) for the two marginal constraints. The inner minimization over nonnegative measures π gives

$$\inf_{\pi \geqslant 0} \left\{ \varepsilon \Phi(\pi) + \int \big(c - (f \oplus g) - \varepsilon \delta\Phi(\xi)\big) \mathrm{d}\pi \right\} = -\varepsilon \Phi^* \left(\delta\Phi(\xi) + \frac{f \oplus g - c}{\varepsilon} \right).$$

Fenchel equality at ξ gives

$$-\Phi(\xi) + \int \delta\Phi(\xi) \mathrm{d}\xi = \Phi^*(\delta\Phi(\xi)),$$

which yields (8.51). Equality in Fenchel's inequality gives the primal–dual relation (8.52). □

The primal–dual relations (8.52) and (8.49) make the distinction precise. Bregman regularization translates the reference measure in the functional dual coordinate $\delta\Phi$, whereas φ-divergence regularization applies the scalar subgradient $\partial\varphi$ pointwise to the density relative to $\alpha \otimes \beta$. For KL these two laws coincide: the logarithmic dual coordinate converts additive potential shifts into multiplicative density scalings.

When Φ is separable with respect to a fixed dominating measure ξ_0, say $\Phi(\pi) = \int h(\mathrm{d}\pi/\mathrm{d}\xi_0)\mathrm{d}\xi_0$ with $\xi \ll \xi_0$, the Bregman optimality condition becomes

$$h'\left(\frac{\mathrm{d}\pi^\star}{\mathrm{d}\xi_0}\right) = h'\left(\frac{\mathrm{d}\xi}{\mathrm{d}\xi_0}\right) + \frac{f^\star \oplus g^\star - c}{\varepsilon}.$$

This is an additive update in entropy coordinates. The density-ratio formulation instead uses the scalar law associated with φ relative to $\alpha \otimes \beta$. These two laws coincide for the KL entropy, where $h'(r) = \log r$ turns additive dual shifts into multiplicative scalings. The next proposition shows that, under natural smoothness assumptions, this is the only overlap.

Proposition 8.38 (KL is the common Bregman and φ case)**.** *Let ω be a finite reference measure on $\mathcal{X}$ with a measurable set A satisfying $0 < \omega(A) < \omega(\mathcal{X})$. Work on probability measures $\alpha = p\omega$ and $\beta = q\omega$ whose densities are bounded above and below away from 0. Assume that Φ is twice Gâteaux differentiable along bounded zero-mass density perturbations, and that $\varphi \in C^2(0, +\infty)$ is convex with $\varphi(1) = 0$. If*

$$B_\Phi(\alpha|\beta) = \mathcal{D}_\varphi(\alpha|\beta)$$

for all such α, β, then there exist $\kappa \geqslant 0$ and $a \in \mathbb{R}$ such that

$$\varphi(t) = \kappa\, t \log t + a(t-1).$$

Hence the common divergence is $\kappa\,\mathrm{KL}(\alpha|\beta)$, and $\Phi(p\omega)$ differs from $\kappa \int p \log p \,\mathrm{d}\omega$ by an affine functional on the positive probability simplex.

Proof. Fix $\beta = q\omega$ and perturb it by $\alpha_t = (q + th)\omega$, where h is bounded, $\int h \mathrm{d}\omega = 0$, and t is small enough that $q + th > 0$. Differentiating twice at $t = 0$ gives

$$D^2\Phi(q)[h, h] = \varphi''(1) \int \frac{h^2}{q} \mathrm{d}\omega.$$

Indeed the left-hand side is the second variation of the Bregman error, while

$$\mathcal{D}_\varphi(\alpha_t|\beta) = \int q\, \varphi\,(1 + th/q)\, \mathrm{d}\omega$$

has second derivative $\varphi''(1) \int h^2/q \,\mathrm{d}\omega$. Setting $\kappa = \varphi''(1) \geqslant 0$, the functional $\Psi(p\omega) = \Phi(p\omega) - \kappa \int p \log p \mathrm{d}\omega$ has zero second variation along every zero-mass line segment in the positive simplex. It is therefore affine there, and $B_\Psi = 0$. Thus $B_\Phi = \kappa\,\mathrm{KL}$.

It remains to identify the generator. Since $\mathcal{D}_\varphi = \kappa\,\mathrm{KL}$, the function $r(t) = \varphi(t) - \kappa t \log t$ generates the zero φ-divergence. Testing on densities for which the ratio p/q takes two values $x < 1 < y$, with weights chosen so that the mean ratio is 1, gives $(y-1)r(x) + (1-x)r(y) = 0$. Hence $r(t)/(t-1)$ is constant on $(0, +\infty) \setminus \{1\}$, so $r(t) = a(t-1)$. This proves the claim. □

Thus the two generalizations lead to different duals and different algorithms. Bregman regularization by $B_\Phi(\pi|\xi)$ keeps the projection geometry of Section 9.1: linear costs tilt the reference in dual coordinates and alternating marginal updates are Bregman projections. A density-ratio penalty $\mathcal{D}_\varphi(\pi|\alpha\otimes\beta)$ instead gives the Fenchel dual (8.48) and, for interior solutions, the pointwise law $r^\star = (\varphi')^{-1}((f\oplus g-c)/\varepsilon)$. Proposition 8.37 makes the distinction explicit at the dual level. The Bregman dual contains the global conjugate Φ^* and the product reference $\alpha\otimes\beta$, whereas the φ-dual integrates a scalar conjugate against the moving product measure $\alpha\otimes\beta$. Only for KL do these two viewpoints coincide and reduce to multiplicative Sinkhorn scalings.

Generalized soft c-transforms and alternate dual maximization method. The two dual formulations above lead to the same algorithmic principle: maximize exactly over one potential while the other is fixed, then exchange their roles. For the φ-divergence dual (8.48), separability with respect to $\alpha\otimes\beta$ makes each block update pointwise. The two φ-soft c-transforms are

$$\begin{aligned} g^{\bar c,\varepsilon,\varphi}(x) &\in \operatorname*{argmin}_{u\in\mathbb{R}} \left\{ \varepsilon \int_{\mathcal{Y}} \varphi^*\left(\frac{u+g(y)-c(x,y)}{\varepsilon}\right) \mathrm{d}\beta(y) - u \right\}, \\ f^{c,\varepsilon,\varphi}(y) &\in \operatorname*{argmin}_{v\in\mathbb{R}} \left\{ \varepsilon \int_{\mathcal{X}} \varphi^*\left(\frac{f(x)+v-c(x,y)}{\varepsilon}\right) \mathrm{d}\alpha(x) - v \right\}. \end{aligned} \tag{8.53}$$

When φ^* is differentiable, the first scalar minimizer is characterized by

$$\int_{\mathcal{Y}} (\varphi^*)'\left(\frac{u+g(y)-c(x,y)}{\varepsilon}\right) \mathrm{d}\beta(y) = 1,$$

and the second satisfies the symmetric equation. Thus each update normalizes one conditional density. For Burg or quadratic penalties, it remains a one-dimensional monotone solve, although it is no longer a log-sum-exp.

The Bregman dual (8.51) has analogous block transforms, but the global conjugate Φ^* prevents a pointwise definition in general. With $\xi = \alpha\otimes\beta$, define

$$\begin{aligned} g^{\bar c,\varepsilon,\Phi} &\in \operatorname*{argmin}_{u\in C(\mathcal{X})} \left\{ \varepsilon\Phi^*\left(\delta\Phi(\xi) + \frac{u\oplus g - c}{\varepsilon}\right) - \int u\,\mathrm{d}\alpha \right\}, \\ f^{c,\varepsilon,\Phi} &\in \operatorname*{argmin}_{v\in C(\mathcal{Y})} \left\{ \varepsilon\Phi^*\left(\delta\Phi(\xi) + \frac{f\oplus v - c}{\varepsilon}\right) - \int v\,\mathrm{d}\beta \right\}. \end{aligned} \tag{8.54}$$

These are exact function-space block minimizers; when Φ is separable, they reduce after disintegration to independent scalar problems of the same type as (8.53).

Alternate dual maximization therefore reads

$$\begin{aligned} f^{(\ell+1)} &= \left(g^{(\ell)}\right)^{\bar c,\varepsilon,\varphi}, \quad g^{(\ell+1)} = \left(f^{(\ell+1)}\right)^{c,\varepsilon,\varphi}, \\ f^{(\ell+1)} &= \left(g^{(\ell)}\right)^{\bar c,\varepsilon,\Phi}, \quad g^{(\ell+1)} = \left(f^{(\ell+1)}\right)^{c,\varepsilon,\Phi}, \end{aligned} \tag{8.55}$$

for the φ and Bregman models, respectively. For KL, which admits both descriptions, the two lines coincide with the usual soft c-transform iteration (8.27), hence recover Sinkhorn's algorithm.

Quadratic regularizers replace the exponential law by positive-part thresholding and can therefore produce sparse plans, as advocated by Blondel, Seguy, and Rolet [65]. In the discrete setting, the quadratic choices $\varphi(r) = \frac{1}{2}(r-1)^2$ and $\Phi(\mathrm{P}) = \frac{1}{2}\|\mathrm{P}\|_{\mathrm{F}}^2$ give, respectively,

$$\mathrm{P}^\star_{i,j} = a_i b_j \left(1 + \frac{f_i^\star + g_j^\star - \mathrm{C}_{i,j}}{\varepsilon}\right)_+, \qquad \mathrm{P}^\star_{i,j} = \left(a_i b_j + \frac{f_i^\star + g_j^\star - \mathrm{C}_{i,j}}{\varepsilon}\right)_+. \tag{8.56}$$

For example, the first row transform finds the unique threshold u satisfying

$$\sum_j b_j \left(1 + \frac{u + g_j - \mathrm{C}_{i,j}}{\varepsilon}\right)_+ = 1.$$

This left-hand side is continuous, nondecreasing, and piecewise affine. Sorting its breakpoints computes the update in $O(m\log m)$ operations for a row with m entries; equivalently, it is a weighted simplex

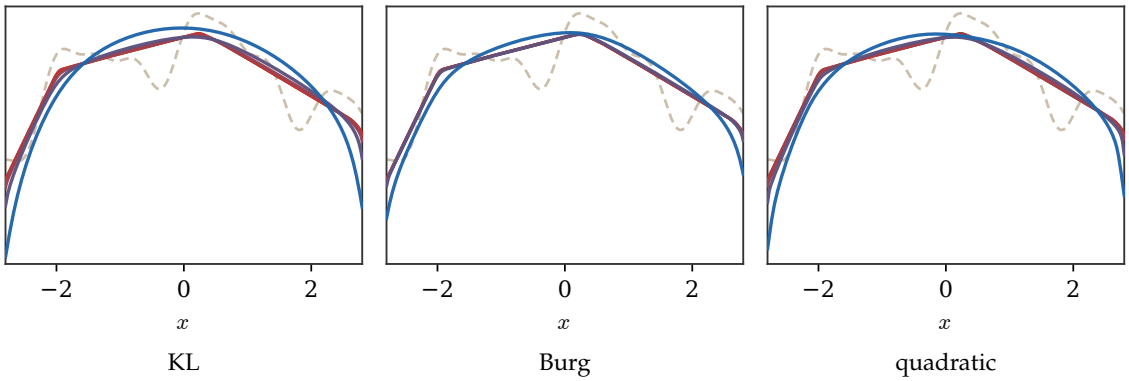


Figure 8.14: Generalized soft double transforms for $c(x, y) = -xy$. Each panel starts from the same dashed non-concave potential f. The dark curve is the hard double c-transform, i.e. the concave envelope associated with the zero-temperature limit, after centering by an additive constant. Colored curves show the centered double transform $(f^{c,\varepsilon,\varphi})^{\bar{c},\varepsilon,\varphi}$ from a near-zero red regularization to larger blue regularization. The numerical ε scale is chosen separately for each generator because KL, Burg, and quadratic density-ratio penalties have different stiffness; in each panel the red curve matches the hard concave envelope. Changing φ changes the scalar soft-normalization law and hence the way the dual ascent step smooths the envelope.

projection. The unweighted quadratic Bregman update is an ordinary Euclidean simplex projection, computable in the same worst-case complexity and in expected linear time by selection methods [65].

Figure 8.14 shows how the φ-transform changes the smoothing of the dual envelope.

Figure 8.15 shows the corresponding change in the concentration and support of the primal coupling; its right panel is the sparse quadratic plan computed by the alternating threshold transforms above.

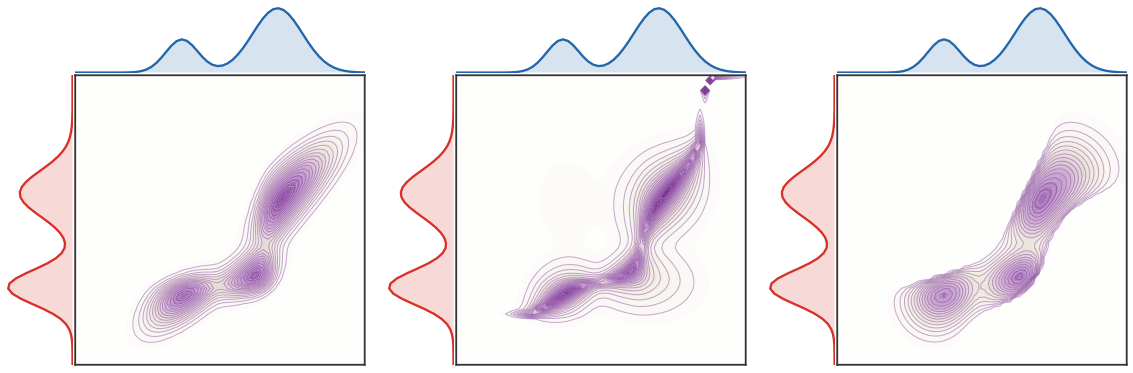


Figure 8.15: Density-ratio regularizers and coupling support. The red and blue side curves are fixed one-dimensional Gaussian-mixture marginals. Dense violet level sets display the square root of the coupling density on a common scale, using the smaller regularization strength $\varepsilon = .06$. KL regularization gives the usual diffuse positive plan, the Burg barrier keeps a positive but differently tailed support, and alternate dual maximization with the quadratic soft c-transform produces the sparse rightmost plan through its positive-part law.

8.10 Sinkhorn Divergences

Sinkhorn divergences remove the entropic self-bias while retaining smoothness. Under the kernel assumptions stated below, they interpolate between OT geometry and a Hilbertian energy distance, which explains their useful statistical behavior [213].

Entropic bias. A basic defect of the raw entropic value is its self-interaction bias. Even when $c(x, x) = 0$, one generally has $\mathcal{L}_c^\varepsilon(\alpha, \alpha) > 0$ for a non-Dirac measure. For a non-atomic α, the diagonal coupling is singular with respect to $\alpha \otimes \alpha$ and has infinite relative entropy; for a discrete non-Dirac measure, it has finite but positive relative entropy. Consequently, a minimizer

$$\beta_\varepsilon \in \operatorname*{argmin}_\beta \mathcal{L}_c^\varepsilon(\alpha, \beta)$$

need not equal α for $\varepsilon > 0$. The following proposition identifies the high-temperature mechanism behind this bias.

Proposition 8.39 (Large-temperature entropic bias)**.** *Let $\mathcal{X}, \mathcal{Y}$ be compact and assume that c is continuous. Then $\mathcal{L}_c^\varepsilon(\alpha,\beta) \to \iint c(x,y)\mathrm{d}\alpha(x)\mathrm{d}\beta(y)$ as $\varepsilon \to +\infty$.*

Proof. Let $(f_\varepsilon, g_\varepsilon)$ be optimal dual potentials, normalized by $\int g_\varepsilon \mathrm{d}\beta = 0$. The soft c-transform equation gives

$$f_\varepsilon(x) = -\varepsilon \log \int \exp\left(\frac{g_\varepsilon(y) - c(x,y)}{\varepsilon}\right) \mathrm{d}\beta(y).$$

For bounded c, the oscillations of normalized entropic potentials are bounded uniformly in ε by the oscillation of c. Hence the log-sum-exp expansion is uniform:

$$f_\varepsilon(x) = -\int (g_\varepsilon(y) - c(x,y))\mathrm{d}\beta(y) + O(\varepsilon^{-1}) = \int c(x,y)\mathrm{d}\beta(y) + O(\varepsilon^{-1}).$$

At optimality the exponential penalty in the dual integrates to zero, so

$$\mathcal{L}_c^\varepsilon(\alpha,\beta) = \int f_\varepsilon \mathrm{d}\alpha + \int g_\varepsilon \mathrm{d}\beta = \iint c(x,y)\mathrm{d}\alpha(x)\mathrm{d}\beta(y) + O(\varepsilon^{-1}).$$

This proves the limit. □

Thus, at large ε, the raw entropic cost approaches the bilinear cross-energy $(\alpha,\beta) \mapsto \iint c\, \mathrm{d}\alpha\mathrm{d}\beta$, not a distance. The following example makes the resulting collapse mechanism explicit and then specializes it to quadratic costs.

Example 8.40 (Large temperature collapse). Suppose that minimizers β_ε are tight and that the large-temperature convergence is uniform enough to pass to their cluster points. The limiting functional is linear in the second argument:

$$\beta \mapsto \int V_\alpha(y)\mathrm{d}\beta(y), \qquad V_\alpha(y) := \int c(x,y)\mathrm{d}\alpha(x).$$

Thus every cluster point is supported on $\operatorname{argmin} V_\alpha$. When this set is the singleton $\{y^\star(\alpha)\}$,

$$\beta_\varepsilon \rightharpoonup \delta_{y^\star(\alpha)}, \qquad y^\star(\alpha) = \operatorname*{argmin}_y V_\alpha(y).$$

For the quadratic cost $c(x,y) = \|x-y\|^2$ on $\mathbb{R}^d$, assuming α has finite second moment, one has $V_\alpha(y) = \|y - \int x\mathrm{d}\alpha(x)\|^2 + \text{const}$, so the collapse is toward the Dirac mass at the mean of α.

Sinkhorn divergences. The raw entropic OT value has a large-temperature attraction toward the product coupling. The standard debiasing of [213] subtracts the two self-interaction energies, in the same spirit as passing from a positive kernel value to the associated squared distance.

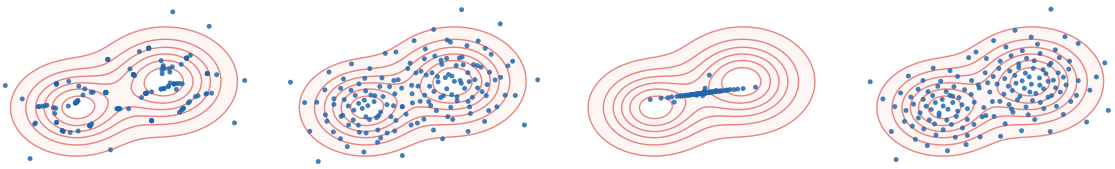
Definition 8.41 (Sinkhorn divergence)**.** Let c be a symmetric cost on a common state space. For $\varepsilon > 0$, the debiased Sinkhorn divergence associated with the entropic OT value $\mathcal{L}_c^\varepsilon$ is

$$\bar{\mathcal{L}}_c^\varepsilon(\alpha,\beta) := \mathcal{L}_c^\varepsilon(\alpha,\beta) - \frac{1}{2}\mathcal{L}_c^\varepsilon(\alpha,\alpha) - \frac{1}{2}\mathcal{L}_c^\varepsilon(\beta,\beta). \tag{8.57}$$

Although this formula is a debiasing by self-costs, its non-negativity is not automatic from the definition; Proposition 8.45 proves it below by a kernel Cauchy–Schwarz argument.

Figure 8.16 demonstrates the role of the two self-cost corrections by optimizing a finite point cloud against a fixed target with and without debiasing.

We first record a fundamental lemma: at an optimal dual pair, the exponential regularization term integrates to zero, so the entropic cost can be read directly from the potentials. The same cancellation holds along Sinkhorn iterations after each exact block update.



$\mathcal{L}_c^\varepsilon$, small ε $\bar{\mathcal{L}}_c^\varepsilon$, small ε $\mathcal{L}_c^\varepsilon$, large ε $\bar{\mathcal{L}}_c^\varepsilon$, large ε

Figure 8.16: Visualization of the debiasing effect by point optimization. The red level sets show a fixed two-Gaussian target density β, while the blue atoms are an optimized empirical measure α_n initialized in the same way in all panels. For small ε, the entropic cost $\mathcal{L}_c^\varepsilon$ and the debiased Sinkhorn divergence $\bar{\mathcal{L}}_c^\varepsilon$ both keep the two overlapping modes. For large ε, minimizing $\mathcal{L}_c^\varepsilon$ collapses the atoms toward the barycenter predicted by the large-temperature bias, whereas the self-cost subtraction in (8.57) keeps a bimodal cloud.

Lemma 8.42 (Entropic dual cost at optimum)**.** *Let $(f_{\alpha,\beta}, g_{\alpha,\beta})$ be optimal dual potentials, normalized arbitrarily. Then*

$$\mathcal{L}_c^\varepsilon(\alpha,\beta) = \langle f_{\alpha,\beta},\, \alpha\rangle + \langle g_{\alpha,\beta},\, \beta\rangle. \tag{8.58}$$

Proof. We first notice that at optimality, the relation

$$f_{\alpha,\beta} = -\varepsilon \log \int_{\mathcal{Y}} e^{\frac{g_{\alpha,\beta}(y)-c(x,y)}{\varepsilon}} \mathrm{d}\beta(y)$$

after taking the exponential, equivalently reads

$$1 = \int_{\mathcal{Y}} e^{\frac{f_{\alpha,\beta}(x)+g_{\alpha,\beta}(y)-c(x,y)}{\varepsilon}} \mathrm{d}\beta(y) \quad \Longrightarrow \quad \int_{\mathcal{X}\times\mathcal{Y}} \left(e^{\frac{f_{\alpha,\beta}\oplus g_{\alpha,\beta}-c}{\varepsilon}} - 1 \right) \mathrm{d}(\alpha\otimes\beta) = 0.$$

Substituting this identity in (8.22) gives the result. □

The next proposition records the two limiting regimes of this debiased quantity.

Proposition 8.43 (Asymptotics of Sinkhorn divergences)**.** *Let $\mathcal{X}$ be a compact metric space, let $\alpha,\beta \in \mathcal{P}(\mathcal{X})$, and assume that $c : \mathcal{X}\times\mathcal{X} \to \mathbb{R}_+$ is continuous and symmetric, with $c(x,x) = 0$. Then $\bar{\mathcal{L}}_c^\varepsilon(\alpha,\beta) \to \mathcal{L}_c(\alpha,\beta)$ when $\varepsilon \to 0$ and*

$$\bar{\mathcal{L}}_c^\varepsilon(\alpha,\beta) \to \frac{1}{2}\int -c\,\mathrm{d}(\alpha-\beta)\otimes\mathrm{d}(\alpha-\beta) \quad \textit{when} \quad \varepsilon \to +\infty.$$

Proof. The discrete convergence result above gives the finite-dimensional analogue; we now pass to continuous measures. **Case $\varepsilon \to 0$.** The first limit is the Γ-convergence of entropic OT to the Kantorovich problem [336, 108]. Non-negativity of the relative entropy and continuity of c give the liminf inequality. For the matching limsup, one approximates an optimal coupling by finite-entropy couplings and chooses the approximation scale so that its entropy multiplied by ε tends to zero. Since $c(x,x) = 0$, the two self-costs in the debiased expression converge to zero, while the cross term converges to $\mathcal{L}_c(\alpha,\beta)$.

Case $\varepsilon \to +\infty$. We denote by $(f_\varepsilon, g_\varepsilon)$ optimal dual potentials. After normalizing them and using boundedness of c, their oscillations stay uniformly bounded, so the following expansion is uniform. The optimality condition on f_ε (equivalently the Sinkhorn fixed point on f_ε) reads

$$\begin{aligned} f_\varepsilon &= -\varepsilon \log \int \exp\left(\frac{g_\varepsilon(y)-c(\cdot,y)}{\varepsilon}\right)\mathrm{d}\beta(y) = -\varepsilon\log\int\left(1+\frac{g_\varepsilon(y)-c(\cdot,y)}{\varepsilon}+o(1/\varepsilon)\right)\mathrm{d}\beta(y) \\ &= -\varepsilon\log\left(1+\frac{1}{\varepsilon}\int (g_\varepsilon(y)-c(\cdot,y))\mathrm{d}\beta(y)+o(1/\varepsilon)\right) = -\int g_\varepsilon \mathrm{d}\beta + \int c(\cdot,y)\mathrm{d}\beta(y) + o(1). \end{aligned}$$

Plugging this relation in the dual expression (8.58)

$$\mathcal{L}_c^\varepsilon(\alpha,\beta) = \int f_\varepsilon \mathrm{d}\alpha + \int g_\varepsilon \mathrm{d}\beta = \iint c(x,y)\mathrm{d}\alpha(x)\mathrm{d}\beta(y) + o(1).$$

Applying this expansion to (α,β), (α,α) and (β,β) gives

$$\bar{\mathcal{L}}_c^\varepsilon(\alpha,\beta) \to \int c\,\mathrm{d}\alpha\otimes\mathrm{d}\beta - \frac{1}{2}\int c\,\mathrm{d}\alpha\otimes\mathrm{d}\alpha - \frac{1}{2}\int c\,\mathrm{d}\beta\otimes\mathrm{d}\beta = -\frac{1}{2}\int c\,\mathrm{d}(\alpha-\beta)\otimes\mathrm{d}(\alpha-\beta).$$

□

> **Remark 8.44 (Large-temperature Hilbertian limit).** If c is conditionally negative definite, equivalently if $-c$ is conditionally positive definite, the large-temperature limit in Proposition 8.43 is a squared Hilbertian seminorm on zero-mass signed measures. A typical example is $c(x, y) = \|x - y\|^p$ for $0 < p < 2$, which yields the energy distance. This kernel norm is the dual of a homogeneous Sobolev norm.

We now show that this debiased Sinkhorn divergence is positive.

Proposition 8.45 (Positivity of Sinkhorn divergences)**.** *Assume that the entropic dual admits optimal potentials and that the symmetric kernel $k_\varepsilon(x, y) = e^{-c(x,y)/\varepsilon}$ is positive semidefinite in the sense of Definition 7.7. Then $\bar{\mathcal{L}}_c^\varepsilon(\alpha, \beta) \geqslant 0$. If, in addition, the common state space is compact, c is continuous, and k_ε is universal in the sense of Definition 7.12, then*

$$\bar{\mathcal{L}}_c^\varepsilon(\alpha, \beta) = 0 \quad \Longleftrightarrow \quad \alpha = \beta,$$

and $\bar{\mathcal{L}}_c^\varepsilon(\alpha_n, \alpha) \to 0$ is equivalent to $\alpha_n \rightharpoonup \alpha$.

Proof. In the following, we denote by $(f_{\alpha,\beta}, g_{\alpha,\beta})$ optimal dual potentials for the dual Schrödinger problem between α and β. We denote by $f_{\alpha,\alpha} = g_{\alpha,\alpha}$ (one can assume they are equal by symmetry) the solution for the problem between α and itself. Using the suboptimal function $(f_{\alpha,\alpha}, g_{\beta,\beta})$ in the dual maximization problem, and using relation (8.58) for the simplified expression of the dual cost, one obtains

$$\mathcal{L}_c^\varepsilon(\alpha, \beta) \geqslant \langle f_{\alpha,\alpha},\, \alpha\rangle + \langle g_{\beta,\beta},\, \beta\rangle - \varepsilon\langle e^{\frac{f_{\alpha,\alpha}\oplus g_{\beta,\beta}-c}{\varepsilon}} - 1,\, \alpha\otimes\beta\rangle$$

Moreover $\langle f_{\alpha,\alpha},\, \alpha\rangle = \frac{1}{2}\mathcal{L}_c^\varepsilon(\alpha, \alpha)$, and similarly for β, so the previous inequality equivalently reads

$$\frac{1}{\varepsilon}\bar{\mathcal{L}}_c^\varepsilon(\alpha, \beta) \geqslant 1 - \langle e^{\frac{f_{\alpha,\alpha}\oplus g_{\beta,\beta}-c}{\varepsilon}},\, \alpha\otimes\beta\rangle = 1 - \langle\tilde{\alpha},\, \tilde{\beta}\rangle_{k_\varepsilon}$$

where $\tilde{\alpha} = e^{f_{\alpha,\alpha}/\varepsilon}\alpha$, $\tilde{\beta} = e^{f_{\beta,\beta}/\varepsilon}\beta$ and we introduced the positive-semidefinite kernel pairing $\langle\tilde{\alpha},\, \tilde{\beta}\rangle_{k_\varepsilon} := \int k_\varepsilon(x, y)\mathrm{d}\tilde{\alpha}(x)\mathrm{d}\tilde{\beta}(y)$. The self Sinkhorn fixed point equation, once exponentiated, reads pointwise

$$e^{f_{\alpha,\alpha}(x)/\varepsilon}\int k_\varepsilon(x, y)\mathrm{d}\tilde{\alpha}(y) = 1 \qquad \text{for } \alpha\text{-a.e. } x,$$

and hence

$$\|\tilde{\alpha}\|_{k_\varepsilon}^2 = \langle\tilde{\alpha},\, \tilde{\alpha}\rangle_{k_\varepsilon} = \int e^{f_{\alpha,\alpha}(x)/\varepsilon}\left(\int k_\varepsilon(x, y)\mathrm{d}\tilde{\alpha}(y)\right)\mathrm{d}\alpha(x) = 1$$

and similarly $\|\tilde{\beta}\|_{k_\varepsilon}^2 = 1$. Therefore, by Cauchy–Schwarz, one has $1 - \langle\tilde{\alpha},\, \tilde{\beta}\rangle_{k_\varepsilon} \geqslant 0$.

Suppose now that k_ε is universal and $\bar{\mathcal{L}}_c^\varepsilon(\alpha, \beta) = 0$. Both inequalities above are then equalities. Equality in Cauchy–Schwarz, together with the unit norms, implies that $\tilde{\alpha}$ and $\tilde{\beta}$ have the same kernel mean embedding; universality gives $\tilde{\alpha} = \tilde{\beta} := \tilde{\mu}$. The two self-consistency equations yield

$$\mathrm{d}\alpha = (k_\varepsilon\tilde{\mu})\,\mathrm{d}\tilde{\mu} = \mathrm{d}\beta, \qquad (k_\varepsilon\tilde{\mu})(x) := \int k_\varepsilon(x, y)\mathrm{d}\tilde{\mu}(y),$$

so $\alpha = \beta$. The converse follows directly from the definition. Finally, on a compact space, continuity of the entropic value under weak convergence and compactness of $\mathcal{P}(\mathcal{X})$ turn this definiteness into the stated equivalence of convergences; see [213]. □

Positive semidefiniteness alone therefore guarantees only non-negativity and may leave a nontrivial null space.

8.11 Complex ε

The Sinkhorn temperature is usually a positive real number: positivity makes the Gibbs kernel positive, gives the entropy a convex meaning, and underlies the Hilbert-metric and monotone convergence arguments of Chapter 9. Once the equations are written as exponential fixed-point equations, however, ε can also be regarded locally as a complex variable. This does not produce a positive coupling or a contraction theorem; it produces a holomorphic branch of the same scaling equations near any positive real temperature.

Measure fixed point. Let $\alpha \in \mathcal{P}(\mathcal{X})$ and $\beta \in \mathcal{P}(\mathcal{Y})$. For $\varepsilon \in \mathbb{C} \setminus \{0\}$, reuse the Gibbs kernel and integral operators of (8.14) whenever the defining integrals exist. The density factorization (8.15), the marginal equations that follow it, and the scaling updates (8.16) then remain valid verbatim for complex-valued u and v, provided that the divisions are well defined. For real $\varepsilon > 0$ these are the usual Sinkhorn equations. For complex ε they are only local analytic identities: they assert neither positivity of π_ε nor convergence of the alternating iteration.

Discrete histograms. For discrete histograms, evaluate the Gibbs kernel entrywise at (x_i, y_j). The scaling factorization (8.2) and Sinkhorn updates (8.5) are then used verbatim over $\mathbb{C}$, with the same marginal constraints and multiplicative gauge, whenever $\varepsilon \neq 0$ and all divisions are defined. This complexified iteration is again a local parametrization of the scaling equations, not a globally convergent algorithm.

The following finite-dimensional result is the scaling-variable counterpart of the analytic-dependence theorem of Carlier, Pegon and Tamanini [114, Theorem 2.1 and Remark 2.2]. For compactly supported marginals and a continuous cost, their proof shows that normalized Schrödinger potentials, and hence the entropic cost, depend analytically on every real $\varepsilon > 0$; their Remark 2.2 explicitly records the resulting local extension to complex temperatures. We state the discrete version directly in (u, v) and also allow the marginals and cost matrix to vary.

Theorem 8.46 (Local holomorphic continuation of Sinkhorn scalings)**.** *Fix positive histograms $\mathrm{a}^0 \in \Sigma_n$, $\mathrm{b}^0 \in \Sigma_m$, a finite real cost matrix C^0, and a real temperature $\varepsilon_0 > 0$. Choose positive scalings $(\mathrm{u}^0, \mathrm{v}^0)$ of the corresponding Sinkhorn coupling. Then there are complex neighborhoods of $(\varepsilon_0, \mathrm{a}^0, \mathrm{b}^0, \mathrm{C}^0)$, inside the affine constraint $\sum_i \mathrm{a}_i = \sum_j \mathrm{b}_j$, and a unique holomorphic map*

$$(\varepsilon, \mathrm{a}, \mathrm{b}, \mathrm{C}) \mapsto (\mathrm{u}_\varepsilon, \mathrm{v}_\varepsilon) \in \mathbb{C}^n \times \mathbb{C}^m$$

satisfying the linear gauge $\sum_i \mathrm{a}_i^0 u_{\varepsilon,i} = \sum_i \mathrm{a}_i^0 u_i^0$, such that $\mathrm{P}_\varepsilon = \operatorname{diag}(\mathrm{u}_\varepsilon)\mathrm{K}_\varepsilon(\mathrm{C})\operatorname{diag}(\mathrm{v}_\varepsilon)$ has marginals (a, b). In particular, the gauge-fixed scalings and P_ε are holomorphic in all four arguments near the base point.

Proof. Apply the holomorphic implicit-function theorem to the n row residuals, the first $m-1$ column residuals, and the gauge residual $G(\mathrm{u}) = \sum_i \mathrm{a}_i^0(u_i - u_i^0)$. There are $n+m$ equations for the $n+m$ scaling coordinates. It remains to prove that their derivative with respect to (u, v) is invertible at the base point.

Let $(\delta\mathrm{u}, \delta\mathrm{v})$ belong to its kernel and write

$$r_i := \frac{\delta u_i}{u_i^0}, \qquad s_j := \frac{\delta v_j}{v_j^0}.$$

Since $\mathrm{P}^0 = \operatorname{diag}(\mathrm{u}^0)\mathrm{K}_{\varepsilon_0}(\mathrm{C}^0)\operatorname{diag}(\mathrm{v}^0)$, the induced perturbation is $\delta\mathrm{P}_{i,j} = \mathrm{P}^0_{i,j}(r_i + s_j)$. The linearized row sums vanish. The first $m-1$ linearized column sums vanish by definition of the residuals, and the last one vanishes because the total row sum equals the total column sum. Therefore

$$0 = \sum_i \overline{r_i} \sum_j \delta\mathrm{P}_{i,j} + \sum_j \overline{s_j} \sum_i \delta\mathrm{P}_{i,j} = \sum_{i,j} \mathrm{P}^0_{i,j} \left| r_i + s_j \right|^2 .$$

Strict positivity of P^0 gives $r_i = \theta$ and $s_j = -\theta$ for some $\theta \in \mathbb{C}$. The linearized gauge is $0 = \sum_i \mathrm{a}_i^0 \delta u_i = \theta \sum_i \mathrm{a}_i^0 u_i^0$; its last factor is positive, hence $\theta = 0$. The derivative is injective and therefore invertible. The implicit-function theorem now gives the asserted holomorphic scalings, and their matrix product gives the holomorphic coupling. □

After shrinking the neighborhoods if necessary, the scaling coordinates remain nonzero. Choosing their local logarithms defines the holomorphic log-scalings $\mathrm{f}_\varepsilon := \varepsilon \log \mathrm{u}_\varepsilon$ and $\mathrm{g}_\varepsilon := \varepsilon \log \mathrm{v}_\varepsilon$. These derived quantities are convenient for visualization but are not needed in the theorem.

> **Remark 8.47 (Local, not global, complex scaling).** The theorem is local around a positive real point. It does not rule out complex singularities, changes of logarithm branch, or zeros in the denominators of the complexified updates (8.5) farther away. Along a compact interval $[\varepsilon_{\min}, \varepsilon_{\max}] \subset (0, +\infty)$, uniqueness lets the local branches agree on overlaps and therefore cover some neighborhood of that interval. This still gives no global single-valued continuation on $\mathbb{C} \setminus \{0\}$.

Figure 8.17 visualizes the continuation at the level of the coupling, without choosing logarithm branches. For a fixed real part ε_0, it displays $\left|\mathrm{P}_{\varepsilon_0 + i\eta}\right|$ at four increasing imaginary parts. The complex

matrix $P_{\varepsilon_0+i\eta}$ always has marginals (a, b), but taking its entrywise modulus destroys these linear identities. The red and blue side profiles therefore show the prescribed marginals, while the violet profiles show the row and column sums of $\left|P_{\varepsilon_0+i\eta}\right|$ and reveal the cancellations hidden by the modulus.

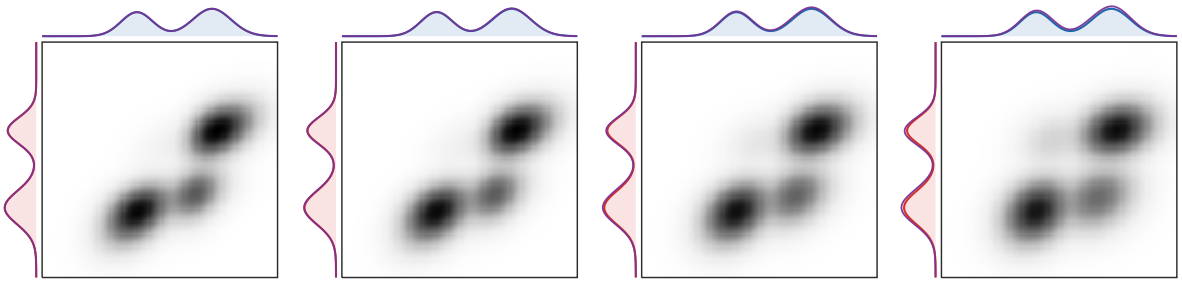


Figure 8.17: The magnitude of a complex Sinkhorn coupling exposes the oscillations created by an imaginary temperature. Two Gaussian-mixture histograms and $\varepsilon_0 = 0.55$ are fixed, while $\varepsilon = \varepsilon_0 + i\eta$ is continued through four nonnegative values of η. Every panel uses the same intensity scale. The attached red and blue profiles are the prescribed marginals (a, b) of the complex coupling; the violet profiles are the marginals of its entrywise modulus and separate from them as cancellation increases.

Example 8.48 (Centered one-dimensional Gaussians). Let $\alpha = \mathcal{N}(0, \sigma_\alpha^2)$ and $\beta = \mathcal{N}(0, \sigma_\beta^2)$ on $\mathbb{R}$, with $\sigma_\alpha, \sigma_\beta > 0$, and take $c(x, y) = (x - y)^2$. Proposition 9.24 shows that, for real $\varepsilon > 0$, the Sinkhorn coupling is Gaussian. Its analytic continuation is obtained by using the same formula: for $\Re(\varepsilon) > 0$, set

$$k_\varepsilon := \frac{\sqrt{\varepsilon^2 + 16\sigma_\alpha^2\sigma_\beta^2} - \varepsilon}{4},$$

where the square root is the holomorphic branch on the right half-plane that is positive for $\varepsilon > 0$. The continued coupling π_ε is a centered complex Gaussian, with covariance parameter

$$\begin{pmatrix} \sigma_\alpha^2 & k_\varepsilon \\ k_\varepsilon & \sigma_\beta^2 \end{pmatrix}.$$

Thus no new Gaussian computation is required: one simply analytically continues the real-temperature expression. This formula is guaranteed to define the finite complex Gaussian coupling throughout $\Re(\varepsilon) > 0$. It can be continued farther along paths on which the Gaussian integral converges and which avoid $\varepsilon = 0$ and the square-root branch points $\varepsilon = \pm 4i\sigma_\alpha\sigma_\beta$.

CHAPTER 9

Entropic Regularization: Convergence

This chapter focuses on algorithmic convergence for entropic optimal transport: the marginals and the temperature are fixed, and one asks how Sinkhorn iterates, soft transforms and related monotone scaling procedures approach their regularized fixed point. Statistical convergence is a separate question, because the empirical marginals then change with the sample size; this is the topic of Chapter 10.

The chapter revisits Sinkhorn convergence through several complementary lenses. Bregman projections explain the alternating-projection geometry, Fortet's order argument gives qualitative fixed-point convergence, and an order-theoretic M-function viewpoint explains why Sinkhorn-like scaling can still converge even when the equations are no longer the gradient of a convex potential. Robust Bregman estimates then give a non-asymptotic $O(1/\ell)$ dual-gap bound, while Hilbert's metric gives a clean linear contraction when the kernel is uniformly positive. The last sections discuss Gaussian closed forms and continuous ε-Sinkhorn flows as model cases where the fixed-point structure becomes explicit.

9.1 Sinkhorn Convergence: Bregman Point of View

This section explains Sinkhorn as alternating Bregman projections. The main message is geometric: each row or column rescaling is the KL projection onto one affine marginal constraint, so convergence follows from the Pythagorean identity for Bregman divergences.

For simplicity, this section is written for discrete measures, but the same ideas carry over to general measures. The robust-rate section later revisits the alternating-projection mechanism through convex duality, with constants expressed through the oscillation of the potentials and of the cost range.

Alternating KL projections. The projection viewpoint explains Sinkhorn as repeated enforcement of one marginal constraint at a time. It is not specific to entropy, although the KL case is the one where the projections reduce to elementary row and column scalings.

The following matrix construction is the finite-dimensional counterpart of the measure-valued Definition 8.35.

Definition 9.1 (Bregman divergence)**.** Let $\Omega \subset \mathbb{R}^{n\times m}$ be convex with nonempty interior, and let $\Phi : \mathbb{R}^{n\times m} \to \mathbb{R} \cup \{+\infty\}$ be a proper lower-semicontinuous Legendre function with $\mathrm{dom}(\Phi) = \Omega$. Thus $\Phi = +\infty$ outside Ω, while Φ is differentiable and strictly convex on $\mathrm{int}(\Omega)$. For $\mathrm{P}, \mathrm{Q} \in \mathrm{int}(\Omega)$, the Bregman divergence generated by Φ is

$$B_\Phi(\mathrm{P}|\mathrm{Q}) := \Phi(\mathrm{P}) - \Phi(\mathrm{Q}) - \langle \nabla\Phi(\mathrm{Q}),\, \mathrm{P} - \mathrm{Q}\rangle.$$

For the quadratic generator $\Phi(\mathrm{P}) = \frac{1}{2}\|\mathrm{P}\|_{\mathrm{F}}^2$ on $\mathbb{R}^{n\times m}$, one recovers half the squared Euclidean distance, $B_\Phi(\mathrm{P}|\mathrm{Q}) = \frac{1}{2}\|\mathrm{P} - \mathrm{Q}\|_{\mathrm{F}}^2$. For the negative entropy $\Phi(\mathrm{P}) = \sum_{i,j} \mathrm{P}_{i,j} \log \mathrm{P}_{i,j}$ on the nonnegative orthant, extended by $+\infty$ outside it, one obtains $B_\Phi(\mathrm{P}|\mathrm{Q}) = \mathrm{KL}(\mathrm{P}|\mathrm{Q})$. For $\mathrm{P} \geqslant 0$ and $\mathrm{Q} > 0$, this identity is understood through the lower-semicontinuous extension, with $0 \log 0 = 0$.

The corresponding projection is the basic operation in alternating Bregman methods.

Definition 9.2 (Bregman projection)**.** For a nonempty convex set C, define

$$\mathrm{Proj}_C^{B_\Phi}(\mathrm{Q}) := \mathrm{argmin}_{\mathrm{P}\in C}\, B_\Phi(\mathrm{P}|\mathrm{Q}),$$

which is set-valued if the minimizer is not unique.

Bregman divergences are useful because their geometry can encode constraints. A Legendre-type generator Φ blows up, or has an infinite derivative, at the boundary of its domain. For negative entropy, positivity is therefore built into the divergence, so one projects onto affine marginal constraints without separately handling non-negativity.

Linear tilts and Gibbs references. The next proposition explains why adding a linear cost to a Bregman penalty merely shifts the reference point in dual coordinates. The usual Gibbs–KL reformulation is the entropy specialization.

Proposition 9.3 (Linear tilts of Bregman penalties)**.** *Let Φ be differentiable and strictly convex, and let B_Φ be its Bregman divergence. Fix a reference point Q in the interior of the domain. Assume that there exists Q^C such that*

$$\nabla\Phi(Q^C) = \nabla\Phi(Q) - C/\varepsilon.$$

Then, for all P in the domain,

$$\langle P, C\rangle + \varepsilon B_\Phi(P|Q) = \varepsilon B_\Phi(P|Q^C) + \mathrm{cst},$$

where the constant does not depend on P.

Proof. Subtract the two Bregman divergences:

$$B_\Phi(P|Q^C) - B_\Phi(P|Q) = \langle \nabla\Phi(Q) - \nabla\Phi(Q^C),\, P\rangle + \mathrm{cst}.$$

Using $\nabla\Phi(Q) - \nabla\Phi(Q^C) = C/\varepsilon$ and multiplying by ε gives the claim. □

For the negative entropy $\Phi(P) = \sum_{i,j} P_{i,j} \log P_{i,j}$, one has $B_\Phi = \mathrm{KL}$. Taking $Q = a \otimes b$ gives the tilted reference

$$K^\varepsilon_{a,b} := (a \otimes b) \odot e^{-C/\varepsilon}.$$

Thus

$$\langle P, C\rangle + \varepsilon \mathrm{KL}(P|a \otimes b) = \varepsilon \mathrm{KL}(P|K^\varepsilon_{a,b}) + \mathrm{cst}.$$

On the transport polytope, scaling $K^\varepsilon_{a,b}$ is equivalent to scaling the Gibbs kernel $K = e^{-C/\varepsilon}$ because the factors a_i and b_j can be absorbed into the Sinkhorn scalings.

Thus the unique solution P_ε of (8.1) is the KL projection of the tilted Gibbs reference onto $U(a, b)$:

$$P_\varepsilon = \mathrm{Proj}^{\mathrm{KL}}_{U(a,b)}(K^\varepsilon_{a,b}) := \operatorname*{argmin}_{P \in U(a,b)} \mathrm{KL}(P|K^\varepsilon_{a,b}). \tag{9.1}$$

Cyclic projection convergence. Given two closed convex constraint sets C_1 and C_2, the cyclic method projects first onto C_1, then onto C_2, and repeats. Full iterates satisfy the second constraint and half-iterates satisfy the first; both constraints are attained together only in the limit.

Algorithm 9.1 records the general iteration, with a stopping rule based on nonnegative defects that vanish on the respective constraint sets.

Algorithm 9.1: Cyclic Bregman projections

Input: Closed convex sets C_1, C_2, Bregman divergence B_Φ, interior point $P^{(0)}$, constraint defects $\mathrm{def}_{C_1}, \mathrm{def}_{C_2}$, tolerance tol, iteration budget $L \geqslant 1$.
Output: Approximate point in $C_1 \cap C_2$ when the intersection is nonempty.
Initialize: Set $r^{(0)} = +\infty$ and $\ell = 0$.
While $r^{(\ell)} \geqslant$ tol and $\ell < L$ **do**:
  Set $\ell \leftarrow \ell + 1$.
  $P^{(\ell-1/2)} = \mathrm{Proj}^{B_\Phi}_{C_1}(P^{(\ell-1)})$ and $P^{(\ell)} = \mathrm{Proj}^{B_\Phi}_{C_2}(P^{(\ell-1/2)})$.
  Set $r^{(\ell)} = \max\{\mathrm{def}_{C_1}(P^{(\ell)}), \mathrm{def}_{C_2}(P^{(\ell)})\}$.
Return $P^{(\ell)}$.

The convergence mechanism is the classical one of Bregman [79]. General convex constraints determine a feasible limit, while affine constraints provide the stronger closest-point characterization.

Proposition 9.4 (Convergence of cyclic Bregman projections)**.** *Let Φ be a Legendre generator on a finite-dimensional convex domain, and let C_1, C_2 be closed convex sets whose intersection meets the interior of the domain. Generate $(P^{(\ell-1/2)}, P^{(\ell)})$ by Algorithm 9.1 from an interior point $P^{(0)}$. Assume that the projections are well-defined and that the full half-step sequence remains in a compact subset of the interior of the domain. Then there exists $\bar{P} \in C_1 \cap C_2$ such that $P^{(\ell)} \to \bar{P}$ and $P^{(\ell-1/2)} \to \bar{P}$.*

If, in addition, C_1 and C_2 are affine, then

$$\bar{P} = \mathrm{Proj}^{B_\Phi}_{C_1 \cap C_2}(P^{(0)}).$$

Proof. For three interior points, the Bregman three-point formula reads

$$B_\Phi(\mathrm{Q}|\mathrm{P}) = B_\Phi(\mathrm{Q}|\mathrm{P}^+) + B_\Phi(\mathrm{P}^+|\mathrm{P}) + \langle \nabla\Phi(\mathrm{P}^+) - \nabla\Phi(\mathrm{P}),\, \mathrm{Q} - \mathrm{P}^+\rangle.$$

If $\mathrm{P}^+ = \mathrm{Proj}_C^{B_\Phi}(\mathrm{P})$ and C is closed and convex, first-order optimality gives

$$\langle \nabla\Phi(\mathrm{P}^+) - \nabla\Phi(\mathrm{P}),\, \mathrm{Q} - \mathrm{P}^+\rangle \geqslant 0 \qquad \forall \mathrm{Q} \in C.$$

Combining these two relations yields the Bregman–Fejér inequality

$$B_\Phi(\mathrm{Q}|\mathrm{P}) \geqslant B_\Phi(\mathrm{Q}|\mathrm{P}^+) + B_\Phi(\mathrm{P}^+|\mathrm{P}) \qquad \forall \mathrm{Q} \in C.$$

Let $(Z^{(r)})_r$ be the half-step sequence, so that $Z^{(2\ell)} = \mathrm{P}^{(\ell)}$ and $Z^{(2\ell+1)} = \mathrm{P}^{(\ell+1/2)}$. Fix $\mathrm{Q} \in C_1 \cap C_2$. Applying the inequality at each half-step gives

$$B_\Phi(\mathrm{Q}|Z^{(r)}) - B_\Phi(\mathrm{Q}|Z^{(r+1)}) \geqslant B_\Phi(Z^{(r+1)}|Z^{(r)}) \geqslant 0.$$

Thus $B_\Phi(\mathrm{Q}|Z^{(r)})$ decreases and $\sum_r B_\Phi(Z^{(r+1)}|Z^{(r)}) < +\infty$. The projection drops tend to zero; compactness and strict convexity imply $\|Z^{(r+1)} - Z^{(r)}\| \to 0$. Every cluster point of either subsequence consequently belongs to both closed sets. Let $\bar{\mathrm{P}} \in C_1 \cap C_2$ be one such cluster point. The sequence $B_\Phi(\bar{\mathrm{P}}|Z^{(r)})$ is nonincreasing and tends to zero along a convergent subsequence, hence tends to zero altogether. Compactness and strict convexity then give $Z^{(r)} \to \bar{\mathrm{P}}$, proving the first claim.

Suppose now that C_1 and C_2 are affine. The first-order inequality above is then an equality because both signs of every feasible direction are admissible. Moreover, each dual displacement $\nabla\Phi(Z^{(r+1)}) - \nabla\Phi(Z^{(r)})$ belongs to the normal space of the affine set used at that half-step. Telescoping and passing to the limit gives

$$\nabla\Phi(\bar{\mathrm{P}}) - \nabla\Phi(\mathrm{P}^{(0)}) \in N_{C_1} + N_{C_2} = N_{C_1 \cap C_2},$$

where the last equality uses affinity and the nonempty intersection. This is precisely the first-order optimality condition for minimizing $\mathrm{Q} \mapsto B_\Phi(\mathrm{Q}|\mathrm{P}^{(0)})$ over $C_1 \cap C_2$, proving the displayed projection identity. □

Row and column scalings. We now apply the preceding Bregman-projection framework to entropic OT. The two affine constraints impose the source and target marginals, while negative entropy turns their cyclic Bregman projections into explicit row and column rescalings.

Denote the two affine marginal sets by

$$C_{\mathrm{a}}^1 := \left\{\mathrm{P} \in \mathbb{R}^{n\times m} \;;\; \mathrm{P}\mathbb{1}_m = \mathrm{a}\right\} \quad \text{and} \quad C_{\mathrm{b}}^2 := \left\{\mathrm{P} \in \mathbb{R}^{n\times m} \;;\; \mathrm{P}^\top \mathbb{1}_n = \mathrm{b}\right\}.$$

Positivity is a separate constraint, so

$$\mathrm{U}(\mathrm{a},\mathrm{b}) = C_{\mathrm{a}}^1 \cap C_{\mathrm{b}}^2 \cap \mathbb{R}_+^{n\times m}.$$

For negative entropy, however, the effective domain of the Legendre generator is $\Omega = \mathbb{R}_+^{n\times m}$ and $\Phi = +\infty$ outside Ω. Consequently, $\mathrm{Proj}_{C_i}^{\mathrm{KL}}$ already minimizes over $C_i \cap \Omega$: positivity is enforced by the generator and need not be repeated in the projector notation. Thus the two half-steps of Algorithm 9.1 specialize to

$$\mathrm{P}^{(\ell+1)} := \mathrm{Proj}_{C_{\mathrm{a}}^1}^{\mathrm{KL}}(\mathrm{P}^{(\ell)}) \quad \text{and} \quad \mathrm{P}^{(\ell+2)} := \mathrm{Proj}_{C_{\mathrm{b}}^2}^{\mathrm{KL}}(\mathrm{P}^{(\ell+1)}). \tag{9.2}$$

These two KL projectors are explicit: they rescale respectively the rows and the columns.

Proposition 9.5 (KL projections are scalings)**.** *One has*

$$\mathrm{Proj}_{C_{\mathrm{a}}^1}^{\mathrm{KL}}(\mathrm{P}) = \mathrm{diag}\left(\frac{\mathrm{a}}{\mathrm{P}\mathbb{1}_m}\right)\mathrm{P} \quad \textit{and} \quad \mathrm{Proj}_{C_{\mathrm{b}}^2}^{\mathrm{KL}}(\mathrm{P}) = \mathrm{P}\,\mathrm{diag}\left(\frac{\mathrm{b}}{\mathrm{P}^\top\mathbb{1}_n}\right).$$

Proof. Consider the problem along each row or column vector to impose a fixed sum $s \in \mathbb{R}_+$ $\min_p \{\mathrm{KL}(p|q) \;;\; \langle p, \mathbb{1}\rangle = s\}$. The Lagrange multiplier equation for this problem reads

$$\log(p/q) + \lambda\mathbb{1} = 0 \quad \Longrightarrow \quad p = uq \quad \text{where} \quad u = e^{-\lambda} > 0.$$

The constraint $\langle p, \mathbb{1}\rangle = s$ is equivalent to $\langle uq, \mathbb{1}\rangle = s$, i.e. $u = s/\sum_i q_i$, which gives the desired scaling formula $p = sq/\sum_i q_i$. □

For positive histograms and a strictly positive Gibbs kernel, the classical iterative-proportional-fitting theorem ensures that the half-steps (9.2) remain in the positive stratum and converge [460, 461]. Since the two marginal sets are affine, Proposition 9.4, applied with $\mathrm{P}^{(0)} = \mathrm{K}^{\varepsilon}_{\mathrm{a,b}}$, identifies the limit as the KL projection (9.1). The Hilbert-metric argument in Section 9.4 gives an independent quantitative proof.

These iterations are equivalent to Sinkhorn iterations (8.5) since defining

$$\mathrm{P}^{(2\ell)} \coloneqq \operatorname{diag}(\mathrm{u}^{(\ell)})\mathrm{K}\operatorname{diag}(\mathrm{v}^{(\ell)}),$$

one has

$$\begin{aligned} \mathrm{P}^{(2\ell+1)} &\coloneqq \operatorname{diag}(\mathrm{u}^{(\ell+1)})\mathrm{K}\operatorname{diag}(\mathrm{v}^{(\ell)}) \\ \text{and} \quad \mathrm{P}^{(2\ell+2)} &\coloneqq \operatorname{diag}(\mathrm{u}^{(\ell+1)})\mathrm{K}\operatorname{diag}(\mathrm{v}^{(\ell+1)}) \end{aligned}$$

In practice, however, one should prefer using (8.5), which only requires manipulating scaling vectors and multiplying by a Gibbs kernel, and can often be accelerated when the kernel has separable, sparse, low-rank or geometric structure.

Such a convergence analysis using Bregman projection is of limited interest because it only works directly in finite dimension. For instance, the linear convergence speed one can obtain from strong convexity degrades with the dimension and with ε. The robust dual analysis below gives a dimension-free qualitative message: the constants are expressed through the oscillation of the potentials and the cost range, and the resulting $O(1/\ell)$ estimate explains what can be guaranteed before any asymptotic linear regime becomes visible. It is also possible to anneal ε during the iterations and to rely on multiscale strategies in low dimensions.

Other divergences. The simplicity of the KL construction relies on negative entropy encoding nonnegativity through its effective domain. Suppose instead that Φ is Legendre on the full space $\mathbb{R}^{n\times m}$. Positivity must then be imposed explicitly in each marginal constraint by defining

$$C^1_{\mathrm{a},+} \coloneqq C^1_{\mathrm{a}} \cap \mathbb{R}^{n\times m}_{+} \quad \text{and} \quad C^2_{\mathrm{b},+} \coloneqq C^2_{\mathrm{b}} \cap \mathbb{R}^{n\times m}_{+}, \tag{9.3}$$

so that $\mathrm{U}(\mathrm{a},\mathrm{b}) = C^1_{\mathrm{a},+} \cap C^2_{\mathrm{b},+}$. These sets are convex but no longer affine, and their B_Φ-projectors generally do not admit a closed-form scaling formula. More importantly, the affine conclusion of Proposition 9.4 no longer applies: ordinary cyclic projections may converge to a feasible point, but not in general to the closest point $\operatorname{Proj}^{B_\Phi}_{\mathrm{U}(\mathrm{a},\mathrm{b})}(\mathrm{P}^{(0)})$.

The Euclidean generator $\Phi(\mathrm{P}) = \frac{1}{2}\|\mathrm{P}\|^2_{\mathrm{F}}$ illustrates both the additional difficulty and the remaining structure. Projection onto $C^1_{\mathrm{a},+}$ is rowwise projection onto a simplex:

$$\left[\operatorname{Proj}^{B_\Phi}_{C^1_{\mathrm{a},+}}(\mathrm{Q})\right]_{i,j} = (\mathrm{Q}_{i,j} - \tau_i)_+, \qquad \sum_j (\mathrm{Q}_{i,j} - \tau_i)_+ = a_i,$$

and the column projector is analogous. The thresholds are not given by a multiplicative scaling, but sorting or selection computes them efficiently [65].

To recover the closest-point projection onto the intersection, one uses Dykstra's algorithm: it alternates the same Bregman projectors while carrying correction variables in dual coordinates. These corrections retain the normal components discarded by plain cyclic projection, and the iterates converge to $\operatorname{Proj}^{B_\Phi}_{C^1_{\mathrm{a},+}\cap C^2_{\mathrm{b},+}}(\mathrm{P}^{(0)})$ under the standard finite-dimensional assumptions [198, 119, 30]. In regularized OT, this construction was developed systematically by Benamou, Carlier, Cuturi, Nenna, and Peyré [37]. Viewed through Fenchel duality, Bregman–Dykstra is iterative block-coordinate ascent on the marginal dual potentials, with the correction variables storing the dual memory [80].

For the quadratic generator and the product reference $\xi = \mathrm{a}\otimes\mathrm{b}$, the Bregman penalty is

$$B_\Phi(\mathrm{P}|\xi) = \frac{1}{2}\sum_{i,j}(\mathrm{P}_{i,j} - a_i b_j)^2.$$

This is closely related, but not identical, to the quadratic φ-divergence generated by $\varphi(r) = \frac{1}{2}(r-1)^2$, which for positive a_i, b_j is

$$\mathcal{D}_\varphi(\mathrm{P}|\mathrm{a}\otimes\mathrm{b}) = \frac{1}{2}\sum_{i,j}\frac{(\mathrm{P}_{i,j} - a_i b_j)^2}{a_i b_j}.$$

The difference is the fixed inverse product-marginal weighting in the second quadratic norm. Both models yield positive-part thresholding and potentially sparse plans. Section 8.9, in particular (8.56) and (8.55), develops their dual laws and alternating dual-maximization algorithms.

9.2 Sinkhorn Convergence: Monotone Point of View

This section isolates the order structure shared by generalized Sinkhorn updates. It first introduces the abstract language of topical maps, then applies it to the generalized soft transforms of Section 8.9, and finally derives a monotone fixed-point argument valid beyond KL regularization.

Variation seminorm and topical maps. Potentials are defined only up to additive constants, so their natural gauge-invariant size is their oscillation.

Definition 9.6 (Variation seminorm)**.** For a bounded real-valued function h, its variation seminorm is

$$\|h\|_V := \sup h - \inf h.$$

It vanishes exactly on constant functions and therefore induces a norm after quotienting by additive constants.

The order maps relevant to Sinkhorn commute with this additive gauge.

Definition 9.7 (Topical map)**.** Let E be an ordered vector space of functions containing the constants. A map $\mathcal{T} : E \to E$ is *topical* if

$$f \leqslant g \Longrightarrow \mathcal{T}(f) \leqslant \mathcal{T}(g), \qquad \mathcal{T}(f+s) = \mathcal{T}(f) + s \quad (s \in \mathbb{R}).$$

Proposition 9.8 (Topical maps are variation-nonexpansive)**.** *Let E be a vector space of bounded real-valued functions, ordered pointwise. Every topical map $\mathcal{T} : E \to E$ satisfies*

$$\|\mathcal{T}(f) - \mathcal{T}(g)\|_V \leqslant \|f - g\|_V.$$

The same conclusion holds if $\mathcal{T}$ is order reversing and additively anti-homogeneous, meaning $\mathcal{T}(f+s) = \mathcal{T}(f) - s$.

Proof. Set $a = \inf(f-g)$ and $b = \sup(f-g)$, so that $g + a \leqslant f \leqslant g + b$. If $\mathcal{T}$ is topical, then

$$\mathcal{T}(g) + a \leqslant \mathcal{T}(f) \leqslant \mathcal{T}(g) + b.$$

Hence the oscillation of $\mathcal{T}(f) - \mathcal{T}(g)$ is at most $b - a = \|f - g\|_V$. In the order-reversing, additively anti-homogeneous case, the corresponding bounds are $\mathcal{T}(g) - b \leqslant \mathcal{T}(f) \leqslant \mathcal{T}(g) - a$ and give the same conclusion. □

> **Remark 9.9 (Topical maps and projective geometry).** Order-preserving additively homogeneous maps are called topical maps in nonlinear Perron–Frobenius theory [335]. Proposition 9.8 is the additive counterpart of projective nonexpansiveness.

Generalized Sinkhorn maps. Consider the φ-divergence regularized transport problem of Section 8.9. Its one-sided block updates, expressed using the Legendre transform φ^* defined in (7.13), are the generalized soft c-transforms of (8.53); their alternating dual-maximization iteration is (8.55). After choosing a consistent extremal minimizer whenever a scalar update is not unique, one full cycle acting on the first potential is

$$\mathcal{A}_\varphi(f) := \left(f^{c,\varepsilon,\varphi}\right)^{\bar{c},\varepsilon,\varphi}. \tag{9.4}$$

For the KL generator, this is the usual double Sinkhorn map.

Proposition 9.10 (Generalized Sinkhorn maps are topical)**.** *Assume that the scalar minimizer sets in (8.53) are nonempty compact intervals, and select either their smallest elements everywhere or their largest elements everywhere. Each selected one-sided transform is order reversing and additively anti-homogeneous:*

$$g \leqslant g' \Longrightarrow g^{\bar{c},\varepsilon,\varphi} \geqslant g'^{\bar{c},\varepsilon,\varphi}, \qquad (g+s)^{\bar{c},\varepsilon,\varphi} = g^{\bar{c},\varepsilon,\varphi} - s,$$

and likewise for $f \mapsto f^{c,\varepsilon,\varphi}$. Consequently, $\mathcal{A}_\varphi$ is topical and

$$\|\mathcal{A}_\varphi(f) - \mathcal{A}_\varphi(h)\|_V \leqslant \|f - h\|_V.$$

No selection convention is needed when the scalar minimizers are unique.

Proof. Since φ is extended by $+\infty$ on $(-\infty, 0)$, its Legendre transform φ^* from (7.13) is convex and nondecreasing. For fixed x, let $H^x_g(u)$ denote the scalar objective in the first line of (8.53). If $g \leqslant g'$, then $u \mapsto H^x_{g'}(u) - H^x_g(u)$ is nondecreasing: this follows by integrating the fact that $s \mapsto \varphi^*(s+\delta) - \varphi^*(s)$ is nondecreasing for every $\delta \geqslant 0$.

Let $I = \operatorname{argmin} H^x_g$ and $I' = \operatorname{argmin} H^x_{g'}$. If $a = \min I$ and $a' = \min I'$ satisfied $a' > a$, the optimality inequalities $H^x_g(a) \leqslant H^x_g(a')$ and $H^x_{g'}(a') \leqslant H^x_{g'}(a)$ would contradict the preceding monotonicity; in the equality case, a would also belong to I', contradicting the definition of a'. Hence $\min I' \leqslant \min I$. The same argument with the maximal minimizers gives $\max I' \leqslant \max I$. Finally, $H^x_{g+s}(u) = H^x_g(u+s) + s$, so either extremal selection is additively anti-homogeneous. The second one-sided transform is treated symmetrically. Their composition is topical, and Proposition 9.8 gives the final estimate. □

Topicality gives nonexpansiveness, not a strict contraction. For KL, positivity of the Gibbs kernel supplies the stronger Hilbert-metric contraction proved in Section 9.4; no comparable strict factor follows from the order properties alone for a general φ.

Monotone convergence for generalized Sinkhorn. The following order argument traces back to Fortet's proof of the Schrödinger system [219, 209, 338]. It is not based on an optimization principle: a subsolution generates an increasing orbit, while a fixed point provides an upper barrier.

Proposition 9.11 (Monotone convergence of generalized Sinkhorn cycles)**.** *Let $\mathcal{X}$ and $\mathcal{Y}$ be compact metric spaces and $c \in C(\mathcal{X} \times \mathcal{Y})$. Suppose that the assumptions of Proposition 9.10 hold, that the selected transforms are finite valued, and that $\mathcal{A}_\varphi$ has a fixed point $f^\star \in C(\mathcal{X})$.*

If $f^{(0)} \in C(\mathcal{X})$ is a subsolution, $f^{(0)} \leqslant \mathcal{A}_\varphi(f^{(0)})$, then the iterates $f^{(\ell+1)} = \mathcal{A}_\varphi(f^{(\ell)})$ converge uniformly to a fixed point of $\mathcal{A}_\varphi$. The analogous conclusion holds for a supersolution, with a decreasing sequence. If the fixed point is unique modulo additive constants, the corresponding potential classes converge to this unique class.

Proof. Because $\mathcal{A}_\varphi$ is additively homogeneous, shifting a subsolution preserves the subsolution inequality. After such a shift, compactness of $\mathcal{X}$ allows us to assume $f^{(0)} \leqslant f^\star$. Topicality then gives

$$f^{(0)} \leqslant f^{(1)} \leqslant \cdots \leqslant f^\star,$$

so the orbit converges pointwise.

The one-sided transforms inherit the modulus of continuity of the cost. Indeed, if $\delta(x, x') := \sup_{y \in \mathcal{Y}} |c(x,y) - c(x',y)|$, order reversal and additive anti-homogeneity give

$$\left| g^{\bar{c},\varepsilon,\varphi}(x) - g^{\bar{c},\varepsilon,\varphi}(x') \right| \leqslant \delta(x, x').$$

Thus $(f^{(\ell)})_{\ell \geqslant 1}$ is uniformly bounded and equicontinuous. Arzelà–Ascoli makes the orbit relatively compact in the uniform norm, and its monotonicity forces every uniformly convergent subsequence to have the same pointwise limit. Hence the full sequence converges uniformly.

Moreover, every selected one-sided transform is nonexpansive for the uniform norm. If $\|g - h\|_\infty \leqslant r$, then $g - r \leqslant h \leqslant g + r$; order reversal and additive anti-homogeneity imply $\|g^{\bar{c},\varepsilon,\varphi} - h^{\bar{c},\varepsilon,\varphi}\|_\infty \leqslant r$, and similarly for the other block. Therefore $\mathcal{A}_\varphi$ is uniformly continuous, and the limit is a fixed point. The supersolution argument reverses all inequalities. Uniqueness modulo constants gives the last claim. □

The subsolution and supersolution conditions are invariant under additive shifts, but an additive shift cannot turn an arbitrary initialization into either one. The proposition therefore isolates the genuinely order-theoretic convergence mechanism; convergence from unrestricted initializations requires an additional compactness, ascent, or contraction argument.

9.3 Sinkhorn Convergence: Sublinear Robust Rate

The preceding projection and monotonicity arguments establish convergence but do not provide a quantitative rate. Section 9.4 will give geometric, or linear, convergence, the canonical asymptotic behavior of a strictly contractive fixed-point iteration; however, the gap between its global contraction factor and one typically becomes exponentially small in the cost oscillation divided by ε. We therefore first establish a slower $O(1/\ell)$ dual-gap rate whose constant grows only as $1/\varepsilon$ [428, 7, 197]. This robust dependence is crucial when Sinkhorn is used to approximate unregularized discrete OT: Corollary 9.13 chooses $\varepsilon = \delta/(2\log(nm))$, so the temperature decreases with the requested accuracy and with the number of support points.

Proposition 9.12 (Robust $O(1/\ell)$ dual rate for discrete Sinkhorn)**.** *Let* $a \in \Sigma_n$ *and* $b \in \Sigma_m$ *be positive histograms, let* $C \in \mathbb{R}^{n\times m}$ *be finite, and write*

$$R := \max_{i,j} C_{i,j} - \min_{i,j} C_{i,j}.$$

Initialize $g^{(0)} = 0$*, perform complete Sinkhorn dual cycles, and denote the resulting potentials by* $(f^{(\ell)}, g^{(\ell)})$ *for* $\ell \geqslant 1$*. Let*

$$\Delta^{(\ell)} := \mathcal{D}_\varepsilon(f^\star, g^\star) - \mathcal{D}_\varepsilon(f^{(\ell)}, g^{(\ell)})$$

be the dual suboptimality gap for the entropic dual objective (8.23)*. Then*

$$0 \leqslant \Delta^{(\ell)} \leqslant \frac{8R^2}{\varepsilon\ell}, \qquad \ell \geqslant 1.$$

Proof. Subtracting $\min_{i,j} C_{i,j}$ from the cost changes all transport values by the same constant and leaves dual gaps and Sinkhorn iterates unchanged up to gauge. We may therefore assume $0 \leqslant C_{i,j} \leqslant R$. For a vector h, introduce the quotient norm

$$|h|_{\mathrm{q}} := \inf_{\lambda\in\mathbb{R}} \|h + \lambda \mathbb{1}\|_\infty = \frac{1}{2}\|h\|_V.$$

Every potential produced by a soft transform has oscillation at most R. The same is true for optimal transformed potentials, hence

$$\max\{|f^{(\ell)}|_{\mathrm{q}}, |g^{(\ell)}|_{\mathrm{q}}\} \leqslant \frac{R}{2}.$$

This is precisely the block-quotient dual radius required by the robust cyclic-projection theorem of [428, Theorem 3.1].

To instantiate that theorem, vectorize P and use the constraint operator $A(P) = (P\mathbb{1}_m, P^\top \mathbb{1}_n)$. Every column of A contains two unit entries, so $\|A\|_{1\to 1} = 2$. After each exact row or column projection, the associated nonnegative matrix has total mass one, giving the uniform primal bound $X = 1$. Substituting $U = R/2$, $X = 1$, and $\|A\|_{1\to 1} = 2$ into the cited theorem yields

$$\Delta^{(\ell)} \leqslant \frac{8XU^2\|A\|_{1\to 1}^2}{\varepsilon\ell} = \frac{8R^2}{\varepsilon\ell}.$$

For completeness, the mechanism behind this substitution is the KL Pythagorean identity. If

$$P^{(\ell)} = (a \otimes b) \odot \exp\left(\frac{f^{(\ell)} \oplus g^{(\ell)} - C}{\varepsilon}\right),$$

then $P^{(\ell)}$ has exact target marginal, hence total mass one, and $\Delta^{(\ell)} = \varepsilon \mathrm{KL}(P^\star | P^{(\ell)})$. A complete cycle gives the exact ascent identity

$$\mathcal{D}_\varepsilon(f^{(\ell+1)}, g^{(\ell+1)}) - \mathcal{D}_\varepsilon(f^{(\ell)}, g^{(\ell)}) = \varepsilon\left[\mathrm{KL}(P^\star|P^{(\ell)}) - \mathrm{KL}(P^\star|P^{(\ell+1)})\right].$$

Pinsker's inequality from Theorem 7.21 controls the marginal residual by this KL ascent, while the quotient-radius bound controls the dual gap by the same residual. Their combination is the scalar recursion

$$\Delta^{(\ell)} - \Delta^{(\ell+1)} \geqslant \frac{\varepsilon}{8R^2}(\Delta^{(\ell)})^2,$$

whose reciprocal telescoping gives the stated $1/\ell$ estimate. When $R = 0$, one complete cycle is already optimal and the conclusion is immediate. □

Corollary 9.13 (Approximating unregularized OT by regularized dual costs)**.** *Under the assumptions of Proposition 9.12, suppose* $nm > 1$*. Let* $\mathcal{D}_{\varepsilon,\ell}$ *be the KL-normalized entropic dual value after* $\ell \geqslant 1$ *Sinkhorn cycles, and define the entropy-corrected lower bound*

$$L_{\varepsilon,\ell} := \mathcal{D}_{\varepsilon,\ell} - \varepsilon \mathrm{H}(a) - \varepsilon \mathrm{H}(b), \qquad \mathrm{H}(a) = -\sum_i a_i \log a_i.$$

Then

$$0 \leqslant \mathrm{L_C}(\mathrm{a},\mathrm{b}) - L_{\varepsilon,\ell} \leqslant \varepsilon \log(nm) + \frac{8R^2}{\varepsilon \ell}.$$

Consequently, the choices

$$\varepsilon = \frac{\delta}{2\log(nm)}, \qquad \ell \geqslant \frac{32R^2 \log(nm)}{\delta^2},$$

produce a lower bound satisfying $0 \leqslant \mathrm{L_C}(\mathrm{a},\mathrm{b}) - L_{\varepsilon,\ell} \leqslant \delta$.

Proof. The KL-normalized objective differs from the entropy convention (8.1) by the constant $\varepsilon\mathrm{H}(\mathrm{a})+\varepsilon\mathrm{H}(\mathrm{b})$ on the transport polytope, because

$$\mathrm{KL}(\mathrm{P}|\mathrm{a}\otimes\mathrm{b}) = -\mathrm{H}(\mathrm{P}) + \mathrm{H}(\mathrm{a}) + \mathrm{H}(\mathrm{b}).$$

Let E_ε be the optimum of the entropy-regularized objective $\langle \mathrm{P}, \mathrm{C}\rangle - \varepsilon\mathrm{H}(\mathrm{P})$. Since $0 \leqslant \mathrm{H}(\mathrm{P}) \leqslant \log(nm)$ for any coupling matrix,

$$\mathrm{L_C}(\mathrm{a},\mathrm{b}) - \varepsilon\log(nm) \leqslant E_\varepsilon \leqslant \mathrm{L_C}(\mathrm{a},\mathrm{b}).$$

The corrected iterate satisfies $L_{\varepsilon,\ell} = E_\varepsilon - \Delta^{(\ell)}$, which gives the displayed value bound after applying Proposition 9.12. The stated choices allocate at most $\delta/2$ to each of the regularization and optimization errors. □

The same identities yield computable diagnostics. If $r = \mathrm{P}\mathbb{1}_m$, the KL projection onto the row constraint satisfies $\mathrm{KL}(\mathrm{Proj}^{\mathrm{KL}}_{C^1_\mathrm{a}}(\mathrm{P})|\mathrm{P}) = \mathrm{KL}(\mathrm{a}|r)$; the analogous column identity uses $s = \mathrm{P}^\top\mathbb{1}_n$. Thus each observed dual ascent is exactly a marginal KL defect, and Theorem 7.21 turns it into an ℓ^1 residual. The residual is not itself the remaining dual gap; a certified upper bound also uses the quotient-radius estimate in the preceding proof.

9.4 Sinkhorn Convergence: Linear Hilbert Metric Rate

Hilbert's projective metric gives a complementary convergence mechanism. Instead of following objective values, it measures distances between positive scaling vectors modulo global multiplication. Positive kernels are contractions in this geometry, yielding a global linear convergence statement.

Projective contraction. As initially explained by [221], the global convergence analysis of Sinkhorn is greatly simplified using Hilbert's projective metric on positive vectors.

Definition 9.14 (Hilbert metric)**.** On $\mathbb{R}^n_{+,*}$, Hilbert's projective metric is

$$\forall\,(\mathrm{u},\mathrm{u}') \in (\mathbb{R}^n_{+,*})^2, \quad d_\mathcal{H}(\mathrm{u},\mathrm{u}') := \|\log(\mathrm{u}) - \log(\mathrm{u}')\|_V. \tag{9.5}$$

Here, $\|\cdot\|_V$ is the variation seminorm of Definition 9.6, specialized to functions on the finite space $\{1,\dots,n\}$: $\|z\|_V = \max_i z_i - \min_i z_i$.

Multiplying both vectors by arbitrary positive constants does not change this quantity, so it is a distance only after passing to projective classes.

Proposition 9.15 (Hilbert metric on the projective cone)**.** *The function $d_\mathcal{H}$ defines a complete distance on the projective cone $\mathbb{R}^n_{+,*}/\!\sim$, where* $\mathrm{u} \sim \mathrm{u}'$ *means that* $\mathrm{u} = s\mathrm{u}'$ *for some* $s > 0$.

Proof. The map $\mathrm{u} \mapsto \log \mathrm{u}$ identifies $\mathbb{R}^n_{+,*}/\!\sim$ with the quotient vector space $\mathbb{R}^n/\mathrm{Span}(\mathbb{1}_n)$, because multiplying u by $s > 0$ adds the constant vector $\log(s)\mathbb{1}_n$. The variation seminorm $\|z\|_V = \max_i z_i - \min_i z_i$ vanishes exactly on constant vectors, so it induces a norm on this quotient. Symmetry, the triangle inequality and separation therefore follow from the corresponding norm properties. Completeness follows because $\mathbb{R}^n/\mathrm{Span}(\mathbb{1}_n)$ is finite-dimensional and all finite-dimensional normed spaces are complete. □

It was introduced independently by [62] and [466] to provide quantitative proofs of the Perron-Frobenius theorem (convergence of iterations of positive matrices). Sinkhorn should be viewed as a nonlinear generalization of Perron-Frobenius.

Theorem 9.16 (Birkhoff contraction theorem)**.** *Let* $\mathrm{K} \in \mathbb{R}^{n\times m}_{+,*}$*. Then, for* $(\mathrm{v}, \mathrm{v}') \in (\mathbb{R}^m_{+,*})^2$*,*

$$d_{\mathcal{H}}(\mathrm{Kv}, \mathrm{Kv}') \leqslant \lambda(\mathrm{K}) d_{\mathcal{H}}(\mathrm{v}, \mathrm{v}'), \qquad \lambda(\mathrm{K}) \coloneqq \frac{\sqrt{\eta(\mathrm{K})}-1}{\sqrt{\eta(\mathrm{K})}+1} < 1, \qquad \eta(\mathrm{K}) \coloneqq \max_{i,j,k,\ell} \frac{\mathrm{K}_{i,k}\mathrm{K}_{j,\ell}}{\mathrm{K}_{j,k}\mathrm{K}_{i,\ell}}.$$

Proof. The Birkhoff–Hopf theorem states that a positive linear map A contracts Hilbert's metric by at most $\tanh(\Delta(A)/4)$, where $\Delta(A) \coloneqq \sup_{u,v>0} d_{\mathcal{H}}(Au, Av)$. This is the classical cone-contraction theorem of Birkhoff [62]. For the matrix K, expand a cross-ratio between output coordinates, $\frac{(\mathrm{K}u)_i (\mathrm{K}v)_j}{(\mathrm{K}u)_j (\mathrm{K}v)_i}$. Every summand in the numerator is bounded by $\eta(\mathrm{K})$ times the corresponding cross-summand in the denominator, so this ratio is at most $\eta(\mathrm{K})$. Conversely, concentrating u and v on two columns approaching a maximizing quadruple attains this bound in the limit. Hence

$$\Delta(\mathrm{K}) = \log \eta(\mathrm{K}), \qquad \eta(\mathrm{K}) = \max_{i,j,k,\ell} \frac{\mathrm{K}_{i,k}\mathrm{K}_{j,\ell}}{\mathrm{K}_{j,k}\mathrm{K}_{i,\ell}}.$$

Therefore $\tanh(\Delta(\mathrm{K})/4) = (\sqrt{\eta(\mathrm{K})}-1)/(\sqrt{\eta(\mathrm{K})}+1)$, which is the claimed factor. □

Nested simplex images. The contraction is already visible on the three-state probability simplex. Adopt the column-vector convention for histograms. Every positive column-stochastic matrix K maps Σ_3 into its interior, and

$$\mathrm{K}^{\ell+1}\Sigma_3 = \mathrm{K}^{\ell}(\mathrm{K}\Sigma_3) \subseteq \mathrm{K}^{\ell}\Sigma_3.$$

To compare isotropic contraction, rotation, and anisotropic contraction, set

$$J_3 \coloneqq \frac{1}{3}\mathbb{1}_3\mathbb{1}_3^{\top}, \qquad A \coloneqq \begin{pmatrix} 0 & -1 & 1 \\ 1 & 0 & -1 \\ -1 & 1 & 0 \end{pmatrix}, \qquad \mathrm{K}_{\rho,\delta} \coloneqq \rho I_3 + (1-\rho)J_3 + \delta A.$$

The matrices are doubly stochastic, and they are strictly positive whenever $|\delta| < (1-\rho)/3$. We use the first two members

$$\rho = .90, \qquad \delta = .014, \qquad \mathrm{K}_1 = \mathrm{K}_{\rho,0}, \qquad \mathrm{K}_2 = \mathrm{K}_{\rho,\delta}.$$

The control K_1 acts as ρI on the zero-sum tangent plane, so its images are parallel homothetic triangles. Since the restriction of A has eigenvalues $\pm \mathrm{i}\sqrt{3}$, the two non-Perron eigenvalues of K_2 are $\rho \pm \mathrm{i}\sqrt{3}\delta$: its tangent action is an isotropic contraction followed by a slow rotation.

To exhibit a genuinely different geometry, take the positive doubly stochastic, non-normal kernel

$$\mathrm{K}_3 = \mathrm{K}_{\mathrm{aniso}} \coloneqq \begin{pmatrix} .950 & .018 & .032 \\ .042 & .880 & .078 \\ .008 & .102 & .890 \end{pmatrix}.$$

Its non-Perron eigenvalues are approximately .9222 and .7978. Their unequal moduli progressively stretch the image in the dominant eigendirection, while non-normality adds shear and makes that direction turn during the transient. Write

$$r_i \coloneqq \max\left\{|z| : z \in \mathrm{spec}(\mathrm{K}_i),\ z \neq 1\right\}$$

for the asymptotic spectral contraction on that plane. Here $r_1 = \rho$, $r_2 = \sqrt{\rho^2 + 3\delta^2} \simeq .9003$, and $r_3 \simeq .9222$. This Euclidean quantity is distinct from the global Birkhoff factor $\lambda(\mathrm{K}_i)$ in Hilbert's metric.

Figure 9.1 contrasts the symmetric control with a rotating normal kernel and an anisotropic non-normal kernel.

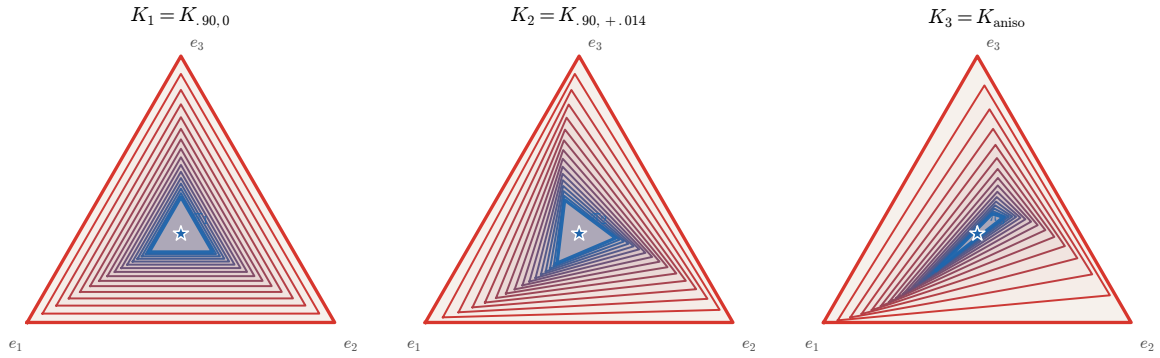


Figure 9.1: Positive Markov kernels contract the three-state simplex and can simultaneously rotate its image. Each colored polygon is $K_i^{\ell}\Sigma_3$: color progresses from red at $\ell = 0$ to blue at $\ell = 15$, and the star is the common stationary vector $\pi_i = \mathbb{1}_3/3$. The isotropic kernel K_1 preserves edge directions, K_2 rotates an essentially homothetic triangle, and the non-normal kernel $K_3 = K_{\text{aniso}}$ both turns and strongly elongates the image because its two tangent modes contract at unequal rates. The red boundary contains zero coordinates and is only a geometric reference: positivity sends the first image into the simplex interior, where Hilbert contraction applies.

The same theorem holds for positive linear maps between proper cones. Order-preserving homogeneous nonlinear maps admit related projective nonexpansiveness and contraction results, but a generic affine map is not covered because projective homogeneity is essential.

The following result, due to Franklin and Lorenz [221], applies Theorem 9.16 to obtain global linear convergence of Sinkhorn's scaling rays.

Theorem 9.17 (Projective linear convergence of Sinkhorn)**.** *Let* $v^{(0)} > 0$*, define the Sinkhorn iterates by* (8.5)*, and fix any optimal scaling pair* $(u^\star, v^\star)$*. For* $\ell \geqslant 1$*, set*

$$P^{(\ell)} := \operatorname{diag}(u^{(\ell)}) K \operatorname{diag}(v^{(\ell)}),$$

and, for $\ell \geqslant 0$*, denote the row-normalized half-step by*

$$P^{(\ell+1/2)} := \operatorname{diag}(u^{(\ell+1)}) K \operatorname{diag}(v^{(\ell)}).$$

Writing $\lambda = \lambda(K)$*, one has*

$$d_{\mathcal{H}}(v^{(\ell)}, v^\star) \leqslant \lambda^{2\ell} d_{\mathcal{H}}(v^{(0)}, v^\star), \qquad d_{\mathcal{H}}(u^{(\ell+1)}, u^\star) \leqslant \lambda^{2\ell+1} d_{\mathcal{H}}(v^{(0)}, v^\star). \tag{9.6}$$

Thus the scaling rays converge linearly. After fixing any continuous gauge, their representatives converge, and $P^{(\ell)} \to P^\star$ *entrywise. The following posterior estimates use the marginal that is not yet exact at the corresponding step; the first holds for* $\ell \geqslant 1$ *and the second for* $\ell \geqslant 0$*:*

$$d_{\mathcal{H}}(u^{(\ell)}, u^\star) \leqslant \frac{d_{\mathcal{H}}(P^{(\ell)}\mathbb{1}_m, a)}{1-\lambda^2} \quad \text{and} \quad d_{\mathcal{H}}(v^{(\ell)}, v^\star) \leqslant \frac{d_{\mathcal{H}}((P^{(\ell+1/2)})^\top \mathbb{1}_n, b)}{1-\lambda^2}. \tag{9.7}$$

Lastly,

$$\|\log(P^{(\ell)}) - \log(P^\star)\|_\infty \leqslant d_{\mathcal{H}}(u^{(\ell)}, u^\star) + d_{\mathcal{H}}(v^{(\ell)}, v^\star), \tag{9.8}$$

where $P^\star$ *is the unique solution of* (8.1)*.*

Proof. Entrywise multiplication by a fixed positive vector and entrywise inversion are isometries of Hilbert's metric. Therefore the row map $R(v) = a/(Kv)$ and the column map $C(u) = b/(K^\top u)$ are both λ-Lipschitz by Theorem 9.16. Their full-cycle compositions $C \circ R$ and $R \circ C$ are λ^2-contractions. Since $v^\star = C(R(v^\star))$ and $u^\star = R(C(u^\star))$, iteration gives (9.6).

For any contraction F with factor $q < 1$ and fixed point $x^\star$, the triangle inequality gives

$$d(x, x^\star) \leqslant d(x, Fx) + d(Fx, Fx^\star) \leqslant d(x, Fx) + q\, d(x, x^\star),$$

hence $d(x, x^\star) \leqslant d(x, Fx)/(1-q)$. Apply this with $q = \lambda^2$ to the two full-cycle maps. The scaling identities imply

$$d_{\mathcal{H}}(u^{(\ell+1)}, u^{(\ell)}) = d_{\mathcal{H}}(a, P^{(\ell)}\mathbb{1}_m),$$

and

$$d_{\mathcal{H}}(\mathrm{v}^{(\ell+1)}, \mathrm{v}^{(\ell)}) = d_{\mathcal{H}}(\mathrm{b}, (\mathrm{P}^{(\ell+1/2)})^{\top} \mathbb{1}_n),$$

which proves (9.7).

Finally, write $\xi_i = \log(u_i^{(\ell)}/u_i^{\star})$ and $\zeta_j = \log(v_j^{(\ell)}/v_j^{\star})$. Then

$$\log(\mathrm{P}_{i,j}^{(\ell)}/\mathrm{P}_{i,j}^{\star}) = \xi_i + \zeta_j, \qquad \operatorname{osc}(\xi \oplus \zeta) = \operatorname{osc}(\xi) + \operatorname{osc}(\zeta).$$

Both plans have total mass one, so $\sum_{i,j} \mathrm{P}_{i,j}^{\star} e^{\xi_i+\zeta_j} = 1$. Consequently $\min(\xi \oplus \zeta) \leqslant 0 \leqslant \max(\xi \oplus \zeta)$, and the sup norm of $\xi \oplus \zeta$ is bounded by its oscillation. This proves (9.8). □

Nonlinear Sinkhorn images of the simplex. The preceding contraction can be visualized simultaneously for every possible scaling ray, rather than along a single initialization. Using the row and column maps from the proof, define the full-cycle map on the left scaling by

$$F_{\mathrm{u}}(\mathrm{u}) \coloneqq R(C(\mathrm{u})) = \mathrm{a} \oslash \left[\mathrm{K}\left(\mathrm{b} \oslash (\mathrm{K}^{\top}\mathrm{u})\right)\right].$$

It is positively homogeneous, $F_{\mathrm{u}}(s\mathrm{u}) = sF_{\mathrm{u}}(\mathrm{u})$ for $s > 0$, and therefore induces the projective self-map

$$\widehat{F}_{\mathrm{u}}(p) \coloneqq \frac{F_{\mathrm{u}}(p)}{\langle F_{\mathrm{u}}(p),\, \mathbb{1}_3\rangle}, \qquad p \in \Sigma_3.$$

Thus $\widehat{\mathrm{u}}^{(\ell)} \coloneqq \mathrm{u}^{(\ell)}/\langle \mathrm{u}^{(\ell)},\, \mathbb{1}_3\rangle$ follows $\widehat{\mathrm{u}}^{(\ell+1)} = \widehat{F}_{\mathrm{u}}(\widehat{\mathrm{u}}^{(\ell)})$ when ℓ indexes complete Sinkhorn cycles. Since $\widehat{F}_{\mathrm{u}}$ maps Σ_3 into itself, its set-valued iterates are nested:

$$\widehat{F}_{\mathrm{u}}^{\ell+1}(\Sigma_3) = \widehat{F}_{\mathrm{u}}^{\ell}\left(\widehat{F}_{\mathrm{u}}(\Sigma_3)\right) \subseteq \widehat{F}_{\mathrm{u}}^{\ell}(\Sigma_3).$$

Unlike the linear images in Figure 9.1, these sets need not be polygons. Theorem 9.17 quantifies their projective shrinkage: since the first image is strictly positive,

$$\operatorname{diam}_{d_{\mathcal{H}}}\left(\widehat{F}_{\mathrm{u}}^{\ell}(\Sigma_3)\right) \leqslant \lambda(\mathrm{K})^{2(\ell-1)} \operatorname{diam}_{d_{\mathcal{H}}}\left(\widehat{F}_{\mathrm{u}}(\Sigma_3)\right), \qquad \ell \geqslant 1.$$

Figure 9.2 reuses the three kernels K_i from Figure 9.1, with the same uniform Sinkhorn marginals $\mathrm{a} = \mathrm{b} = \mathbb{1}_3/3$ in every panel. This isolates the passage from the linear action K_i^{ℓ} to the nonlinear balancing map $\widehat{F}_{\mathrm{u},i}^{\ell}$. The symmetric control remains aligned, K_2 turns the curved images, and the unequal tangent rates of K_3 additionally produce a pronounced anisotropic collapse. Here the kernels are doubly stochastic, so the normalized Sinkhorn fixed ray and stationary Markov vector both equal $\mathbb{1}_3/3$; they need not coincide for general kernels and marginals.

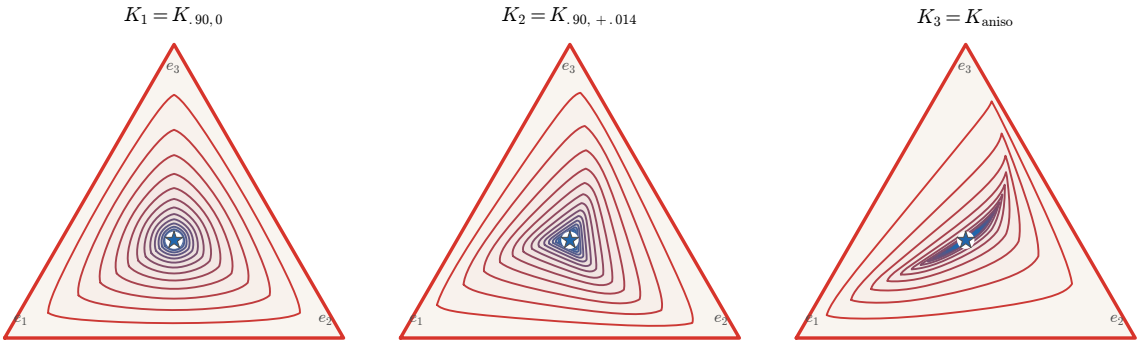


Figure 9.2: Complete Sinkhorn cycles curve, turn and contract the simplex of normalized left scalings. Panel i uses the kernel K_i identified in Figure 9.1 and the common marginals $\mathrm{a} = \mathrm{b} = \mathbb{1}_3/3$. Color shows the densely sampled boundary of $\widehat{F}_{\mathrm{u},i}^{\ell}(\Sigma_3)$, progressing from the red triangle at $\ell = 0$ to the blue curve at $\ell = 15$; the star marks $\widehat{\mathrm{u}}_i^{\star} = \mathbb{1}_3/3$. Coordinatewise reciprocals curve every image; K_2 adds a gradual turn, whereas the non-normal K_3 combines turning with a much stronger collapse across one tangent direction. Since these kernels are invertible, the projective maps are injective and the sampled curves are the actual boundaries of the nested images. The outer boundary contains zeros; the Hilbert estimate starts after one positive cycle.

Dual-potential form of the contraction. By (8.20), the KL-normalized potentials satisfy $\mathrm{f}^{(\ell)} = \varepsilon \log(\mathrm{u}^{(\ell)} \oslash \mathrm{a})$ and $\mathrm{g}^{(\ell)} = \varepsilon \log(\mathrm{v}^{(\ell)} \oslash \mathrm{b})$. The Hilbert contraction can therefore be read directly on potential classes because

$$\|\mathrm{f}^{(\ell)} - \mathrm{f}^\star\|_V = \varepsilon d_{\mathcal{H}}(\mathrm{u}^{(\ell)}, \mathrm{u}^\star), \qquad \|\mathrm{g}^{(\ell)} - \mathrm{g}^\star\|_V = \varepsilon d_{\mathcal{H}}(\mathrm{v}^{(\ell)}, \mathrm{v}^\star).$$

Thus a complete cycle contracts both potential classes at the rate $\lambda(\mathrm{K})^2$. Moreover, the temperature factor must be retained when passing from potentials to densities:

$$\|\log \frac{\mathrm{d}\pi^{(\ell)}}{\mathrm{d}\pi^\star}\|_\infty = \frac{1}{\varepsilon}\|(\mathrm{f}^{(\ell)} - \mathrm{f}^\star) \oplus (\mathrm{g}^{(\ell)} - \mathrm{g}^\star)\|_\infty \leqslant \frac{\|\mathrm{f}^{(\ell)} - \mathrm{f}^\star\|_V + \|\mathrm{g}^{(\ell)} - \mathrm{g}^\star\|_V}{\varepsilon}.$$

Here the last inequality uses the common mass normalization exactly as in the proof of (9.8). For the Gibbs kernel $\mathrm{K}_\varepsilon = e^{-\mathrm{C}/\varepsilon}$, let $R = \max_{i,j} \mathrm{C}_{i,j} - \min_{i,j} \mathrm{C}_{i,j}$. With $\eta = \eta(\mathrm{K}_\varepsilon)$ as in Theorem 9.16,

$$\lambda = \frac{\sqrt{\eta} - 1}{\sqrt{\eta} + 1} \leqslant \tanh(R/(2\varepsilon)) < 1.$$

Indeed, every logarithmic kernel cross-ratio is bounded by $2R/\varepsilon$. The posterior estimates (9.7) become

$$\|\mathrm{f}^{(\ell)} - \mathrm{f}^\star\|_V \leqslant \frac{\varepsilon}{1 - \lambda^2}\|\log((\mathrm{P}^{(\ell)}\mathbb{1}_m) \oslash \mathrm{a})\|_V,$$

and

$$\|\mathrm{g}^{(\ell)} - \mathrm{g}^\star\|_V \leqslant \frac{\varepsilon}{1 - \lambda^2}\|\log(((\mathrm{P}^{(\ell+1/2)})^\top \mathbb{1}_n) \oslash \mathrm{b})\|_V.$$

These are genuine posterior bounds because they use the marginal that has not yet been normalized.

The marginal violations $\|\mathrm{P}^{(\ell)}\mathbb{1}_m - \mathrm{a}\|_1$ and $\|(\mathrm{P}^{(\ell+1/2)})^\top \mathbb{1}_n - \mathrm{b}\|_1$ remain useful practical stopping criteria. At each stage one must monitor the marginal that has not just been normalized; these two residuals vanish only as the iteration approaches a common fixed point. Theorem 9.17 shows global linear convergence in projective geometry, but its worst-case rate becomes exponentially bad as $\varepsilon \to 0$: $1 - \lambda(\mathrm{K}_\varepsilon)^2$ can be of order $e^{-R/\varepsilon}$. In practice, one often observes a much better local linear regime after enough iterations, and sharper continuous analyses can replace this exponential dependence by polynomial rates under additional semiconcavity or log-concavity assumptions [139]. The same Hilbert-metric mechanism extends beyond finite matrices to positive integral operators under suitable compactness and positivity assumptions. An important limitation of this analysis is that it requires a uniformly bounded cost and a kernel bounded away from degeneracy; Gaussian distributions with quadratic cost therefore require a different approach.

9.5 Local Convergence Analysis and Acceleration

The projective analysis of Section 9.4 gives an initialization-independent global rate from the kernel alone, but this robust bound can be pessimistic near the solution. Linearization reveals a more precise structure: the dual Hessian, the Jacobian of a complete Sinkhorn cycle and the curvature of the entropic semi-dual are all controlled by the same conditional-expectation operator. This operator gives the sharp local factor and motivates two accelerations: over-relaxing the alternating soft transforms, or optimizing the semi-dual after eliminating one potential.

Conditional operator and maximal correlation. The local geometry is intrinsic to the coupling and does not depend on whether it is parameterized by potentials or scalings. Let $L^2_0(\alpha) := \{h \in L^2(\alpha) \,;\, \int h \mathrm{d}\alpha = 0\}$ and define $L^2_0(\beta)$ similarly.

Definition 9.18 (Conditional coupling operator)**.** For a coupling $\pi \in \mathcal{U}(\alpha, \beta)$, let $\pi(\mathrm{d}x, \mathrm{d}y) = \alpha(\mathrm{d}x)\pi_x(\mathrm{d}y) = \beta(\mathrm{d}y)\pi^y(\mathrm{d}x)$ be its two disintegrations. Define $T_\pi : L^2(\beta) \to L^2(\alpha)$ and $T^*_\pi : L^2(\alpha) \to L^2(\beta)$ by

$$(T_\pi k)(x) := \int_{\mathcal{Y}} k(y)\mathrm{d}\pi_x(y), \qquad (T^*_\pi h)(y) := \int_{\mathcal{X}} h(x)\mathrm{d}\pi^y(x). \tag{9.9}$$

The centered maximal correlation of π is

$$\sigma(\pi) := \sup_{\substack{k \in L^2_0(\beta) \\ \|k\|_{L^2(\beta)} = 1}} \|T_\pi k\|_{L^2(\alpha)} = \sup_{\substack{h \in L^2_0(\alpha),\, k \in L^2_0(\beta) \\ \|h\|_{L^2(\alpha)} = \|k\|_{L^2(\beta)} = 1}} \int h(x)k(y)\mathrm{d}\pi(x, y). \tag{9.10}$$

For the optimal entropic coupling π_ε of (8.13), set $T_\varepsilon := T_{\pi_\varepsilon}$ and $\sigma_\varepsilon := \sigma(\pi_\varepsilon)$.

The notation T_π^* is literal: the two operators are adjoint for the $L^2(\alpha)$ and $L^2(\beta)$ inner products. Conditional Jensen's inequality gives $\|T_\pi\| \leqslant 1$, while $T_\pi 1 = T_\pi^* 1 = 1$, so both operators preserve centered functions and $0 \leqslant \sigma(\pi) \leqslant 1$. In the discrete setting of (8.1), assuming the weights are positive, the optimizer P_ε gives

$$T_\varepsilon = \operatorname{diag}(\mathrm{a})^{-1}\mathrm{P}_\varepsilon, \qquad T_\varepsilon^* = \operatorname{diag}(\mathrm{b})^{-1}\mathrm{P}_\varepsilon^\top, \tag{9.11}$$

where adjointness is understood in the a- and b-weighted Euclidean inner products. Thus σ_ε is the largest nonconstant singular value of the normalized optimal coupling. Equivalently, it is the second singular value of $\operatorname{diag}(\mathrm{a})^{-1/2}\mathrm{P}_\varepsilon \operatorname{diag}(\mathrm{b})^{-1/2}$, whose first singular value is one.

Dual Hessian and exact local rate. Recall that $\mathcal{D}_\varepsilon$ is the continuous dual objective (8.23), its optimizers $(f_\varepsilon, g_\varepsilon)$ are unique up to the gauge $(f, g) \mapsto (f + s, g - s)$ by Proposition 8.22, and Sinkhorn alternates the soft transforms (8.27). The following proposition identifies the linearization of all these objects without redefining them.

Proposition 9.19 (Dual Hessian and local Sinkhorn rate)**.** *Assume the setting of Proposition 8.20, and that differentiation under the Gibbs integrals and the stated L^2 Taylor expansions are valid. These requirements are automatic in finite spaces and form the local regularity hypothesis in continuous settings. For bounded directions (h, k) and (h', k'),*

$$-\varepsilon D^2\mathcal{D}_\varepsilon(f_\varepsilon, g_\varepsilon)[(h, k), (h', k')] = \int_{\mathcal{X}\times\mathcal{Y}} (h(x) + k(y))(h'(x) + k'(y))\mathrm{d}\pi_\varepsilon(x, y). \tag{9.12}$$

Hence the Hessian form extends continuously to $L^2(\alpha) \oplus L^2(\beta)$, where the positive dual curvature operator is

$$-\nabla^2\mathcal{D}_\varepsilon(f_\varepsilon, g_\varepsilon) = \frac{1}{\varepsilon}\begin{pmatrix} \mathrm{Id} & T_\varepsilon \\ T_\varepsilon^* & \mathrm{Id} \end{pmatrix}. \tag{9.13}$$

After quotienting by the gauge direction $(1, -1)$ and using the product-L^2 quotient norm, its spectrum is contained in

$$\left[\frac{1 - \sigma_\varepsilon}{\varepsilon}, \frac{2}{\varepsilon}\right].$$

Both endpoints belong to the spectrum, although the lower one need not be an eigenvalue. In particular, when $\sigma_\varepsilon < 1$, the spectral condition number of the full dual is $2/(1 - \sigma_\varepsilon)$.

The derivatives of the two soft-transform updates (8.25)–(8.26) at the optimum are

$$D[g \mapsto g^{\bar{c},\varepsilon}](g_\varepsilon) = -T_\varepsilon, \qquad D[f \mapsto f^{c,\varepsilon}](f_\varepsilon) = -T_\varepsilon^*. \tag{9.14}$$

Consequently, the Jacobian of a complete cycle $g \mapsto (g^{\bar{c},\varepsilon})^{c,\varepsilon}$ is $T_\varepsilon^* T_\varepsilon$. If the gauges are fixed so that $e^{(\ell)} := g^{(\ell)} - g_\varepsilon \in L_0^2(\beta)$, then

$$e^{(\ell+1)} = T_\varepsilon^* T_\varepsilon e^{(\ell)} + o\left(\|e^{(\ell)}\|_{L^2(\beta)}\right), \qquad \limsup_{\ell\to\infty} \frac{\|e^{(\ell+1)}\|_{L^2(\beta)}}{\|e^{(\ell)}\|_{L^2(\beta)}} \leqslant \sigma_\varepsilon^2. \tag{9.15}$$

If σ_ε^2 is an isolated eigenvalue and the asymptotic error has a nonzero projection onto its eigenspace, the ratio in (9.15) converges to σ_ε^2.

Proof. Twice differentiating the exponential term in (8.23) gives (9.12); the density law (8.24) identifies its reference measure with π_ε. Its two diagonal terms are the L^2 inner products of the marginals, and its cross term is represented by T_ε, which proves (9.13).

Each quotient class has a unique representative orthogonal to the gauge direction, hence of the form $(h_0 + m, k_0 + m)$ with $h_0 \in L_0^2(\alpha)$ and $k_0 \in L_0^2(\beta)$. On the centered component, Cauchy–Schwarz and (9.10) give

$$(1 - \sigma_\varepsilon)\big(\|h_0\|_{L^2(\alpha)}^2 + \|k_0\|_{L^2(\beta)}^2\big) \leqslant \|h_0\|_{L^2(\alpha)}^2 + \|k_0\|_{L^2(\beta)}^2 + 2\langle h_0, T_\varepsilon k_0\rangle_{L^2(\alpha)} \leqslant (1 + \sigma_\varepsilon)\big(\|h_0\|_{L^2(\alpha)}^2 + \|k_0\|_{L^2(\beta)}^2\big).$$

The common-constant direction $(1, 1)$ has eigenvalue $2/\varepsilon$, whereas $(1, -1)$ is the gauge kernel. Approximate maximizers in (9.10) show that the lower spectral endpoint is $(1 - \sigma_\varepsilon)/\varepsilon$, while the constant

direction gives the upper endpoint $2/\varepsilon$. This proves the condition-number formula without requiring a singular-vector basis.

Differentiating either log-partition formula in Definition 8.21 produces minus the corresponding conditional expectation under π_ε, proving (9.14). The chain rule and a Taylor expansion then give (9.15). Since $\|T_\varepsilon^* T_\varepsilon\|_{L_0^2(\beta)\to L_0^2(\beta)} = \sigma_\varepsilon^2$, the rate follows. □

In finite spaces, this recovers the second-singular-value description of the asymptotic matrix-scaling rate [309, 333].

The same spectral gap has a direct probabilistic interpretation. The law of total variance gives

$$1 - \sigma_\varepsilon^2 = \inf_{\substack{k \in L_0^2(\beta) \\ \|k\|_{L^2(\beta)}=1}} \int_{\mathcal{X}} \mathrm{Var}_{\pi_{\varepsilon,x}}\big(k(Y)\big)\,\mathrm{d}\alpha(x). \tag{9.16}$$

Thus Sinkhorn slows down when the optimal conditional laws $Y|X = x$ become nearly deterministic. This is exactly what happens as $\varepsilon \to 0$ in a smooth Monge regime.

Proposition 9.20 (The global Hilbert factor controls the local rate)**.** *In the discrete positive-kernel setting of Theorem 9.17, let $\lambda(\mathrm{K})$ be the Birkhoff contraction factor defined in Theorem 9.16. Then*

$$\sigma_\varepsilon \leqslant \lambda(\mathrm{K}) < 1, \qquad \sigma_\varepsilon^2 \leqslant \lambda(\mathrm{K})^2. \tag{9.17}$$

Proof. Let $R(\mathrm{v}) = \mathrm{a} \oslash (\mathrm{K}\mathrm{v})$ and $C(\mathrm{u}) = \mathrm{b} \oslash (\mathrm{K}^\top \mathrm{u})$ be the row and column scaling maps used in the proof of Theorem 9.17. Entrywise inversion and multiplication by a fixed positive vector are isometries of Hilbert's metric. Theorem 9.16, applied successively to K and $\mathrm{K}^\top$, therefore gives the pairwise estimate

$$d_{\mathcal{H}}\big(C(R(\mathrm{v})), C(R(\widetilde{\mathrm{v}}))\big) \leqslant \lambda(\mathrm{K})^2 d_{\mathcal{H}}(\mathrm{v}, \widetilde{\mathrm{v}}),$$

where $\lambda(\mathrm{K}^\top) = \lambda(\mathrm{K})$ follows directly from its cross-ratio definition. Under the logarithmic change of variables $g = \varepsilon \log \mathrm{v}$, Hilbert's metric becomes $\varepsilon^{-1}\|\cdot\|_V$ by (9.5). Hence the complete dual update $\mathcal{S}(g) := (g^{\bar{c},\varepsilon})^{c,\varepsilon}$ is $\lambda(\mathrm{K})^2$-Lipschitz for the variation norm on the quotient by constants:

$$\|\mathcal{S}(g) - \mathcal{S}(\widetilde{g})\|_V \leqslant \lambda(\mathrm{K})^2 \|g - \widetilde{g}\|_V.$$

Differentiate this inequality at the fixed point g_ε. Proposition 9.19 gives $D\mathcal{S}(g_\varepsilon) = T_\varepsilon^* T_\varepsilon$, and therefore

$$\|T_\varepsilon^* T_\varepsilon h\|_V \leqslant \lambda(\mathrm{K})^2 \|h\|_V \qquad \text{for every } h \in L_0^2(\beta).$$

In the present finite-dimensional setting, $T_\varepsilon^* T_\varepsilon$ is positive and self-adjoint for the β-weighted inner product, and its largest eigenvalue on $L_0^2(\beta)$ is σ_ε^2. If $\sigma_\varepsilon > 0$, take a corresponding eigenvector $h_\star$. It is nonconstant, so $\|h_\star\|_V > 0$, and the preceding bound gives

$$\sigma_\varepsilon^2 \|h_\star\|_V = \|T_\varepsilon^* T_\varepsilon h_\star\|_V \leqslant \lambda(\mathrm{K})^2 \|h_\star\|_V.$$

The conclusion is immediate; the case $\sigma_\varepsilon = 0$ is trivial. □

The same argument applies when the Gibbs kernel is continuous and strictly positive on compact marginal supports: the corresponding optimal coupling has a density relative to $\alpha \otimes \beta$ that is bounded above and below, the conditional operator is compact, and the continuous Birkhoff–Hopf coefficient bounds σ_ε strictly below one.

The proposition makes the relation with the preceding section precise: maximal correlation is the sharp local quantity, whereas Birkhoff contraction supplies an input-level global upper bound. The latter can become exponentially close to one as $\varepsilon \to 0$. In a smooth nondegenerate Monge regime, the conditional laws have width $O(\sqrt{\varepsilon})$. If a uniform Laplace expansion turns the conditional-variance form in (9.16) into ε times a coercive weighted Dirichlet form, then $1 - \sigma_\varepsilon^2 \asymp \varepsilon$. This explains why observed local rates can be polynomially better than the universal Hilbert estimate. Related non-asymptotic continuous bounds with polynomial dependence on ε are proved under regularity and log-concavity hypotheses in [139]; Proposition 9.26 gives an explicit Gaussian instance of the local calculation.

Blockwise over-relaxation. Sinkhorn performs exact maximization in one dual block after the other. A minimal acceleration extrapolates each block toward its soft-transform maximizer [517, 333]: for $\omega > 0$,

$$f^{(\ell+1)} = (1-\omega)f^{(\ell)} + \omega(g^{(\ell)})^{\bar{c},\varepsilon}, \qquad g^{(\ell+1)} = (1-\omega)g^{(\ell)} + \omega(f^{(\ell+1)})^{c,\varepsilon}. \tag{9.18}$$

The choice $\omega = 1$ is the original iteration (8.27); $0 < \omega < 1$ damps the updates, while $1 < \omega < 2$ over-relaxes them.

Proposition 9.21 (Optimal local block relaxation)**.** *Assume the hypotheses of Proposition 9.19 and $\sigma_\varepsilon < 1$. For every $0 < \omega < 2$, the iteration (9.18) is locally linearly convergent modulo the additive gauge. Its optimal relaxation parameter and corresponding asymptotic factor are*

$$\omega_\star = \frac{2}{1+\sqrt{1-\sigma_\varepsilon^2}}, \qquad r_\star = \omega_\star - 1 = \frac{1-\sqrt{1-\sigma_\varepsilon^2}}{1+\sqrt{1-\sigma_\varepsilon^2}}. \tag{9.19}$$

For $\omega = 1$, the factor is σ_ε^2, as in Proposition 9.19.

Proof. Use (9.14). The polar decomposition of T_ε and the spectral theorem for $(T_\varepsilon^* T_\varepsilon)^{1/2}$ reduce the centered linearized iteration to scalar spectral values $s \in [0, \sigma_\varepsilon]$; in finite dimensions these are the usual singular values. On the corresponding two-dimensional mode, one relaxed cycle has matrix

$$\mathsf{M}_\omega(s) = \begin{pmatrix} 1-\omega & -\omega s \\ -\omega s(1-\omega) & 1-\omega+\omega^2 s^2 \end{pmatrix},$$

whose characteristic polynomial is

$$z^2 - \left(2(1-\omega) + \omega^2 s^2\right) z + (1-\omega)^2.$$

The supremum of the root moduli is reached, or approached, as $s \to \sigma_\varepsilon$. The two roots of this limiting mode have equal modulus precisely when $\omega^2\sigma_\varepsilon^2 = 4(\omega - 1)$. The smaller solution is $\omega_\star$ in (9.19); at this value both root moduli equal $\omega_\star - 1$. Below this threshold the dominant real root decreases with ω, while above it the conjugate-root modulus $\omega - 1$ increases. This proves optimality and local convergence for $0 < \omega < 2$. The nongauge constant mode has factor $(1-\omega)^2$ and never dominates these centered modes. □

Set $\delta_\varepsilon = 1 - \sigma_\varepsilon^2$. When δ_ε is small, the exact expression in (9.19) gives the complete-cycle factor

$$r_\star = \frac{1-\sqrt{\delta_\varepsilon}}{1+\sqrt{\delta_\varepsilon}} = 1 - 2\sqrt{\delta_\varepsilon} + 2\delta_\varepsilon O(\delta_\varepsilon^{3/2}),$$

instead of $\sigma_\varepsilon^2 = 1 - \delta_\varepsilon$ for ordinary Sinkhorn. The coefficient two is specific to measuring one complete two-block cycle. Per half-step, the corresponding factor is $\sqrt{r_\star} = 1 - \sqrt{\delta_\varepsilon} + O(\delta_\varepsilon)$. This is a local prescription. In the finite positive-kernel setting, the Hilbert estimate used in Theorem 9.17 yields global convergence of the relaxed iteration for the conservative range [333] $0 < \omega < 2/(1+\lambda(\mathrm{K}))$; larger choices require the adaptive safeguards analyzed in [517, 333].

Figure 9.3 compares this local prediction with nonlinear iterations. A common ordinary-Sinkhorn warm start places all five runs in the local regime before their relaxation parameters are changed.

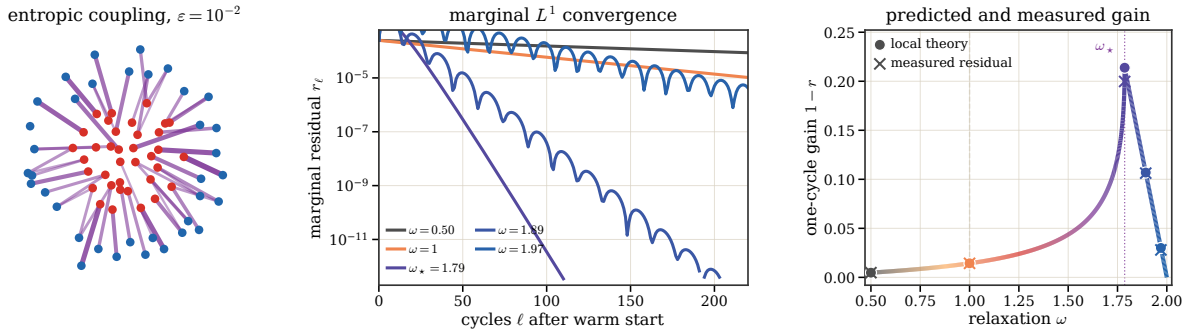


Figure 9.3: Blockwise over-relaxation accelerates local Sinkhorn convergence. Left: the entropic coupling at $\varepsilon = 10^{-2}$; only entries larger than 35% of the maximal mass are drawn, with segment widths proportional to their masses. Middle: after 300 common ordinary-Sinkhorn warm-up cycles, the residual $r_\ell := \frac{1}{2}(\|\mathrm{P}^{(\ell)}\mathbb{1} - \mathrm{a}\|_1 + \|(\mathrm{P}^{(\ell)})^\top\mathbb{1} - \mathrm{b}\|_1)$ is shown for $\omega \in \{0.5, 1, \omega_\star, (\omega_\star + 2)/2, 1.97\}$; $\omega = 1$ is ordinary Sinkhorn, and curves stop at numerical roundoff. Right: the continuous curve is the theoretical one-cycle gain $1 - r_{\mathrm{loc}}(\omega)$, where $r_{\mathrm{loc}}(\omega)$ is the spectral radius of the centered linearization obtained from the maximal correlation $\sigma_\varepsilon \simeq 0.993$ of the limiting coupling. Filled circles mark this prediction at the five displayed parameters; crosses show the empirical gains $1 - \widehat{r}_{\mathrm{marg}}(\omega)$ obtained by fitting $\log r_\ell$ affinely over the final 80 available iterates above the numerical floor. Equation (9.19) gives $\omega_\star \simeq 1.79$.

Entropic semi-dual and variable projection. The hard semi-dual $\mathcal{E}_0$ in (6.2) eliminates one potential from the full dual $\mathcal{D}_0$. At positive temperature, the full dual is $\mathcal{D}_\varepsilon$ from (8.23), and the soft $\bar{c}$-transform $g^{\bar{c},\varepsilon}$ is defined by (8.26). Applying the same partial maximization gives the entropic semi-dual

$$\mathcal{E}_\varepsilon(g) := \mathcal{D}_\varepsilon(g^{\bar{c},\varepsilon}, g) = \int_{\mathcal{X}} g^{\bar{c},\varepsilon}\mathrm{d}\alpha + \int_{\mathcal{Y}} g\mathrm{d}\beta. \tag{9.20}$$

The exponential term disappears in the last equality because the first marginal is normalized by the soft transform. This exact block elimination is an instance of variable projection (VarPro) [244, 245]. The VarPro principle is to optimize out one block of variables exactly whenever its conditional optimizer is available at an acceptable cost. At the level of optimization geometry, this reduction is always favorable: it removes variables and cannot worsen the local spectral conditioning, as quantified below. Its computational value depends on the cost of the inner solve.

VarPro was introduced for separable nonlinear least-squares problems. A representative later application is the training of a two-layer neural network with nonlinear hidden parameters θ, feature matrix $\Phi_\theta(X)$ and linear output weights A:

$$\min_{\theta,A} \frac{1}{2}\|A\Phi_\theta(X) - Y\|_{\mathrm{F}}^2.$$

For fixed θ, optimizing A is a convex linear least-squares problem that can be solved by QR or SVD, or through its normal-equation linear system, and is therefore practical for moderate-size problems; one then optimizes the reduced, generally nonconvex, objective in θ [302]. The present use is structurally different. The Sinkhorn dual $\mathcal{D}_\varepsilon(f, g)$ is jointly concave in (f, g), or equivalently $-\mathcal{D}_\varepsilon$ is jointly convex, and the eliminated block does not require a least-squares solve: its exact maximizer is the closed-form soft transform $f = g^{\bar{c},\varepsilon}$.

For a potential g, let π_g be the Gibbs measure obtained from the density law (8.24) by replacing $(f_\varepsilon, g_\varepsilon)$ with $(g^{\bar{c},\varepsilon}, g)$. Write $\pi_g(\mathrm{d}x, \mathrm{d}y) = \alpha(\mathrm{d}x)\pi_{g,x}(\mathrm{d}y)$ and denote its second marginal by β_g. The first marginal is exactly α by construction.

Proposition 9.22 (Semi-dual curvature and conditioning)**.** *For bounded directions k, k',*

$$D\mathcal{E}_\varepsilon(g)[k] = \int_{\mathcal{Y}} k\mathrm{d}(\beta - \beta_g), \qquad D^2\mathcal{E}_\varepsilon(g)[k, k'] = -\frac{1}{\varepsilon}\int_{\mathcal{X}} \mathrm{Cov}_{\pi_{g,x}}\big(k(Y), k'(Y)\big)\mathrm{d}\alpha(x). \tag{9.21}$$

At $g = g_\varepsilon$, its positive curvature on $L^2_0(\beta)$ is

$$-\nabla^2\mathcal{E}_\varepsilon(g_\varepsilon) = \frac{1}{\varepsilon}(\mathrm{Id} - T_\varepsilon^* T_\varepsilon), \qquad \frac{1 - \sigma_\varepsilon^2}{\varepsilon}\mathrm{Id} \preceq -\nabla^2\mathcal{E}_\varepsilon(g_\varepsilon) \preceq \frac{1}{\varepsilon}\mathrm{Id}. \tag{9.22}$$

Hence its spectral condition number is at most $1/(1-\sigma_\varepsilon^2)$. *Gradient ascent on the local quadratic model, with the minimax constant step based only on these spectral bounds,* $2\varepsilon/(2-\sigma_\varepsilon^2)$, *has factor*

$$r_{\mathrm{grad}} \leqslant \frac{\sigma_\varepsilon^2}{2-\sigma_\varepsilon^2}. \tag{9.23}$$

Proof. The envelope theorem gives the first derivative in (9.21). Differentiating the normalized conditional Gibbs law gives the covariance formula. At the optimum, total covariance yields

$$\int_{\mathcal{X}} \mathrm{Cov}_{\pi_{\varepsilon,x}}(k(Y), k'(Y))\mathrm{d}\alpha(x) = \int_{\mathcal{Y}} kk'\mathrm{d}\beta - \int_{\mathcal{X}} (T_\varepsilon k)(T_\varepsilon k')\mathrm{d}\alpha,$$

which is (9.22). Equivalently, this operator is the Schur complement of the first identity block in (9.13). Its spectrum on centered functions lies in $[(1-\sigma_\varepsilon^2)/\varepsilon, 1/\varepsilon]$. The standard optimal-step formula for a quadratic with spectral interval $[\alpha, L]$ gives factor $(L-\alpha)/(L+\alpha)$, which here is (9.23). □

Variable projection cannot worsen local spectral conditioning [244, 245]: the Schur-complement quadratic form is the minimum of the full quadratic form over the eliminated block, so its smallest eigenvalue cannot decrease and its largest eigenvalue cannot increase. Here the gain has a simple explicit bound. The full-dual condition number from Proposition 9.19 is $2/(1-\sigma_\varepsilon)$, whereas the semi-dual condition number is at most $1/(1-\sigma_\varepsilon^2)$; this upper bound is smaller by the factor $2(1+\sigma_\varepsilon)$. This comparison is available because both diagonal blocks in (9.13) are identities.

There is also a useful algorithmic comparison. Writing $r_g := \mathrm{d}\beta_g/\mathrm{d}\beta$ and using the two soft transforms gives

$$r_g = \exp\left(\frac{g-(g^{\bar{c},\varepsilon})^{c,\varepsilon}}{\varepsilon}\right), \qquad \nabla\mathcal{E}_\varepsilon(g) = 1 - r_g. \tag{9.24}$$

Here the gradient is the Riesz gradient for the $L^2(\beta)$ inner product. Therefore one ordinary Sinkhorn cycle satisfies

$$(g^{\bar{c},\varepsilon})^{c,\varepsilon} - g = -\varepsilon \log r_g = \varepsilon\nabla\mathcal{E}_\varepsilon(g) + o\left(\|g-g_\varepsilon\|_{L^2(\beta)}\right).$$

Sinkhorn is thus locally semi-dual gradient ascent with step ε, although the two nonlinear algorithms are not globally identical. Quasi-Newton methods instead learn an approximation of $\varepsilon(\mathrm{Id}-T_\varepsilon^* T_\varepsilon)^{-1}$ from successive gradients. This explains the effectiveness of BFGS and L-BFGS on regularized semi-duals [170, 171, 404]: they reuse curvature information, at the cost of storing curvature pairs and, typically, extra evaluations during line search. Stochastic first-order methods instead exploit the same semi-dual gradient formula using sampled marginals [238].

9.6 Entropic Optimal Transport between Gaussians

Gaussian marginals provide an explicit finite-dimensional model of Sinkhorn's behavior. The soft c-transform preserves quadratic potentials, the optimal entropic coupling is Gaussian, and the value can be written with matrix square roots [291]. This is the entropic counterpart of the Gaussian $\mathcal{W}_2$ formula in Proposition 2.54.

Proposition 9.23 (Quadratic closure of Sinkhorn iterates)**.** *Let* $\beta = \mathcal{N}(m_\beta, \Sigma_\beta)$ *on* $\mathbb{R}^d$ *and take* $c(x,y) = \|x-y\|^2$. *If* $g(y)$ *is a quadratic polynomial such that the Gaussian integral below is finite, then the soft transform*

$$f(x) = -\varepsilon \log \int \exp\left(\frac{g(y)-\|x-y\|^2}{\varepsilon}\right)\mathrm{d}\beta(y)$$

is a quadratic polynomial in x. *In particular, starting Sinkhorn from* $g_0 = 0$ *gives*

$$f_1(x) = \frac{\varepsilon}{2}\log\det\left(\mathrm{Id} + \frac{2\Sigma_\beta}{\varepsilon}\right) + \varepsilon\langle x-m_\beta, (\varepsilon\mathrm{Id}+2\Sigma_\beta)^{-1}(x-m_\beta)\rangle.$$

Proof. The exponent is the sum of a quadratic polynomial in y and the logarithm of the Gaussian density of β. Completing the square in y evaluates the integral as a positive constant times the exponential of a quadratic polynomial in x. Taking $-\varepsilon$ log therefore gives a quadratic polynomial.

For $g_0 = 0$, let $Y \sim \beta$. The Gaussian identity

$$\mathbb{E}\exp\left(-\frac{\|x-Y\|^2}{\varepsilon}\right) = \det\left(\mathrm{Id} + \frac{2\Sigma_\beta}{\varepsilon}\right)^{-1/2} \exp\left(-\langle x - m_\beta,\, (\varepsilon\mathrm{Id} + 2\Sigma_\beta)^{-1}(x - m_\beta)\rangle\right)$$

gives the displayed expression. □

Proposition 9.24 (Balanced entropic OT between Gaussians)**.** *Let $\alpha = \mathcal{N}(m_\alpha, \Sigma_\alpha)$ and $\beta = \mathcal{N}(m_\beta, \Sigma_\beta)$ with positive-definite covariances, and let*

$$\Sigma_\alpha^{1/2}\Sigma_\beta^{1/2} = U\,\mathrm{diag}(\sigma_i)V^\top$$

be a singular-value decomposition. For the balanced objective

$$\min_{\pi\in\mathcal{U}(\alpha,\beta)} \int \|x-y\|^2 \mathrm{d}\pi(x,y) + \varepsilon\,\mathrm{KL}(\pi|\alpha\otimes\beta),$$

the optimizer is Gaussian with cross-covariance

$$K_\varepsilon = \Sigma_\alpha^{1/2} U\,\mathrm{diag}(s_i)V^\top\Sigma_\beta^{1/2}, \qquad s_i = \frac{\sqrt{\varepsilon^2 + 16\sigma_i^2} - \varepsilon}{4\sigma_i}.$$

The optimal value is

$$\|m_\alpha - m_\beta\|^2 + \mathrm{tr}(\Sigma_\alpha) + \mathrm{tr}(\Sigma_\beta) + \sum_i \left(-2\sigma_i s_i - \frac{\varepsilon}{2}\log(1-s_i^2)\right).$$

As $\varepsilon \downarrow 0$, $s_i \to 1$ and the full covariance contribution, including the two trace terms, converges to $\mathcal{B}(\Sigma_\alpha, \Sigma_\beta)^2$.

Proof. Let (X, Y) be any coupling with finite second moments and cross-covariance $K = \mathbb{E}\big[(X - m_\alpha)(Y - m_\beta)^\top\big]$. Replacing (X, Y) by the Gaussian vector with the same mean and covariance leaves the quadratic cost unchanged. Since the marginals are fixed,

$$\mathrm{KL}(\pi|\alpha\otimes\beta) = -h(X,Y) + h(\alpha) + h(\beta),$$

where h denotes differential entropy when it is finite. Among laws with a fixed covariance, the Gaussian maximizes entropy; if the entropy is not finite, the relative entropy is already $+\infty$. Thus the Gaussian replacement cannot increase the objective, and it is enough to optimize over Gaussian couplings.

Any such coupling has covariance

$$\begin{pmatrix} \Sigma_\alpha & K \\ K^\top & \Sigma_\beta \end{pmatrix}.$$

Write $K = \Sigma_\alpha^{1/2} S \Sigma_\beta^{1/2}$. The block covariance constraint is equivalent to the singular values of S being at most one, and finite entropy forces them to be strictly smaller than one. The cost depends on K through

$$\|m_\alpha - m_\beta\|^2 + \mathrm{tr}(\Sigma_\alpha) + \mathrm{tr}(\Sigma_\beta) - 2\,\mathrm{tr}(K),$$

while

$$\mathrm{KL}(\pi|\alpha\otimes\beta) = -\frac{1}{2}\log\det(\mathrm{Id} - SS^\top).$$

By von Neumann's trace inequality, the minimizer aligns S with the singular vectors of $\Sigma_\alpha^{1/2}\Sigma_\beta^{1/2}$, so $S = U\,\mathrm{diag}(s_i)V^\top$. The problem separates into scalar minimizations

$$\min_{0\leqslant s<1} -2\sigma_i s - \frac{\varepsilon}{2}\log(1-s^2).$$

The first-order condition is $2\sigma_i = \varepsilon s/(1-s^2)$, whose positive solution is the displayed s_i. Substitution gives the value formula. Since $s_i \to 1$ and $\varepsilon\log(1-s_i^2) \to 0$ as $\varepsilon \downarrow 0$, the spectral sum converges to $-2\sum_i \sigma_i$. The identity

$$\sum_i \sigma_i = \mathrm{tr}\left((\Sigma_\alpha^{1/2}\Sigma_\beta\Sigma_\alpha^{1/2})^{1/2}\right)$$

then gives the Bures–Wasserstein covariance contribution. □

Corollary 9.25 (Gaussian Sinkhorn divergence and smoothed Bures term)**.** *Let $\alpha = \mathcal{N}(m_\alpha, \Sigma_\alpha)$ and $\beta = \mathcal{N}(m_\beta, \Sigma_\beta)$ have positive-definite covariances. For $r > 0$, define*

$$\tau_\varepsilon(r) := \frac{\sqrt{\varepsilon^2 + 16r^2} - \varepsilon}{4r}, \qquad \psi_\varepsilon(r) := -2r\,\tau_\varepsilon(r) - \frac{\varepsilon}{2}\log\bigl(1 - \tau_\varepsilon(r)^2\bigr).$$

If $\sigma_i(\Sigma, \Lambda)$ denotes the singular values of $\Sigma^{1/2}\Lambda^{1/2}$ and $\lambda_i(\Sigma)$ the eigenvalues of Σ, then the debiased Sinkhorn divergence (8.57) is

$$\bar{\mathcal{L}}^{\varepsilon}_{\|\cdot-\cdot\|^2}(\alpha, \beta) = \|m_\alpha - m_\beta\|^2 + \mathcal{B}_\varepsilon(\Sigma_\alpha, \Sigma_\beta)^2,$$

where the Gaussian covariance contribution is the debiased smoothed Bures term

$$\mathcal{B}_\varepsilon(\Sigma, \Lambda)^2 := \sum_i \psi_\varepsilon\bigl(\sigma_i(\Sigma, \Lambda)\bigr) - \frac{1}{2}\sum_i \psi_\varepsilon\bigl(\lambda_i(\Sigma)\bigr) - \frac{1}{2}\sum_i \psi_\varepsilon\bigl(\lambda_i(\Lambda)\bigr).$$

Moreover $\mathcal{B}_\varepsilon(\Sigma, \Lambda)^2 \to \mathcal{B}(\Sigma, \Lambda)^2$ as $\varepsilon \downarrow 0$, where $\mathcal{B}$ is the Bures–Wasserstein metric of Proposition 2.54.

Proof. Proposition 9.24 writes the raw entropic value as the squared mean displacement plus trace terms and a spectral sum. With the notation above, the spectral part is exactly $\sum_i \psi_\varepsilon(\sigma_i(\Sigma_\alpha, \Sigma_\beta))$. Applying the same formula to the self-costs (α, α) and (β, β) replaces these singular values by the eigenvalues of Σ_α and Σ_β. In the polarization formula (8.57), the trace terms cancel:

$$\operatorname{tr}\Sigma_\alpha + \operatorname{tr}\Sigma_\beta - \frac{1}{2}(2\operatorname{tr}\Sigma_\alpha) - \frac{1}{2}(2\operatorname{tr}\Sigma_\beta) = 0,$$

while the polarization of the squared mean terms leaves exactly $\|m_\alpha - m_\beta\|^2$. This gives the displayed formula. Since $\tau_\varepsilon(r) \to 1$ and $\varepsilon\log(1 - \tau_\varepsilon(r)^2) \to 0$, one has $\psi_\varepsilon(r) \to -2r$. The limit is therefore

$$\operatorname{tr}\Sigma + \operatorname{tr}\Lambda - 2\sum_i \sigma_i(\Sigma, \Lambda) = \mathcal{B}(\Sigma, \Lambda)^2,$$

which is the Bures formula (2.17). □

Proposition 9.26 (One-dimensional Gaussian Sinkhorn rate)**.** *Consider $\alpha = \beta = \mathcal{N}(0, 1)$ on $\mathbb{R}$ with $c(x, y) = (x - y)^2$. If a dual potential has the form $g_q(y) = qy^2 + \mathrm{cst}$, then one soft transform has quadratic coefficient*

$$\mathsf{Q}_\varepsilon(q) = 1 - \frac{1}{1 - q + \varepsilon/2}, \qquad q < 1 + \varepsilon/2,$$

and one full Sinkhorn cycle acts as $q \mapsto \mathsf{Q}_\varepsilon(\mathsf{Q}_\varepsilon(q))$. The fixed point $q_\star = \mathsf{Q}_\varepsilon(q_\star)$ is determined by

$$A_\star^2 - \frac{\varepsilon}{2}A_\star - 1 = 0, \qquad A_\star := 1 - q_\star + \frac{\varepsilon}{2} = \frac{\varepsilon + \sqrt{\varepsilon^2 + 16}}{4}.$$

The optimal Gaussian coupling of Proposition 9.24 has correlation $s_\varepsilon = A_\star^{-1}$. Hence the maximal correlation in Definition 9.18 is $\sigma_\varepsilon = s_\varepsilon$, and the exact local L^2 factor is $A_\star^{-2}$. The quadratic-potential mode contracts faster, with factor

$$\rho_\varepsilon = A_\star^{-4} = \left(\frac{4}{\varepsilon + \sqrt{\varepsilon^2 + 16}}\right)^4.$$

Proof. Completing the square in

$$\int \exp\left(\frac{qy^2 - (x - y)^2}{\varepsilon}\right) \mathrm{d}\mathcal{N}(0, 1)(y)$$

gives the coefficient $\mathsf{Q}_\varepsilon(q)$. The fixed-point equation $q_\star = 1 - 1/A_\star$, together with $q_\star = 1 + \varepsilon/2 - A_\star$, gives $A_\star^2 - \frac{\varepsilon}{2}A_\star - 1 = 0$. The positive solution is the displayed $A_\star$. Since $\mathsf{Q}_\varepsilon'(q) = -\frac{1}{(1-q+\varepsilon/2)^2}$, the derivative of the full-cycle map at the fixed point is $\mathsf{Q}_\varepsilon'(q_\star)^2 = A_\star^{-4}$. The correlation formula is the one-dimensional specialization of Proposition 9.24. Conditional expectation under a correlated Gaussian diagonalizes in the Hermite basis, with largest nonconstant singular value s_ε; Proposition 9.19 then gives the full factor $s_\varepsilon^2 = A_\star^{-2}$. □

This calculation illustrates the general Gaussian convergence picture of Chizat, Delalande and Vaškevičius [139]: the rate improves when ε is large or the covariance scales overlap well, and deteriorates in the small-temperature limit where the entropic coupling approaches a deterministic Brenier map.

9.7 Continuous ε-Sinkhorn Flow

This section studies a simultaneous high-resolution, many-iteration limit of Sinkhorn. It is not the continuous-time interpolation of a fixed-temperature algorithm: the grid is refined while the temperature and the fictitious time step both vanish as $1/k$.

Parabolic Monge–Ampere limit. The previous arguments view Sinkhorn as a discrete fixed-point iteration. A complementary picture, developed by Berman [46], considers the quadratic torus cost $c(x,y) = d_{\mathbb{T}^d}(x,y)^2/2$, discretizes both marginals on a grid of mesh $1/k$, sets the entropic temperature to $\varepsilon_k = 1/k$, and assigns the same duration $\Delta t = 1/k$ to each Sinkhorn update. The ℓ-th log-potential is therefore observed at time $t = \ell/k$. In this coupled limit, the iterates converge to a fully nonlinear parabolic flow whose stationary equation is the Monge–Ampere equation for the optimal map. This connects matrix scaling with real and complex Monge–Ampere geometry.

Definition 9.27 (Continuous ε-Sinkhorn flow)**.** Let α and β be probability measures on the unit-volume flat torus $\mathbb{T}^d$ with smooth positive densities $\rho_\alpha = e^{-F}$ and $\rho_\beta = e^{-G}$. For the cost $c(x,y) = d_{\mathbb{T}^d}(x,y)^2/2$, a gauge-fixed potential $u_t : \mathbb{T}^d \to \mathbb{R}$, with $\int_{\mathbb{T}^d} u_t \mathrm{d}x = 0$, follows the continuous ε-Sinkhorn flow if

$$\partial_t u_t(x) = \log\det\big(\mathrm{Id} + \nabla^2 u_t(x)\big) - G\big(x + \nabla u_t(x)\big) + F(x) - \bar{r}_t, \tag{CES}$$

where $x + \nabla u_t(x)$ is understood modulo $\mathbb{Z}^d$, $\mathrm{Id} + \nabla^2 u_t \succ 0$, and $\bar{r}_t$ is the unique scalar preserving the mean-zero gauge.

Explicitly,

$$\bar{r}_t = \int_{\mathbb{T}^d} \Big[\log\det\big(\mathrm{Id} + \nabla^2 u_t(x)\big) - G\big(x + \nabla u_t(x)\big) + F(x)\Big]\mathrm{d}x.$$

Its subtraction fixes the additive constant of the potential. The following operators make the discrete scaling behind this equation explicit. Let $\alpha^{[k]}$ and $\beta^{[k]}$ be the positive grid discretizations of α and β, and define

$$v_k[u](y) := \frac{1}{k}\log\int_{\mathbb{T}^d} e^{-k(c(x,y)+u(x))}\mathrm{d}\alpha^{[k]}(x),$$

followed by the double log-Sinkhorn operator

$$(S_k u)(x) := \frac{1}{k}\log\int_{\mathbb{T}^d} e^{-k(c(x,y)+v_k[u](y))}\mathrm{d}\beta^{[k]}(y).$$

Its multiplicative increment is

$$\rho_{k,u}(x) := e^{k(S_k u(x) - u(x))} = e^{-ku(x)}\int_{\mathbb{T}^d} \frac{e^{-kc(x,y)}}{\int_{\mathbb{T}^d} e^{-k(c(x',y)+u(x'))}\mathrm{d}\alpha^{[k]}(x')}\mathrm{d}\beta^{[k]}(y). \tag{9.25}$$

The normalized update is $u^{[k],(\ell+1)} = S_k u^{[k],(\ell)} - \int_{\mathbb{T}^d} S_k u^{[k],(\ell)}\mathrm{d}x$.

Proposition 9.28 (Scaled log-Sinkhorn limit)**.** *Let $\varepsilon_k = 1/k$, let $\alpha^{[k]}, \beta^{[k]}$ be the regular-grid density discretizations above, and generate the mean-zero iterates with the normalized operator S_k. Define $u^{[k]}(t) := u^{[k],(\lfloor kt \rfloor)}$. Assume that, on every compact time interval, $u^{[k]}(t)$ converges smoothly to u_t, that $\mathrm{Id} + \nabla^2 u_t \succ 0$, and that the Laplace expansion below is uniform along the sequence. Then u_t solves the continuous ε-Sinkhorn flow* (CES).

Proof. By (9.25), the normalized update satisfies

$$u^{[k],(\ell+1)} - u^{[k],(\ell)} = k^{-1}\left[\log\rho_{k,u^{[k],(\ell)}} - \int_{\mathbb{T}^d}\log\rho_{k,u^{[k],(\ell)}}\mathrm{d}x\right],$$

where the subtraction enforces the mean-zero gauge. Berman's discrete Laplace estimate gives

$$\rho_{k,u}(x) = \det(\mathrm{Id} + \nabla^2 u(x))e^{F(x)-G(x+\nabla u(x))}\big(1 + O(k^{-1})\big),$$

uniformly for the smooth potentials under consideration. Taking logarithms gives

$$\log \rho_{k,u}(x) = \log\det(\mathrm{Id} + \nabla^2 u(x)) - G(x + \nabla u(x)) + F(x) + O(k^{-1}).$$

Dividing the increment by the time step $1/k$, passing to the smooth limit $u^{[k]}(t) \to u_t$, and keeping the mean-zero normalization gives exactly (CES). □

The word "epsilon" therefore refers to the vanishing entropic temperature before this rescaling; no explicit ε remains in the normalized limiting PDE. The equation is a real parabolic Monge–Ampere equation. Its Kähler analogue is the parabolic complex Monge–Ampere equation, which explains the link with the geometric Monge–Ampere literature emphasized by Berman.

Proposition 9.29 (Stationary continuous Sinkhorn potentials)**.** *Assume that a smooth stationary solution of* (CES) *satisfies* $\mathrm{Id} + \nabla^2 u > 0$*, and define* $T(x) = x + \nabla u(x)$ *modulo* $\mathbb{Z}^d$*. Then* $T_\sharp \alpha = \beta$*. Conversely, any smooth map of this form satisfying* $\mathrm{Id} + \nabla^2 u > 0$ *and* $T_\sharp \alpha = \beta$ *solves the stationary equation, up to the additive gauge of* u*.*

Proof. At stationarity, the bracket in (CES) is a constant c:

$$\log\det\big(\mathrm{Id} + \nabla^2 u(x)\big) - G(T(x)) + F(x) = c.$$

Exponentiating, and using $\nabla T = \mathrm{Id} + \nabla^2 u$, gives

$$e^{-G(T(x))}\det(\nabla T(x)) = e^{-F(x)}e^{c}.$$

Because T is homotopic to the identity and $\mathrm{Id} + \nabla^2 u > 0$, it is an orientation-preserving local diffeomorphism of degree one; hence it is a diffeomorphism of the torus. Since α and β have total mass one, integration and the change-of-variables formula give $e^c = 1$. The same identity then becomes precisely the Jacobian equation for $T_\sharp(e^{-F}\mathrm{d}x) = e^{-G}\mathrm{d}x$. The converse follows by reversing this argument. □

Gaussian closure. Gaussian marginals give a finite-dimensional test case for the continuous flow. The theorem above is stated on the flat torus, but the same local Laplace calculation can be read formally on $\mathbb{R}^d$ for confining Gaussian densities. Write $\alpha = \mathcal{N}(m_\alpha, \Sigma_\alpha)$ and $\beta = \mathcal{N}(m_\beta, \Sigma_\beta)$, and restrict the potential to the quadratic ansatz for which

$$T_t(x) := x + \nabla u_t(x) = q_t + B_t(x - m_\alpha), \qquad B_t \in \mathbb{S}^d_{++}.$$

Taking the spatial gradient of (CES) removes the additive gauge. Since

$$F(x) = \frac{1}{2}\langle x - m_\alpha,\, \Sigma_\alpha^{-1}(x - m_\alpha)\rangle + \mathrm{cst}, \qquad G(y) = \frac{1}{2}\langle y - m_\beta,\, \Sigma_\beta^{-1}(y - m_\beta)\rangle + \mathrm{cst},$$

coefficient matching in the identity $\partial_t T_t = \nabla \partial_t u_t$ gives

$$\dot B_t = \Sigma_\alpha^{-1} - B_t\Sigma_\beta^{-1}B_t, \qquad \dot q_t = -B_t\Sigma_\beta^{-1}(q_t - m_\beta).$$

Thus the parabolic Monge–Ampere equation reduces, on the Gaussian ansatz, to a Riccati evolution for the linear part of the transport. The image mean is q_t, and the image covariance is

$$\Sigma_t = B_t\Sigma_\alpha B_t, \qquad \dot\Sigma_t = \dot B_t\Sigma_\alpha B_t + B_t\Sigma_\alpha \dot B_t.$$

At equilibrium,

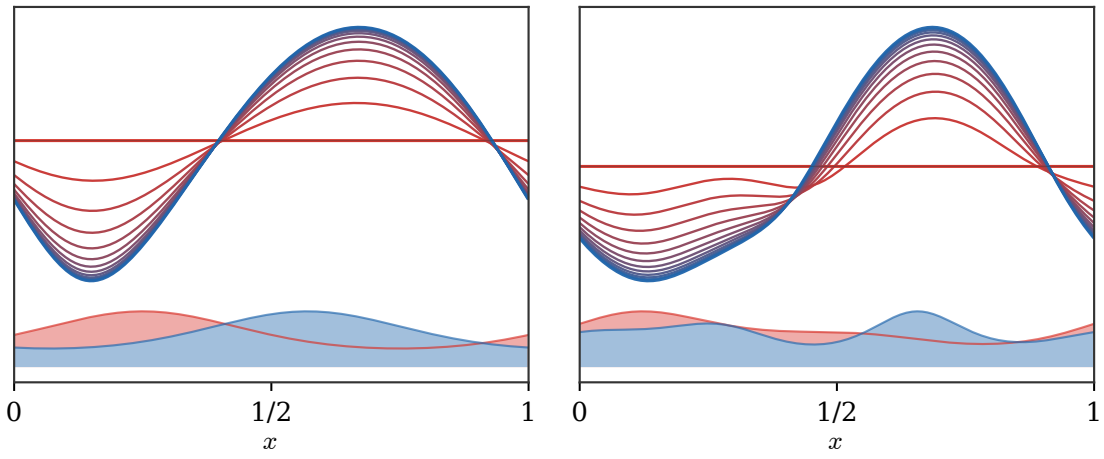
$$B\Sigma_\beta^{-1}B = \Sigma_\alpha^{-1}, \qquad q = m_\beta,$$

which is equivalent to $B\Sigma_\alpha B = \Sigma_\beta$. The stationary map is therefore the Gaussian Brenier map, and the endpoint covariance is governed by the same Bures–Wasserstein geometry as in Proposition 2.54 and Section 9.6. This Gaussian reduction should be viewed as the finite-dimensional covariance shadow of the vanishing-temperature continuous Sinkhorn limit, not as the fixed-temperature Gaussian Sinkhorn formula itself.

In one dimension, the flow reduces to

$$\partial_t u_t(x) = \log\big(1 + u_t''(x)\big) - G\big(x + u_t'(x)\big) + F(x) - \bar r_t,$$

as long as $1 + u_t'' > 0$. This scalar case is useful for visualization because the potential curves can be plotted directly and the positivity condition is exactly the monotonicity of $x \mapsto x + u_t'(x)$. Figure 9.4 shows this evolution from the zero initialization for two smooth pairs of marginals.



translated smooth bump oscillatory split density

Figure 9.4: Continuous ε-Sinkhorn flow in one dimension. The curves are snapshots of the gauge-fixed potential u_t, initialized at $u_0 = 0$, under the parabolic Monge–Ampere equation (CES) obtained from Berman's high-resolution, vanishing-temperature Sinkhorn scaling. Time is encoded from red to blue. The faint bottom silhouettes show the source density in red and the target density in blue. The examples are deliberately smooth so that the finite-difference evolution remains inside the monotone regime $1 + u_t'' > 0$.

9.8 Monotone Clearing Beyond Variational Sinkhorn

The preceding sections mostly used variational structure: Sinkhorn is alternating KL projection, coordinate ascent on a dual objective, or a contraction in a projective metric. There is another, more algebraic, convergence mechanism which keeps the scaling form but discards the existence of an objective. In Galichon's equilibrium-flow viewpoint, and in related work on substitutability and inverse isotonicity, one studies nonlinear market-clearing equations whose Jacobian has a substitute sign structure [227, 229]. This is the nonlinear analogue of the classical theory of nonsingular M-matrices and M-functions [390, 437]. The relevance for OT is that Sinkhorn is the canonical scaling example, but the same monotone-clearing proof also covers fixed-point equations that are not first-order conditions of any convex regularized transport problem.

Two-block clearing maps. Write the unknowns as two blocks $z = (u, v) \in \mathbb{R}^n \times \mathbb{R}^m$, where u and v will later be signed log-scalings. Let

$$Q(z) = \big(Q^\alpha(u, v), Q^\beta(u, v)\big) \in \mathbb{R}^n \times \mathbb{R}^m$$

be a continuous clearing map. The parallel coordinate-clearing map T is defined by solving, for each coordinate,

$$Q_\ell\big(T_\ell(z), z_{-\ell}\big) = 0, \qquad 1 \leqslant \ell \leqslant n + m.$$

This is the Jacobi version: all coordinates are cleared against the old values of the other coordinates. In two-block scaling problems, all coordinates in u decouple when v is fixed, and all coordinates in v decouple when u is fixed. The more common alternating Sinkhorn sweep is the Gauss–Seidel composition of these two block clearings; the order argument below is stated for the parallel map to keep the notation short. The same construction extends to any finite number of blocks.

Definition 9.30 (Z-functions and M-functions)**.** Let $D \subset \mathbb{R}^N$ be an order interval. A map $Q : D \to \mathbb{R}^N$ is diagonally isotone if $Q_\ell(z_\ell, z_{-\ell})$ is nondecreasing in z_ℓ for every ℓ. It is a Z-function if increasing the other coordinates cannot increase the ℓ-th component:

$$z_{-\ell} \leqslant z'_{-\ell} \quad \Longrightarrow \quad Q_\ell(z_\ell, z_{-\ell}) \geqslant Q_\ell(z_\ell, z'_{-\ell}).$$

For a C^1 map this means $\partial Q_\ell / \partial z_k \leqslant 0$ for $k \neq \ell$. An M-function is a Z-function which is inverse

isotone:

$$Q(z) \leqslant Q(z') \quad \Longrightarrow \quad z \leqslant z'.$$

A vector z is a subsolution if $Q(z) \leqslant 0$, and a supersolution if $Q(z) \geqslant 0$.

The M-function assumption says that cross-effects have the sign of substitutes, and that own effects dominate them strongly enough to prevent a global reversal of order. Galichon, Samuelson and Vernet formulate this idea through nonreversingness and unified gross substitutes; for single-valued maps this is the inverse-isotone structure used below. A degenerate M_0-function keeps the same order structure but allows a null gauge direction, which is exactly what happens for balanced Sinkhorn before a potential normalization is imposed.

Theorem 9.31 (Monotone convergence of coordinate clearing)**.** *Let $D = [\underline{z}, \overline{z}] \subset \mathbb{R}^N$ be a closed order interval, with $\underline{z}$ a subsolution and $\overline{z}$ a supersolution. Assume that $Q : D \to \mathbb{R}^N$ is continuous, diagonally isotone and an M-function. Assume also that, for every $z \in D$ and every ℓ, the scalar clearing equation $Q_\ell(\xi, z_{-\ell}) = 0$ has a unique solution $\xi \in [\underline{z}_\ell, \overline{z}_\ell]$. Then $Q(z) = 0$ has a unique solution $z^\star \in D$. The Jacobi coordinate-clearing iterates starting from any subsolution in D increase monotonically to $z^\star$, while those starting from any supersolution in D decrease monotonically to $z^\star$.*

Proof. Let T be the coordinate-clearing map. If $z \in D$ is a subsolution, then $Q_\ell(z_\ell, z_{-\ell}) \leqslant 0 = Q_\ell(T_\ell(z), z_{-\ell})$. Diagonal isotonicity and uniqueness of the scalar zero give $z_\ell \leqslant T_\ell(z)$ for every ℓ, hence $z \leqslant T(z)$. Since Q is a Z-function,

$$Q_\ell(T(z)) \leqslant Q_\ell(T_\ell(z), z_{-\ell}) = 0,$$

so $T(z)$ is again a subsolution. Since every subsolution lies below every supersolution by inverse isotonicity, $T(z) \leqslant \overline{z}$, and therefore $T(z) \in D$. The supersolution argument is the same with all inequalities reversed. The lower and upper iterates are thus monotone and trapped in the compact interval D. Their limits exist, and passing to the limit in the identities $Q_\ell(z_\ell^{(k+1)}, z_{-\ell}^{(k)}) = 0$ gives $Q(z^\star) = 0$. If z and z' were two zeros in D, then $Q(z) \leqslant Q(z')$ and $Q(z') \leqslant Q(z)$; inverse isotonicity gives $z = z'$. □

Proposition 9.32 (Smooth M-matrix certificate)**.** *Let $D \subset \mathbb{R}^N$ be a convex order interval and let $Q : D \to \mathbb{R}^N$ be C^1. Assume that $DQ(z)$ has nonpositive off-diagonal entries for every $z \in D$. If there exists $\lambda \gg 0$ such that*

$$\lambda^\top DQ(z) \gg 0 \qquad \text{for all } z \in D,$$

or equivalently

$$\lambda_\ell \frac{\partial Q_\ell}{\partial z_\ell}(z) > \sum_{k \neq \ell} \lambda_k \left| \frac{\partial Q_k}{\partial z_\ell}(z) \right|, \qquad 1 \leqslant \ell \leqslant N,$$

then Q is an M-function on D.

Proof. The off-diagonal sign gives the Z-property by integrating the partial derivatives along coordinate-wise increasing segments. The displayed inequalities imply that each $DQ(z)$ is a strictly weighted column diagonally dominant Z-matrix, hence a nonsingular M-matrix; in particular $DQ(z)^{-1} \geqslant 0$ for each fixed z [437]. For two points $z, z' \in D$, set $A = \int_0^1 DQ\big(z' + t(z - z')\big)\,\mathrm{d}t$. The matrix A has the same Z-sign pattern and the same strict weighted diagonal dominance, so it is again a nonsingular M-matrix and $A^{-1} \geqslant 0$. Since $Q(z) - Q(z') = A(z - z')$, the implication $Q(z) \leqslant Q(z')$ gives $z - z' = A^{-1}(Q(z) - Q(z')) \leqslant 0$. This is inverse isotonicity. The positive diagonal entries also give diagonal isotonicity. □

Sinkhorn as the canonical M_0-system. Let $\mathrm{K} = \exp(-\mathrm{C}/\varepsilon) > 0$, and write the entropic plan as

$$\mathrm{P}_{i,j} = r_i \mathrm{K}_{i,j} s_j, \qquad r_i = e^{u_i}, \qquad s_j = e^{-v_j}.$$

The signed convention $v = -\log s$ makes the clearing map

$$Q_i^\alpha(u, v) = \sum_j \mathrm{K}_{i,j} e^{u_i - v_j} - a_i, \qquad Q_j^\beta(u, v) = b_j - \sum_i \mathrm{K}_{i,j} e^{u_i - v_j}.$$

The equations $Q = 0$ are the row and column constraints $\mathrm{P}\mathbb{1} = \mathrm{a}$ and $\mathrm{P}^\top \mathbb{1} = \mathrm{b}$, and the coordinate-clearing equations give the usual Sinkhorn scalings

$$r_i^+ = \frac{a_i}{\sum_j \mathrm{K}_{i,j} s_j}, \qquad s_j^+ = \frac{b_j}{\sum_i \mathrm{K}_{i,j} r_i}.$$

The Jacobian has positive diagonal terms and nonpositive off-diagonal terms:

$$\frac{\partial Q_i^\alpha}{\partial u_i} = \sum_j P_{i,j}, \qquad \frac{\partial Q_i^\alpha}{\partial v_j} = -\mathrm{P}_{i,j}, \qquad \frac{\partial Q_j^\beta}{\partial v_j} = \sum_i P_{i,j}, \qquad \frac{\partial Q_j^\beta}{\partial u_i} = -\mathrm{P}_{i,j}.$$

Its column sums vanish, which reflects the gauge invariance $(u, v) \mapsto (u + c\mathbb{1}, v + c\mathbb{1})$. Thus balanced Sinkhorn is naturally an M_0-system. Fixing one log-scaling coordinate removes this null direction and turns the reduced Jacobian into a principal minor of the weighted bipartite graph Laplacian. Under the connected-support condition, which is automatic here because $\mathrm{K} > 0$, this minor is a nonsingular M-matrix, so the M-function convergence mechanism applies on the gauge-fixed variables. This order-theoretic reading complements the variational and Hilbert-metric proofs above, and also appears in the link between Sinkhorn scaling and choice models [442].

Example 9.33 (Lossy Sinkhorn clearing with outside options). The following OT-shaped clearing model is deliberately minimal. It keeps a positive transport kernel and multiplicative scalings, but introduces outside options and lossy arrivals. It should be read as a toy absorptive matching model rather than as a new transport distance. Let $0 < \eta_{i,j} \leqslant 1$, $\sigma_i > 0$, $\tau_j > 0$, and set

$$\mathrm{P}_{i,j} = r_i \mathrm{K}_{i,j} s_j, \qquad r_i = e^{u_i}, \qquad s_j = e^{-v_j}.$$

The clearing equations are

$$\sigma_i r_i + \sum_j P_{i,j} = a_i, \qquad \tau_j s_j + \sum_i \eta_{i,j} \mathrm{P}_{i,j} = b_j.$$

They say that source mass can exit through an outside sink, while target demand can be met either by effective arrivals or by a local outside source. The coordinate-clearing update is still Sinkhorn-like:

$$r_i^+ = \frac{a_i}{\sigma_i + \sum_j \mathrm{K}_{i,j} s_j}, \qquad s_j^+ = \frac{b_j}{\tau_j + \sum_i \eta_{i,j} \mathrm{K}_{i,j} r_i}.$$

In the variables (u, v), the off-diagonal derivatives again have the Z-sign. On the invariant domain $r_i \leqslant a_i/\sigma_i$, the simple sufficient condition

$$\tau_j > \sum_i (1 - \eta_{i,j}) \mathrm{K}_{i,j} \frac{a_i}{\sigma_i}, \qquad 1 \leqslant j \leqslant m,$$

implies the weighted diagonal-dominance certificate $\mathbb{1}^\top DQ \gg 0$, hence Proposition 9.32 shows that the system is an M-function on that domain. Indeed, for the clearing map $Q_i^\alpha = \sigma_i r_i + \sum_j P_{i,j} - a_i$ and $Q_j^\beta = b_j - \tau_j s_j - \sum_i \eta_{i,j} \mathrm{P}_{i,j}$, the column sums of the Jacobian are

$$\sigma_i r_i + \sum_j (1 - \eta_{i,j}) \mathrm{P}_{i,j} > 0, \qquad \tau_j s_j - \sum_i (1 - \eta_{i,j}) \mathrm{P}_{i,j} = s_j \left(\tau_j - \sum_i (1 - \eta_{i,j}) \mathrm{K}_{i,j} r_i \right) > 0.$$

The model is generally non-variational. If $Q = \nabla \mathcal{E}$ for a C^2 potential, the cross-partials would satisfy

$$\frac{\partial Q_i^\alpha}{\partial v_j} = \frac{\partial Q_j^\beta}{\partial u_i}.$$

Here this would force $-\mathrm{P}_{i,j} = -\eta_{i,j} \mathrm{P}_{i,j}$ on every active arc, hence $\eta_{i,j} = 1$. Lossy arrivals therefore destroy exactness of the one-form $\sum_\ell Q_\ell \, \mathrm{d}z_\ell$, while preserving the monotone clearing structure.

Figure 9.5 shows this mechanism on two empirical Gaussian mixtures in $\mathbb{R}^2$. In the figure, the outside coefficients are written as $\sigma_i = \rho \bar{\sigma}_i$ and $\tau_j = \rho \bar{\tau}_j$, and the columns increase the common scale ρ. The visual point is not an optimization benchmark but the persistence of a well-behaved Sinkhorn-shaped fixed point outside the variational OT setting.

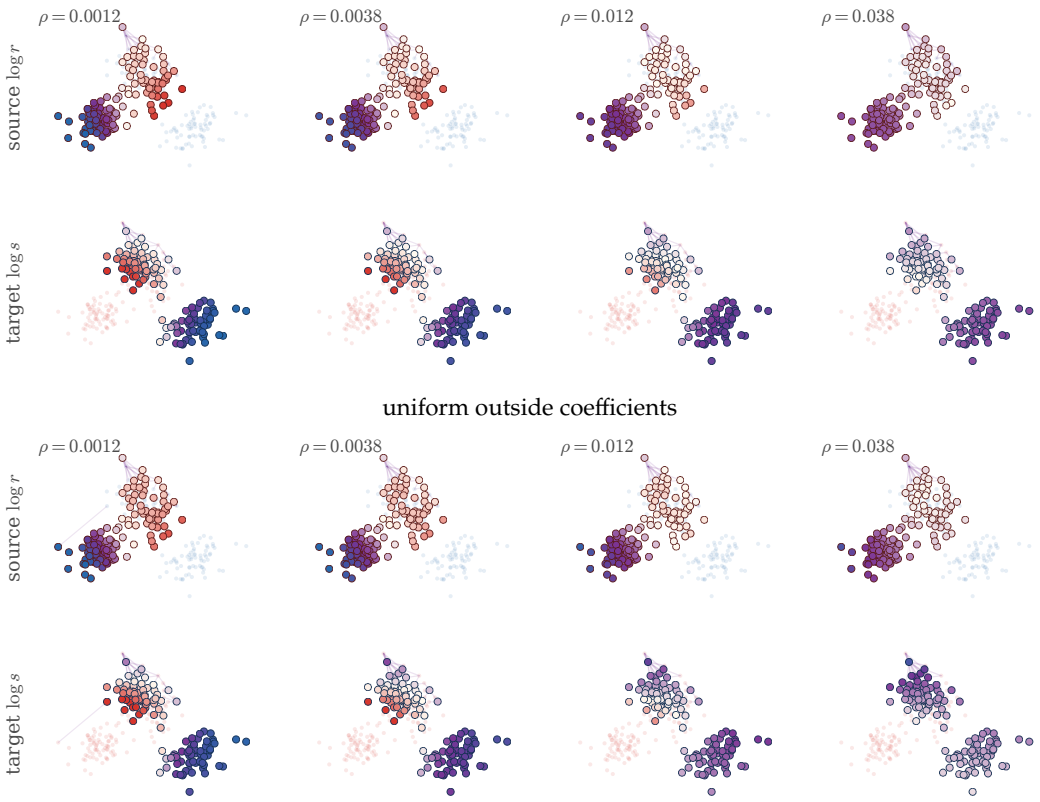


Figure 9.5: Non-variational lossy Sinkhorn scaling on two Gaussian-mixture point clouds. Across columns, the outside-option scale ρ increases in σ and τ. Colors show centered log-scalings, $\log r - \langle \log r \rangle$ on the source row and $\log s - \langle \log s \rangle$ on the target row; faint violet links mark the largest entries of the induced effective plan. The upper panel uses uniform outside coefficients, while the lower panel uses spatially varying outside coefficients and directional loss factors. The updates are Sinkhorn-like row and column clearings, but the loss factors $\eta_{i,j} \neq 1$ break the cross-partial symmetry required by a convex potential.

CHAPTER 10

Statistical Optimal Transport

Optimal transport is rarely evaluated on population measures directly. In machine learning and statistics, the inputs are usually empirical laws, histograms, simulated particles or minibatches. The central question is therefore not only how to compute OT, but also how OT behaves as a random estimator. This chapter studies this statistical layer: qualitative consistency of empirical measures, non-asymptotic sample-complexity rates, and asymptotic bias–variance decompositions for exact and regularized transport costs.

This statistical convergence is conceptually different from the algorithmic convergence studied in the previous chapter. There the marginals and the temperature were fixed and one asked how Sinkhorn iterates approach a regularized optimizer. Here the number of samples grows, the empirical measures themselves move, and the regularization parameter may either remain fixed or be sent to zero. The resulting picture explains why exact OT is statistically expensive in high intrinsic dimension, why fixed-temperature Sinkhorn has smoother parametric fluctuations, and why approximating exact OT with entropy always involves a bias–variance tradeoff.

10.1 Law of Large Numbers and Central Limit Theorem

This section separates qualitative empirical convergence from fluctuation-scale limit theorems.

Empirical laws. Before discussing sample complexity, it is useful to separate consistency from rates. If $X_1, \dots, X_n$ are i.i.d. samples with common law α, the associated empirical measure is the random probability measure

$$\hat{\alpha}_n := \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i}. \tag{10.1}$$

The ordinary law of large numbers says that empirical averages converge to expectations. In measure language this means that $\hat{\alpha}_n$ converges weakly toward α, because testing $\hat{\alpha}_n$ against a bounded continuous function φ gives the sample average $n^{-1} \sum_i \varphi(X_i)$. Wasserstein distances strengthen this statement by also recording moment convergence. Thus, if α has a finite p-th moment, the empirical law converges to α in $\mathcal{W}_p$, almost surely and in p-th mean in the sense that $\mathbb{E}\, \mathcal{W}_p(\hat{\alpha}_n, \alpha)^p \to 0$. This is the qualitative consistency statement behind empirical OT plug-in estimators: empirical transport distances converge to their population counterparts once the sampled laws themselves converge in Wasserstein distance. It says nothing yet about the speed, which is the topic of Section 10.2.

Proposition 10.1 (Empirical law of large numbers in $\mathcal{W}_p$)**.** *Let $(\mathcal{X}, d)$ be a Polish metric space, let $p \geqslant 1$, and let $\alpha \in \mathcal{P}_p(\mathcal{X})$. Let $(X_i)_{i \geqslant 1}$ be i.i.d. random variables with law α, and define $\hat{\alpha}_n$ by (10.1). Then*

$$\hat{\alpha}_n \rightharpoonup \alpha \quad \text{and} \quad \mathcal{W}_p(\hat{\alpha}_n, \alpha) \longrightarrow 0$$

almost surely. Moreover,

$$\mathbb{E}\, \mathcal{W}_p(\hat{\alpha}_n, \alpha)^p \longrightarrow 0.$$

Proof. Fix a reference point $x_0 \in \mathcal{X}$, and write $r(x) = d(x, x_0)$. Since $\mathcal{X}$ is Polish, the weak topology on $\mathcal{P}(\mathcal{X})$ admits a countable convergence-determining class $(\varphi_k)_{k \geqslant 1} \subset C_b(\mathcal{X})$. For each fixed k, the strong law of large numbers gives

$$\int \varphi_k \mathrm{d}\hat{\alpha}_n = \frac{1}{n} \sum_{i=1}^{n} \varphi_k(X_i) \longrightarrow \mathbb{E}\varphi_k(X_1) = \int \varphi_k \mathrm{d}\alpha$$

almost surely. Intersecting the corresponding probability-one events over the countable set of indices gives, on a single event of probability one, convergence against every φ_k. By the convergence-determining property, this is exactly $\hat{\alpha}_n \rightharpoonup \alpha$.

The moment condition $\alpha \in \mathcal{P}_p(\mathcal{X})$ means $\int r^p \mathrm{d}\alpha < +\infty$. Applying the strong law again to the integrable random variable $r(X_1)^p$ gives

$$\int r^p \mathrm{d}\hat{\alpha}_n = \frac{1}{n}\sum_{i=1}^n r(X_i)^p \longrightarrow \int r^p \mathrm{d}\alpha$$

almost surely. Proposition 4.25 states that weak convergence plus convergence of the p-th moments is equivalent to $\mathcal{W}_p$-convergence on $\mathcal{P}_p(\mathcal{X})$. Hence $\mathcal{W}_p(\hat{\alpha}_n, \alpha) \to 0$ almost surely.

It remains to justify convergence in expectation. Set $A_n = \int r^p \mathrm{d}\hat{\alpha}_n$ and $M = \int r^p \mathrm{d}\alpha$. The triangle inequality through the Dirac mass δ_{x_0}, followed by $(a+b)^p \leqslant 2^{p-1}(a^p + b^p)$, gives the deterministic bound

$$\mathcal{W}_p(\hat{\alpha}_n, \alpha)^p \leqslant 2^{p-1}\big(A_n + M\big).$$

The family $(A_n)_n$ is uniformly integrable. Indeed, by the de la Vallée–Poussin criterion, choose a convex superlinear function Ψ such that $\mathbb{E}\Psi(r(X_1)^p) < +\infty$; Jensen's inequality gives

$$\mathbb{E}\Psi(A_n) \leqslant \frac{1}{n}\sum_{i=1}^n \mathbb{E}\Psi(r(X_i)^p) = \mathbb{E}\Psi(r(X_1)^p).$$

Thus $(A_n + M)_n$, and hence $(\mathcal{W}_p(\hat{\alpha}_n, \alpha)^p)_n$, is uniformly integrable. Together with almost-sure convergence to zero, this implies $\mathbb{E}\,\mathcal{W}_p(\hat{\alpha}_n, \alpha)^p \to 0$. □

Central-limit fluctuations. The previous proposition is a law-of-large-numbers statement: a random empirical measure converges to the law that generated it. The central limit theorem describes a different, fluctuation-scale limit. As recalled in Remark 4.23, if $(X_i)_{i\geqslant 1}$ are centered i.i.d. random vectors with identity covariance, the law of $n^{-1/2}\sum_i X_i$ converges weakly toward a Gaussian. Equivalently, if α is the common law of the X_i, this law is the rescaled convolution $(D_{1/\sqrt{n}})_\sharp \alpha^{*n}$. Wasserstein distances make this qualitative convergence quantitative. The next result is a $\mathcal{W}_1$ form of the Berry–Esseen theorem: it controls the error uniformly over all 1-Lipschitz test functions. The qualitative Bernoulli example was visualized earlier in Figure 4.3.

The following estimate is the Wasserstein form of the classical Berry–Esseen theorem [48, 208]. The proof sketch uses Stein's method, whose modern normal-approximation formulation is developed in [126].

Proposition 10.2 (Berry–Esseen bound in $\mathcal{W}_1$). *Let $(X_i)_{i=1}^n$ be i.i.d. real random variables such that $\mathbb{E}X_i = 0$, $\mathbb{E}X_i^2 = 1$ and $\mathbb{E}|X_i|^3 < +\infty$. If α_n is the law of $n^{-1/2}\sum_i X_i$ and γ is the standard Gaussian law, then*

$$\mathcal{W}_1(\alpha_n, \gamma) \leqslant \frac{C\,\mathbb{E}|X_1|^3}{\sqrt{n}},$$

where C is a universal constant.

Proof. By Kantorovich–Rubinstein duality,

$$\mathcal{W}_1(\alpha_n, \gamma) = \sup_{\mathrm{Lip}(h)\leqslant 1} \left|\mathbb{E}h(S_n) - \mathbb{E}h(G)\right|, \qquad S_n = n^{-1/2}\sum_i X_i,\ G \sim \gamma.$$

For each such h, solve Stein's equation $f_h'(x) - x f_h(x) = h(x) - \mathbb{E}h(G)$. The solution satisfies $\|f_h'\|_\infty + \|f_h''\|_\infty \leqslant C$ for a universal constant. Hence

$$\mathbb{E}h(S_n) - \mathbb{E}h(G) = \mathbb{E}[f_h'(S_n) - S_n f_h(S_n)].$$

Write $S_n^{(i)} = S_n - X_i/\sqrt{n}$. Independence and $\mathbb{E}X_i = 0$ give

$$\mathbb{E}[S_n f_h(S_n)] = \frac{1}{\sqrt{n}}\sum_{i=1}^n \mathbb{E}\Big[X_i\big(f_h(S_n^{(i)} + X_i/\sqrt{n}) - f_h(S_n^{(i)})\big)\Big].$$

Taylor's formula with integral remainder and $\mathbb{E}X_i^2 = 1$ imply

$$\left|\mathbb{E}[S_n f_h(S_n)] - \frac{1}{n}\sum_{i=1}^n \mathbb{E}f_h'(S_n^{(i)})\right| \leqslant \frac{C\,\mathbb{E}|X_1|^3}{\sqrt{n}}.$$

Moreover, the Lipschitz bound on f_h' yields

$$\left| \mathbb{E} f_h'(S_n) - \frac{1}{n} \sum_{i=1}^{n} \mathbb{E} f_h'(S_n^{(i)}) \right| \leqslant \frac{C\, \mathbb{E}|X_1|}{\sqrt{n}} \leqslant \frac{C\, \mathbb{E}|X_1|^3}{\sqrt{n}},$$

where the last inequality uses $\mathbb{E}X_1^2 = 1$. Combining the two bounds and taking the supremum over h proves the result. Sharper constants and higher-order transport-distance refinements are studied in [67, 451]. □

The universal $n^{-1/2}$ bound is sharp over broad classes of input laws, but it need not describe the asymptotic behavior of a fixed law. For symmetric inputs, the smooth skewness correction vanishes. What replaces it depends decisively on whether the law is lattice-valued or has a density.

Proposition 10.3 (Sharp symmetric lattice asymptotic)**.** *Let X be centered, symmetric, of unit variance, and supported on a finite subset of $a + h\mathbb{Z}$, where $h > 0$ is the maximal lattice span. If α_n is the law of $n^{-1/2} \sum_{i=1}^n X_i$ and $\gamma = \mathcal{N}(0,1)$, then*

$$\sqrt{n}\, \mathcal{W}_1(\alpha_n, \gamma) \longrightarrow \frac{h}{4}.$$

Proof. Denote by F_n the CDF of α_n, and by Φ and φ the standard Gaussian CDF and density. With the centered periodic sawtooth $\psi(u) = \frac{1}{2} - \{u\}$, the integrated lattice Edgeworth expansion gives

$$F_n(x) - \Phi(x) = \frac{h}{\sqrt{n}}\, \psi\!\left(\frac{\sqrt{n}\, x - na}{h} \right) \varphi(x) + r_n(x), \qquad \|r_n\|_{L^1(\mathbb{R})} = o(n^{-1/2});$$

see [424, Chapter VII], [55, Chapter 5], and [311]. Vallender's one-dimensional identity [528] states that $\mathcal{W}_1(\alpha_n, \gamma) = \int_{\mathbb{R}} |F_n - \Phi|$. Therefore $\big||u+v| - |u|\big| \leqslant |v|$ reduces the claim to

$$\int_{\mathbb{R}} \left| \psi\!\left(\frac{\sqrt{n}\, x - na}{h} \right) \right| \varphi(x) \mathrm{d}x \longrightarrow \int_0^1 |\psi(u)| \mathrm{d}u = \frac{1}{4}.$$

The convergence is periodic averaging: prove it first for compactly supported step functions, where it is a Riemann sum over periods, and conclude by L^1-approximation of φ. □

The absence of a lattice correction reveals the next smooth Edgeworth term. A density automatically satisfies Cramér's non-lattice condition because its characteristic function vanishes at infinity.

Proposition 10.4 (Sharp symmetric density asymptotic)**.** *Let X be centered, symmetric, of unit variance, with a density and moments of every order. Set $\kappa_4 = \mathbb{E}[X^4] - 3$ and $H_3(x) = x^3 - 3x$. If α_n is the law of $n^{-1/2} \sum_{i=1}^n X_i$ and $\gamma = \mathcal{N}(0,1)$, then*

$$\mathcal{W}_1(\alpha_n, \gamma) = \frac{|\kappa_4|}{24n} \int_{\mathbb{R}} |H_3(x)| \varphi(x) \mathrm{d}x + o(n^{-1}).$$

Proof. The non-lattice Edgeworth expansion through order n^{-1} has an $n^{-1/2}$ term proportional to the third cumulant, and n^{-1} terms proportional to the fourth cumulant and to the square of the third one. Symmetry makes the third cumulant vanish and gives

$$F_n(x) - \Phi(x) = -\frac{\kappa_4}{24n} H_3(x) \varphi(x) + r_n(x), \qquad \|r_n\|_{L^1(\mathbb{R})} = o(n^{-1});$$

see [424, Chapter VI] and [55, Chapter 4]. Inserting this expansion into Vallender's identity and using $\big||u+v| - |u|\big| \leqslant |v|$ proves the formula. □

These two results explain the different behaviors in Figure 10.1. For the symmetric Bernoulli law, the maximal span is $h = 2$, hence $\mathcal{W}_1(\alpha_n, \gamma) = 1/(2\sqrt{n}) + o(n^{-1/2})$. Proposition 10.4 cannot be used: this law has no density and its characteristic function is $\mathbb{E}[e^{itX}] = \cos t$, whose modulus returns to one at arbitrarily large frequencies. Thus symmetry removes the smooth skewness term but not the lattice sawtooth. Conversely, $X \sim \mathrm{Unif}[-\sqrt{3}, \sqrt{3}]$ has a density, characteristic function $\sin(\sqrt{3}t)/(\sqrt{3}t) \to 0$, and $\kappa_4 = -6/5$. Since

$$\int_{\mathbb{R}} |H_3(x)| \varphi(x) \mathrm{d}x = 2\varphi(0) + 8\varphi(\sqrt{3}) = \frac{2 + 8e^{-3/2}}{\sqrt{2\pi}},$$

one obtains the two sharp equivalents

$$\mathcal{W}_1(\alpha_n, \gamma) \sim \begin{cases} \dfrac{1}{2\sqrt{n}}, & X \sim \tfrac{1}{2}(\delta_{-1} + \delta_1), \\ \dfrac{1 + 4e^{-3/2}}{10\sqrt{2\pi}} \dfrac{1}{n}, & X \sim \mathrm{Unif}[-\sqrt{3}, \sqrt{3}]. \end{cases} \tag{10.2}$$

Figure 10.1 confronts these equivalents with exact one-dimensional computations. For a centered, unit-variance input law α_0, define

$$\alpha_n = (D_{1/\sqrt{n}})_\sharp \alpha_0^{*n}, \qquad n \geqslant 1,$$

so that $\alpha_1 = \alpha_0$. For Bernoulli input, the distance is evaluated by the exact quantile formula and atom masses are divided by the current lattice spacing in the density display. For continuous-uniform input, the normalized convolution is an affine image of the Irwin–Hall distribution and Vallender's CDF formula is integrated numerically. Neither computation uses Monte Carlo sampling.

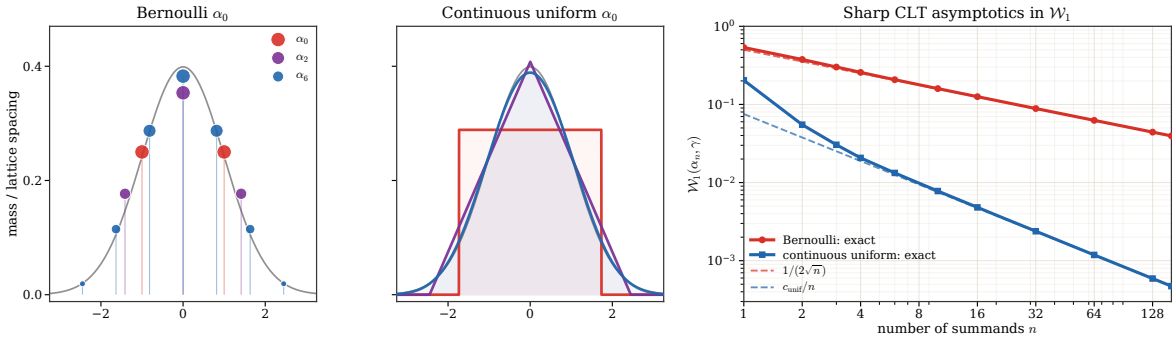


Figure 10.1: Sharp lattice and density central-limit asymptotics in $\mathcal{W}_1$. The two left panels show α_0, α_2, and α_6 for symmetric Bernoulli and continuous-uniform inputs; the gray curve is the standard Gaussian density. The right panel compares the exact numerical distances (solid) with the sharp equivalents in (10.2) (dashed). The Bernoulli curve follows its lattice rate $1/(2\sqrt{n})$, while the continuous-uniform curve approaches $(1 + 4e^{-3/2})/(10\sqrt{2\pi}\, n)$, confirming the faster density-law asymptotic.

Empirical-process fluctuations. The central-limit object needed for statistical OT is not only the normalized sum of random vectors above, but the fluctuation of the empirical law itself. There is no canonical Gaussian probability measure on an unrestricted infinite-dimensional space of signed measures, so the precise statement is made through test functions. For a finite collection, the result is the multivariate central limit theorem; uniform convergence over an infinite class requires that class to be Donsker [529].

Proposition 10.5 (Finite-dimensional empirical-process CLT)**.** *Let $X_1, X_2, \ldots$ be i.i.d. with law α, and let $\varphi_1, \ldots, \varphi_q \in L^2(\alpha)$. Then*

$$\sqrt{n} \left(\int \varphi_r \mathrm{d}(\hat{\alpha}_n - \alpha) \right)_{r=1}^{q} \Longrightarrow G_\alpha,$$

where G_α is centered Gaussian in $\mathbb{R}^q$ with covariance

$$\mathbb{E}[(G_\alpha)_r (G_\alpha)_s] = \int \left(\varphi_r - \int \varphi_r \mathrm{d}\alpha \right) \left(\varphi_s - \int \varphi_s \mathrm{d}\alpha \right) \mathrm{d}\alpha.$$

Proof. Apply the multivariate central limit theorem to the i.i.d. centered vectors $(\varphi_r(X_i) - \int \varphi_r \mathrm{d}\alpha)_{r=1}^{q}$. □

This proposition is the linear input to the delta-method arguments in Section 10.3. The difficulty is that an OT value need not be differentiable as a function of its marginals: exact finite-space OT can have only a directional derivative, while fixed-temperature entropic OT is smooth in the interior of the simplices.

10.2 Sample Complexity

This section compares four statistical regimes. Unregularized OT resolves geometry at all spatial scales and pays rates controlled by the metric entropy, hence by the intrinsic dimension, of the support. MMD replaces transport by a kernel mean embedding and therefore has a parametric Monte-Carlo rate for bounded kernels. Fixed-temperature entropic OT and Sinkhorn divergences sit between these pictures: they smooth the dual potentials and recover parametric fluctuations, but introduce a regularization bias. The section closes with sliced Wasserstein, whose projected one-dimensional structure gives a different route to favorable empirical rates.

Unregularized OT. Exact Wasserstein distances are geometrically sharp, but this sharpness is statistically expensive in worst-case nonparametric regimes. The previous section proves qualitative convergence of empirical laws; the sample-complexity question is how fast this convergence happens. The sample complexity of unregularized OT suffers from the curse of dimensionality, but the relevant dimension is geometric. If the distributions are supported on a regular lower-dimensional set, for instance a d'-dimensional submanifold of $\mathbb{R}^d$, the empirical rate is governed by d' rather than by the ambient dimension d. In the high-dimensional regime for $\mathcal{W}_p$, namely $d' > 2p$, this gives the characteristic rate $n^{-1/d'}$. Exact OT is therefore dimension-adaptive: it sees the intrinsic dimension of the data support through its covering numbers [192, 548]. Related two-sample-testing viewpoints are developed in [448].

For compactly supported measures, it is convenient to record the worst-case empirical Wasserstein scale once and reuse it throughout the book. For $n \geqslant 2$, $p \geqslant 1$, and $d \geqslant 1$, define

$$r_{n,p,d} := \begin{cases} n^{-1/(2p)}, & d < 2p, \\ n^{-1/(2p)}(\log(1+n))^{1/p}, & d = 2p, \\ n^{-1/d}, & d > 2p. \end{cases} \tag{10.3}$$

Indeed, if $\widehat{\alpha}_n = n^{-1}\sum_{i=1}^n \delta_{X_i}$ for i.i.d. samples $X_i \sim \alpha$ and $\mathcal{X} \subset \mathbb{R}^d$ is compact, then

$$\sup_{\alpha \in \mathcal{P}(\mathcal{X})} \mathbb{E}\big[\mathcal{W}_p(\widehat{\alpha}_n, \alpha)^p\big]^{1/p} \leqslant C_{\mathcal{X},p,d}\, r_{n,p,d}. \tag{10.4}$$

These worst-case rates follow from empirical Wasserstein moment estimates [185, 220]. They can improve under additional regularity or lower-dimensional support assumptions, as discussed below.

The elementary mechanism behind this phenomenon is a multiscale coupling. One compares the two measures first at a coarse resolution, then corrects the residual imbalance inside smaller and smaller cells. The next proposition gives a convenient version of this idea for $\mathcal{W}_1$ on the unit cube.

Figure 10.2 first gives a numerical overview of the statistical regimes compared in this section: exact OT exhibits dimension-dependent empirical fluctuations, whereas MMD and fixed-temperature Sinkhorn divergences are much closer to the parametric $n^{-1/2}$ scale.

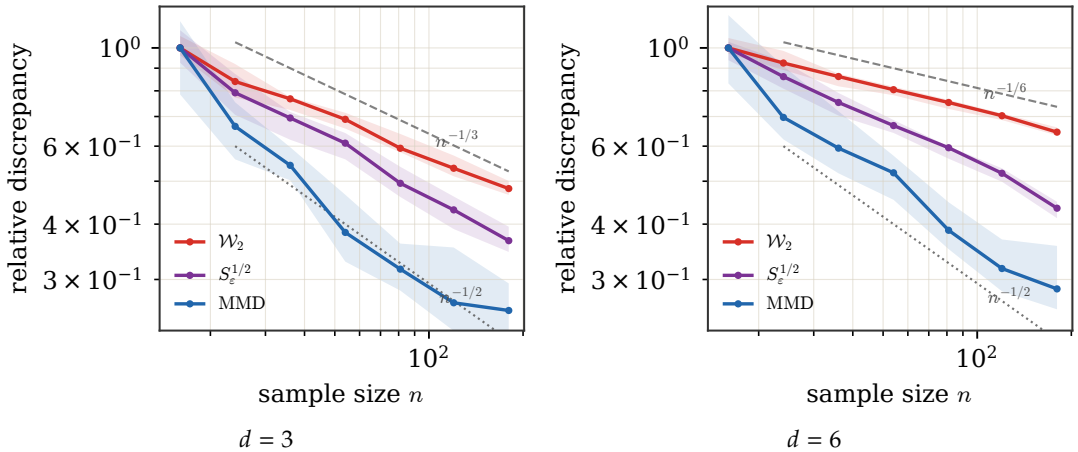


Figure 10.2: Empirical fluctuations in dimensions three and six. For each sample size n, two independent empirical measures α_n and α'_n are drawn from the same standard Gaussian law in $\mathbb{R}^d$. The curves show the median of $D(\alpha_n, \alpha'_n)$, normalized by its value at the smallest displayed n, with interquartile bands. Exact OT follows a slower dimension-dependent scale, while MMD and the fixed-ε Sinkhorn divergence behave closer to the parametric $n^{-1/2}$ guide. This is a statistical illustration, not a solver benchmark.

Proposition 10.6 (Dyadic partition bound for $\mathcal{W}_1$)**.** *Let $\mathcal{Q}_j$ be nested dyadic partitions of $[0,1]^d$ into half-open cubes of side length 2^{-j}, with cubes on the outer face closed along the boundary of $[0,1]^d$. For all $\alpha, \beta \in \mathcal{P}([0,1]^d)$ and every integer $J \geqslant 0$,*

$$\mathcal{W}_1(\alpha,\beta) \leqslant \sqrt{d} \sum_{j=0}^{J-1} 2^{-j} \sum_{Q \in \mathcal{Q}_{j+1}} |\alpha(Q) - \beta(Q)| + \sqrt{d}\, 2^{-J}.$$

Proof. The proof is constructive: one decomposes the two measures into pieces which can be coupled at different spatial scales. Write $\Omega = [0,1]^d$. We build positive measures $(\alpha_j, \beta_j)_{j=0}^J$ such that

$$\sum_{j=0}^{J} \alpha_j = \alpha, \qquad \sum_{j=0}^{J} \beta_j = \beta, \qquad \alpha_j(Q) = \beta_j(Q) \quad \text{for every } Q \in \mathcal{Q}_j.$$

Assuming such a decomposition for a moment, each pair (α_j, β_j) can be coupled inside the cells of $\mathcal{Q}_j$. Indeed, on every cell $Q \in \mathcal{Q}_j$ with positive mass, take the product coupling between the normalized restrictions of α_j and β_j to Q, multiplied by the common mass $\alpha_j(Q) = \beta_j(Q)$. Summing over cells and over j gives a coupling of α and β. Since the diameter of a cell in $\mathcal{Q}_j$ is at most $\sqrt{d}\, 2^{-j}$, this coupling has cost bounded by

$$\mathcal{W}_1(\alpha,\beta) \leqslant \sqrt{d} \sum_{j=0}^{J} 2^{-j} \alpha_j(\Omega).$$

It remains to construct the decomposition and control the masses $\alpha_j(\Omega)$. Start at the finest scale. For $Q \in \mathcal{Q}_J$, define α_J and β_J on Q by keeping the common part of the two masses:

$$\alpha_J(A \cap Q) = \frac{\alpha(Q) \wedge \beta(Q)}{\alpha(Q)}\, \alpha(A \cap Q), \qquad \beta_J(A \cap Q) = \frac{\alpha(Q) \wedge \beta(Q)}{\beta(Q)}\, \beta(A \cap Q),$$

with the convention that the corresponding term is zero when the denominator vanishes. For $j = J-1, \dots, 1$, assume that the pieces at all finer levels $k > j$ have been constructed and define the residuals

$$\alpha'_j = \alpha - \sum_{k=j+1}^{J} \alpha_k, \qquad \beta'_j = \beta - \sum_{k=j+1}^{J} \beta_k.$$

On each $Q \in \mathcal{Q}_j$, set α_j and β_j equal to the common part of α'_j and β'_j inside Q, exactly as above. Finally set $\alpha_0 = \alpha - \sum_{k=1}^{J} \alpha_k$ and $\beta_0 = \beta - \sum_{k=1}^{J} \beta_k$. Each step removes only common residual mass, so all measures are positive, and by construction $\alpha_j(Q) = \beta_j(Q)$ on all cells $Q \in \mathcal{Q}_j$.

We now estimate how much mass is first matched at each scale. For notational consistency, set $\alpha'_0 = \alpha_0$, $\beta'_0 = \beta_0$, $\alpha'_J = \alpha$, and $\beta'_J = \beta$. Let $0 \leqslant j < J$ and $R \in \mathcal{Q}_j$. Since all pieces constructed at levels finer than $j+1$ have equal mass on every child $Q \in \mathcal{Q}_{j+1}$, the residual imbalance on Q just before the matching at level $j+1$ is the original imbalance:

$$\alpha'_{j+1}(Q) - \beta'_{j+1}(Q) = \alpha(Q) - \beta(Q).$$

After removing the common part at level $j+1$, the remaining source residual inside Q has mass

$$\alpha'_j(Q) = \alpha'_{j+1}(Q) - \alpha_{j+1}(Q) = \big(\alpha'_{j+1}(Q) - \beta'_{j+1}(Q)\big)_+.$$

Because α_j is itself a common submeasure of the residual α'_j, the mass matched at level j inside R satisfies

$$\alpha_j(R) = \beta_j(R) \leqslant \alpha'_j(R) = \sum_{Q \subset R} \big(\alpha'_{j+1}(Q) - \beta'_{j+1}(Q)\big)_+ \leqslant \sum_{Q \subset R} |\alpha(Q) - \beta(Q)|\,,$$

where the sums are over $Q \in \mathcal{Q}_{j+1}$. Summing over $R \in \mathcal{Q}_j$ gives

$$\alpha_j(\Omega) \leqslant \sum_{Q \in \mathcal{Q}_{j+1}} |\alpha(Q) - \beta(Q)|\,.$$

The last level satisfies $\alpha_J(\Omega) \leqslant 1$. Substituting these bounds into the cost estimate proves the claim. □

Proposition 10.7 (Empirical OT has intrinsic-dimension value rates)**.** *Let α and β be probability measures supported on $[0,1]^d$, and let $\hat{\alpha}_n$ and $\hat{\beta}_m$ be independent empirical measures. Then*

$$\mathbb{E}\left|\mathcal{W}_1(\hat{\alpha}_n,\hat{\beta}_m)-\mathcal{W}_1(\alpha,\beta)\right| \lesssim_d r_{n,1,d}+r_{m,1,d},$$

In particular, for $d \geqslant 3$, the worst-case one-sample empirical error for exact $\mathcal{W}_1$ has the characteristic high-dimensional scale $n^{-1/d}$. More generally, for $\mathcal{W}_p$ on regular d'-dimensional supports, the corresponding high-dimensional rate is $n^{-1/d'}$ when $d' > 2p$, under the usual volume-growth and moment assumptions. Thus the relevant dimension is the intrinsic dimension of the support rather than necessarily the ambient dimension.

Proof. By the triangle inequality,

$$\left|\mathcal{W}_1(\hat{\alpha}_n,\hat{\beta}_m)-\mathcal{W}_1(\alpha,\beta)\right| \leqslant \mathcal{W}_1(\hat{\alpha}_n,\alpha)+\mathcal{W}_1(\hat{\beta}_m,\beta).$$

It is therefore enough to bound $\mathbb{E}\,\mathcal{W}_1(\hat{\alpha}_N,\alpha)$. Apply Proposition 10.6 to $(\hat{\alpha}_N,\alpha)$. If $Q \in \mathcal{Q}_{j+1}$, then $N\hat{\alpha}_N(Q)$ is a binomial random variable with parameter $\alpha(Q)$, and hence

$$\mathbb{E}\,|\hat{\alpha}_N(Q)-\alpha(Q)| \leqslant \sqrt{\mathbb{E}\,|\hat{\alpha}_N(Q)-\alpha(Q)|^2} \leqslant \sqrt{\frac{\alpha(Q)}{N}}.$$

Summing over the $2^{(j+1)d}$ cells and using the Cauchy–Schwarz inequality gives

$$\sum_{Q\in\mathcal{Q}_{j+1}} \mathbb{E}\,|\hat{\alpha}_N(Q)-\alpha(Q)| \leqslant \frac{1}{\sqrt{N}} \sum_{Q\in\mathcal{Q}_{j+1}} \sqrt{\alpha(Q)} \leqslant \frac{2^{(j+1)d/2}}{\sqrt{N}}.$$

Therefore, for every $J \geqslant 0$,

$$\mathbb{E}\,\mathcal{W}_1(\hat{\alpha}_N,\alpha) \lesssim_d 2^{-J}+\frac{1}{\sqrt{N}}\sum_{j=0}^{J-1} 2^{j(d/2-1)}.$$

If $d=1$, the sum is uniformly bounded and we choose J such that $2^{-J} \lesssim N^{-1/2}$. If $d=2$, the sum is J, and choosing $2^J \simeq N^{1/2}$ gives $(\log N)N^{-1/2}$. If $d \geqslant 3$, the sum is dominated by its last term; choosing $2^J \simeq N^{1/d}$ gives

$$2^{-J}+N^{-1/2}2^{J(d/2-1)} \simeq N^{-1/d}.$$

This proves the displayed $\mathcal{W}_1$ rates. The extension to a regular d'-dimensional support uses nested partitions whose cells have diameter $O(2^{-j})$ and cardinality $O(2^{jd'})$; the general $\mathcal{W}_p$ statement follows from the standard multiscale estimates of [192, 220, 548]. The elementary dyadic estimate is not sharp for every law in the critical case $d=2$. For example, for the uniform law, and more generally under suitable density regularity, finer matching arguments give the smaller order $(\log N/N)^{1/2}$. The proposition deliberately retains the distribution-free $(\log N)N^{-1/2}$ bound supplied by the dyadic proof. □

Lower bounds and minimax optimality. The empirical upper bound above is not merely a weakness of the plug-in estimator. In fully nonparametric classes, the $N^{-1/d}$ scale is a genuine minimax obstruction: from N samples one can resolve only about N spatial boxes, and the Wasserstein cost of an unresolved signed imbalance is proportional to the box diameter. This idea goes back to Dudley's entropy method and is now understood sharply for empirical measures and minimax distribution estimation in Wasserstein loss [192, 220, 488, 548, 135]. The following elementary Assouad argument gives the basic obstruction already for measures with bounded densities.

Proposition 10.8 (Minimax lower bound for bounded densities)**.** *Let $d \geqslant 3$. Let $\mathcal{A}$ be the class of probability measures on $[0,1]^d$ admitting a density ρ with respect to Lebesgue measure such that*

$$\frac{1}{2} \leqslant \rho(x) \leqslant \frac{3}{2} \qquad \text{for a.e. } x \in [0,1]^d.$$

There exists $c_d > 0$ such that, for every $N \geqslant 1$,

$$\inf_{\widetilde{\alpha}_N} \sup_{\alpha\in\mathcal{A}} \mathbb{E}_\alpha\, \mathcal{W}_1(\widetilde{\alpha}_N,\alpha) \geqslant c_d\, N^{-1/d}.$$

Here the infimum is over all estimators based on N i.i.d. samples from α. In particular, the high-dimensional rate of Proposition 10.7 is minimax sharp for $\mathcal{W}_1$, even when one allows arbitrary estimators rather than the empirical measure.

Proof. We use a standard hypercube construction. Choose a smooth function ψ compactly supported in $(0,1)^d$, with $\int \psi = 0$, $\|\psi\|_\infty \leqslant 1$, and whose positive and negative parts are separated in the first coordinate. There exists a compactly supported Lipschitz function φ such that $\int \varphi\psi \mathrm{d}x > 0$. We extend both functions by zero outside $(0,1)^d$, and absorb $\mathrm{Lip}(\varphi)$ into the constants below.

Let m be an integer, $h = 1/m$, and partition $[0,1]^d$ into $M = m^d$ cubes $Q_k = z_k + h[0,1)^d$. Define the rescaled bumps

$$\psi_k(x) = \psi\Big(\frac{x - z_k}{h}\Big)\mathbf{1}_{Q_k}(x).$$

For $\theta = (\theta_1, \dots, \theta_M) \in \{-1,1\}^M$ and a fixed small $\eta > 0$, set

$$\rho_\theta(x) = 1 + \eta \sum_{k=1}^{M} \theta_k \psi_k(x), \qquad \alpha_\theta = \rho_\theta \mathrm{d}x.$$

For η small enough, all these densities belong to $\mathcal{A}$, and they integrate to one because every ψ_k has zero mean.

If θ and θ' differ on a set S of coordinates, the dual formula for $\mathcal{W}_1$ gives an additive lower bound. Indeed, rescale φ as $\varphi_k(x) = h\,\varphi\big(\frac{x-z_k}{h}\big)$. Each φ_k is supported in a fixed strict subcube of Q_k. Consequently every signed sum $\sum_k s_k\varphi_k$, $s_k \in \{-1,0,1\}$, has a Lipschitz constant bounded by a constant depending only on φ, uniformly in m and in the signs. After dividing by this constant, Kantorovich–Rubinstein duality and the choice $s_k = \mathrm{sign}(\theta_k - \theta'_k)$ on S give

$$\mathcal{W}_1(\alpha_\theta, \alpha_{\theta'}) \geqslant c\,\eta\,h^{d+1}\,\mathrm{Ham}(\theta, \theta'),$$

where Ham denotes Hamming distance and $c > 0$ depends only on the chosen bump.

Neighboring vertices of the hypercube are statistically close. If $\theta^{[k]}$ is obtained from θ by flipping only the k-th sign, then the one-sample Kullback–Leibler divergence is bounded by

$$\mathrm{KL}(\alpha_\theta | \alpha_{\theta^{[k]}}) \leqslant C\eta^2 h^d,$$

because the two densities differ only on Q_k, the perturbation has size $O(\eta)$, and the densities are uniformly bounded away from zero. For N independent samples the divergence is multiplied by N. Choose m so that $m^d \simeq N$. Then Nh^d is bounded, and, by taking η sufficiently small, Theorem 7.21 gives a uniform bound

$$\mathrm{TV}\left(\alpha_\theta^{\otimes N}, \alpha_{\theta^{[k]}}^{\otimes N}\right) \leqslant q < 1.$$

Assouad's lemma therefore implies

$$\inf_{\widetilde{\alpha}_N} \sup_{\theta \in \{-1,1\}^M} \mathbb{E}_\theta\, \mathcal{W}_1(\widetilde{\alpha}_N, \alpha_\theta) \geqslant c'\eta\, M h^{d+1}(1-q) = c'\eta(1-q)h \simeq N^{-1/d}.$$

Decreasing the constant handles the finitely many small values of N, and the claim follows. □

The same packing mechanism gives the corresponding high-dimensional lower bounds for $\mathcal{W}_p$: in the regime $d > 2p$, one obtains the scale $N^{-1/d}$ for $\mathcal{W}_p$ itself, or $N^{-p/d}$ for $\mathcal{W}_p^p$; see [488, 548]. For estimating a scalar transport value such as $\mathcal{W}_p(\alpha, \beta)$ at a fixed nondegenerate pair, the minimax picture is more nuanced because the value can be differentiable and have $N^{-1/2}$ fluctuations, as in Section 10.3. The lower bound above nevertheless captures the nonparametric obstruction near the diagonal and for distribution recovery: without structural assumptions beyond bounded density or low intrinsic dimension, the curse-of-dimensionality scale cannot be removed by changing the estimator.

Leveraging smoothness. The preceding minimax obstruction is a worst-case statement over rough density classes. If the unknown laws have smooth densities, one can do better by smoothing the empirical measure before computing OT. This is a genuine statistical improvement: for $s > 0$, the right smoothing scale removes the finest empirical oscillations instead of transporting them at scale $N^{-1/d}$. The following representative statement is a convenient $\mathcal{W}_1$ corollary of the minimax theory of Niles-Weed and Berthet [401]; related $\mathcal{W}_1$ results can also be viewed through Besov integral-probability metrics [525].

Proposition 10.9 (Smoothed plug-in rates)**.** *Let $d \geqslant 3$, $s > 0$, and let $\alpha = \rho_\alpha \mathrm{d}x$, $\beta = \rho_\beta \mathrm{d}x$ be probability measures on $[0,1]^d$. Assume that ρ_α and ρ_β are bounded above and below by positive constants and belong to a Hölder–Zygmund, equivalently Besov $B^s_{\infty,\infty}$, ball of radius L. Let $\hat{\alpha}_n$ and $\hat{\beta}_m$ be independent empirical measures. There exist normalized nonnegative density estimators*

$$\tilde{\alpha}_n = \rho_{\tilde{\alpha}_n} \mathrm{d}x, \qquad \tilde{\beta}_m = \rho_{\tilde{\beta}_m} \mathrm{d}x,$$

obtained for instance by a projected wavelet estimator at resolution $h_N \simeq N^{-1/(d+2s)}$, such that

$$\mathbb{E}\left|\mathcal{W}_1(\tilde{\alpha}_n, \tilde{\beta}_m) - \mathcal{W}_1(\alpha, \beta)\right| \leqslant C\left(n^{-\frac{s+1}{d+2s}} + m^{-\frac{s+1}{d+2s}}\right).$$

The constant depends on d, s, L, on the chosen wavelet system and on the lower and upper density bounds. Thus any positive smoothness improves on the empirical $N^{-1/d}$ exponent when $d > 2$; if s is of order d, the exponent is of order $1/3$, and as $s \to \infty$ it approaches the parametric exponent $1/2$.

Proof sketch. It is enough, by the triangle inequality, to bound $\mathbb{E}\,\mathcal{W}_1(\tilde{\alpha}_N, \alpha)$. For densities bounded below and above, $\mathcal{W}_1$ is controlled by a negative smoothness norm of the density difference; in a wavelet notation one may write, schematically,

$$\mathcal{W}_1(\tilde{\alpha}_N, \alpha) \lesssim \|\rho_{\tilde{\alpha}_N} - \rho_\alpha\|_{B^{-1}_{1,1}}.$$

Truncating the empirical wavelet expansion at scale J, or equivalently smoothing at bandwidth $h = 2^{-J}$, gives two terms. The deterministic approximation error of an s-smooth density is $O(2^{-J(s+1)})$ in the negative B^{-1} norm, while the stochastic contribution of the retained coefficients is

$$N^{-1/2} \sum_{j \leqslant J} 2^{j(d/2-1)} \simeq N^{-1/2} 2^{J(d/2-1)} \qquad (d \geqslant 3).$$

Balancing these two quantities gives $2^J \simeq N^{1/(d+2s)}$ and hence the rate $N^{-(s+1)/(d+2s)}$. Projecting the linear wavelet estimate onto the cone of probability densities in the same negative Besov norm preserves the order of this bound. Applying the same estimate to both marginals proves the displayed two-sample bound. The full minimax theorem, including $\mathcal{W}_p$, Besov classes, lower bounds and endpoint logarithms in dimensions 1 and 2, is proved in [401]. □

This result should not be confused with a free computational cure for high-dimensional OT. Literal kernel smoothing, $\tilde{\alpha}_{n,h} = K_h * \hat{\alpha}_n$, replaces each atom by a continuous bump, so the subsequent transport problem is no longer an n-point discrete OT problem. A regular grid at bandwidth h contains $O(h^{-d})$ degrees of freedom; re-discretizing the smoothed law introduces an additional sampling error that must itself be controlled. These direct representations therefore retain an exponential dependence on d at fixed spatial resolution. Vacher, Muzellec, Rudi, Bach and Vialard use an infinite-dimensional sum-of-squares representation to obtain computational upper bounds with dimension-free exponents in sufficiently smooth regimes [527]. The resulting semidefinite-type procedures are theoretically important, but remain substantially more expensive than an ordinary discrete OT solve.

MMD. Kernel discrepancies smooth the problem before any transport is computed. Instead of matching mass spatially, MMD embeds each distribution into a Hilbert space of features and compares the two mean embeddings. For bounded kernels this turns the plug-in estimator into an ordinary Monte-Carlo average of Hilbert-valued random variables, which gives the dimension-free parametric rate below. The hidden constants still depend on the kernel and its bandwidth, so this is not a free replacement for OT geometry.

Proposition 10.10 (MMD has a parametric value rate)**.** *Let k be a bounded positive definite kernel with RKHS $\mathcal{H}_k$, and define*

$$\mathrm{MMD}_k(\alpha, \beta) = \Big\| \int k(x, \cdot) \mathrm{d}(\alpha - \beta)(x) \Big\|_{\mathcal{H}_k}.$$

If $\hat{\alpha}_n$ and $\hat{\beta}_m$ are independent empirical measures, then

$$\mathbb{E}\left|\mathrm{MMD}_k(\hat{\alpha}_n, \hat{\beta}_m) - \mathrm{MMD}_k(\alpha, \beta)\right| \leqslant \kappa\left(\frac{1}{\sqrt{n}} + \frac{1}{\sqrt{m}}\right)$$

when $k(x,x) \leqslant \kappa^2$.

Proof. Let $\Phi(x) = k(x,\cdot)$ be the feature map and $m_\alpha = \mathbb{E}\Phi(X)$. The reverse triangle inequality for the RKHS norm gives

$$\left|\mathrm{MMD}_k(\hat\alpha_n,\hat\beta_m) - \mathrm{MMD}_k(\alpha,\beta)\right| \leqslant \mathrm{MMD}_k(\hat\alpha_n,\alpha) + \mathrm{MMD}_k(\hat\beta_m,\beta).$$

Independence cancels the cross terms after taking the squared norm and expectation, giving

$$\mathbb{E}\,\mathrm{MMD}_k(\hat\alpha_n,\alpha)^2 = \frac{1}{n}\mathbb{E}\|\Phi(X) - m_\alpha\|^2_{\mathcal{H}_k} = \frac{1}{n}\Big(\mathbb{E}k(X,X) - \mathbb{E}k(X,X')\Big).$$

The same estimate applies to $\hat\beta_m$, and Jensen's inequality together with $k(x,x) \leqslant \kappa^2$ gives the displayed bound. □

Entropic OT. Entropic regularization keeps the transport geometry but smooths the dual potentials at scale $\sqrt{\varepsilon}$. For fixed $\varepsilon > 0$, this smoothing reduces empirical fluctuations to a parametric scale, in the same spirit as smooth losses in classical statistics [237, 56, 379]. The large-ε limit connects to MMD-type kernel tests [213, 253]; the small-ε limit recovers exact OT but makes the constants deteriorate. The statistical gain is therefore inseparable from the regularization bias.

Proposition 10.11 (Fixed-temperature Sinkhorn divergence has a parametric rate)**.** *Let $c(x,y) = \|x-y\|^2/2$, and assume that α and β are σ^2-subgaussian probability measures on $\mathbb{R}^d$, in the sense that*

$$\int e^{\|x\|^2/(2d\sigma^2)}\mathrm{d}\alpha(x) \leqslant 2, \qquad \int e^{\|y\|^2/(2d\sigma^2)}\mathrm{d}\beta(y) \leqslant 2.$$

Set

$$q_d := \left\lceil \frac{5d}{2} \right\rceil + 6, \qquad b_d := \left\lceil \frac{5d}{4} \right\rceil + 3, \qquad \Lambda_{d,\sigma}(\varepsilon) := \varepsilon\left(1 + \frac{\sigma^{q_d}}{\varepsilon^{b_d}}\right).$$

Let $\hat\alpha_n$ and $\hat\beta_m$ be independent empirical measures. Then, for every $\varepsilon > 0$,

$$\mathbb{E}\left|\bar{\mathcal{L}}^\varepsilon_c(\hat\alpha_n,\hat\beta_m) - \bar{\mathcal{L}}^\varepsilon_c(\alpha,\beta)\right| \leqslant C_d\,\Lambda_{d,\sigma}(\varepsilon)\left(\frac{1}{\sqrt{n}} + \frac{1}{\sqrt{m}}\right).$$

In particular, for fixed $\varepsilon > 0$ the empirical fluctuation is parametric, while for small ε the prefactor deteriorates only polynomially in $1/\varepsilon$, not as $e^{C/\varepsilon}$. Compactly supported measures are covered by this statement after a common translation of the two supports, with σ controlled by the diameter of their union.

Proof. We give the argument for the raw entropic value $\mathcal{L}^\varepsilon_c$ and then pass to the debiased expression. The original proof of Genevay–Chizat–Bach–Cuturi–Peyré [237] also reduces the problem to an empirical process indexed by entropic dual potentials, but it bounds these potentials on a compact domain in a way that produces an exponential factor $e^{D^2/\varepsilon}$. The refinement of Mena–Niles-Weed [379] keeps the same reduction and replaces this step by polynomial Hölder estimates for the quadratic-cost Sinkhorn potentials.

For probability measures α_0, β_0, write the dual objective, following (8.23), as

$$\mathcal{D}^{\alpha_0,\beta_0}_\varepsilon(u,v) := \int u\mathrm{d}\alpha_0 + \int v\mathrm{d}\beta_0 - \varepsilon\iint\left(\exp\left(\frac{u(x)+v(y)-c(x,y)}{\varepsilon}\right) - 1\right)\mathrm{d}\alpha_0(x)\mathrm{d}\beta_0(y).$$

Following the extension argument of Mena–Niles-Weed, choose optimal potentials in the normalized class for which the Sinkhorn equations hold pointwise on $\mathbb{R}^d$. Thus, at an optimal pair (u,v),

$$\int \exp\left(\frac{u(x)+v(y)-c(x,y)}{\varepsilon}\right)\mathrm{d}\beta_0(y) = 1 \qquad \text{for all } x.$$

Equivalently, the potentials are fixed points of the soft c-transform, but only the dual normalization is needed below. Let (u,v) and (u_n,v_n) be optimal dual potentials for (α_0,β_0) and (α_n,β_0). By optimality,

$$\mathcal{D}^{\alpha_0,\beta_0}_\varepsilon(u_n,v_n) \leqslant \mathcal{D}^{\alpha_0,\beta_0}_\varepsilon(u,v), \qquad \mathcal{D}^{\alpha_n,\beta_0}_\varepsilon(u,v) \leqslant \mathcal{D}^{\alpha_n,\beta_0}_\varepsilon(u_n,v_n).$$

Hence the difference of the two optimal values is bounded by evaluating the two objectives at either pair of potentials. For the pair (u, v) associated with (α_0, β_0), the preceding Sinkhorn equation cancels the exponential term in the change of the first marginal:

$$\mathcal{D}_\varepsilon^{\alpha_0,\beta_0}(u,v) - \mathcal{D}_\varepsilon^{\alpha_n,\beta_0}(u,v) = \int u \mathrm{d}(\alpha_0 - \alpha_n).$$

The same identity holds for (u_n, v_n), using its normalization with respect to β_0. Consequently, up to the harmless factor 2 used in [379],

$$|\mathcal{L}_c^\varepsilon(\alpha_n,\beta_0) - \mathcal{L}_c^\varepsilon(\alpha_0,\beta_0)| \leqslant 2 \sup_{w \in \mathcal{F}_{\varepsilon,\sigma}} \left| \int w \mathrm{d}(\alpha_n - \alpha_0) \right|,$$

where $\mathcal{F}_{\varepsilon,\sigma}$ denotes the class of first-marginal entropic potentials arising from pairs of σ^2-subgaussian marginals.

It remains to bound this empirical process. For $\varepsilon = 1$, Mena–Niles-Weed prove that the normalized quadratic-cost potentials belong to a polynomial-growth Hölder class. More precisely, after subtracting the quadratic part $x \mapsto \|x\|^2/2$, these potentials have derivatives up to order $s_0 = \lceil d/2 \rceil + 1$ whose local Hölder norms are bounded by polynomial functions of the subgaussian proxy. Denote by $\mathcal{F}^s_{\sigma,L}$ such a class, where L is a random envelope controlled by the empirical subgaussian moment. Its empirical L_2-covering numbers satisfy a bound of the form

$$\log N(\delta, \mathcal{F}^s_{\sigma,L}, L_2(\alpha_n)) \leqslant C_d\, L^{d/(2s)} \delta^{-d/s} (1 + \sigma^{2d}).$$

This holds for $s > s_0 - 1$, with $\mathbb{E}L \leqslant C$ after the normalization of the subgaussian moment. Since $d/(2s_0) < 1$, Dudley's entropy integral is finite, and the standard Rademacher symmetrization and chaining bound yields

$$\mathbb{E} \sup_{w \in \mathcal{F}_{1,\sigma}} \left| \int w \mathrm{d}(\hat{\alpha}_n - \alpha) \right| \leqslant C_d (1 + \sigma^{q_d})\, n^{-1/2}.$$

This is exactly the step which avoids the exponential $e^{D^2/\varepsilon}$ factor: one controls the regularity class of the dual potentials polynomially instead of bounding their sup-norm through the worst value of the Gibbs kernel.

The general ε case follows by scaling. If $T_\varepsilon(x) = x/\sqrt{\varepsilon}$, $\alpha_0^\varepsilon = (T_\varepsilon)_\sharp \alpha_0$, and $\beta_0^\varepsilon = (T_\varepsilon)_\sharp \beta_0$, then

$$\mathcal{L}_c^\varepsilon(\alpha_0,\beta_0) = \varepsilon\, \mathcal{L}_c^1(\alpha_0^\varepsilon, \beta_0^\varepsilon).$$

The push-forward of a σ^2-subgaussian law by T_ε is (σ^2/ε)-subgaussian, so the standard-deviation proxy σ is replaced by $\sigma/\sqrt{\varepsilon}$. The preceding $\varepsilon = 1$ estimate therefore gives

$$\varepsilon\, C_d \left(1 + \left(\frac{\sigma}{\sqrt{\varepsilon}}\right)^{q_d}\right) n^{-1/2} \leqslant C_d\, \varepsilon \left(1 + \frac{\sigma^{q_d}}{\varepsilon^{b_d}}\right) n^{-1/2},$$

where $b_d = \lceil q_d/2 \rceil = \lceil 5d/4 \rceil + 3$ is the rounded exponent used in [379]. Thus

$$\mathbb{E}\,|\mathcal{L}_c^\varepsilon(\hat{\alpha}_n,\beta) - \mathcal{L}_c^\varepsilon(\alpha,\beta)| \leqslant C_d\, \Lambda_{d,\sigma}(\varepsilon)\, n^{-1/2}.$$

Changing both marginals is handled by the triangle inequality,

$$\left|\mathcal{L}_c^\varepsilon(\hat{\alpha}_n,\hat{\beta}_m) - \mathcal{L}_c^\varepsilon(\alpha,\beta)\right| \leqslant \left|\mathcal{L}_c^\varepsilon(\hat{\alpha}_n,\hat{\beta}_m) - \mathcal{L}_c^\varepsilon(\alpha,\hat{\beta}_m)\right| + \left|\mathcal{L}_c^\varepsilon(\alpha,\hat{\beta}_m) - \mathcal{L}_c^\varepsilon(\alpha,\beta)\right|.$$

The preceding reduction is pathwise: for each realization, it only requires the population and empirical measures appearing in the two terms to share a finite subgaussian proxy. Empirical measures drawn from a subgaussian law have such a random proxy, with moments controlled by the population proxy. Integrating over this proxy gives the Mena–Niles-Weed two-sample extension. The argument does not require the two empirical marginals to have the same sample size, so the two contributions scale as $n^{-1/2}$ and $m^{-1/2}$.

Finally, use the definition

$$\bar{\mathcal{L}}_c^\varepsilon(\alpha,\beta) = \mathcal{L}_c^\varepsilon(\alpha,\beta) - \frac{1}{2}\mathcal{L}_c^\varepsilon(\alpha,\alpha) - \frac{1}{2}\mathcal{L}_c^\varepsilon(\beta,\beta).$$

The cross term is controlled by the two-marginal estimate. For the self term, use the telescoping inequality

$$|\mathcal{L}_c^\varepsilon(\hat\alpha_n,\hat\alpha_n)-\mathcal{L}_c^\varepsilon(\alpha,\alpha)| \leqslant |\mathcal{L}_c^\varepsilon(\hat\alpha_n,\hat\alpha_n)-\mathcal{L}_c^\varepsilon(\alpha,\hat\alpha_n)| + |\mathcal{L}_c^\varepsilon(\alpha,\hat\alpha_n)-\mathcal{L}_c^\varepsilon(\alpha,\alpha)|\,.$$

Each term is a one-marginal perturbation in the same pathwise sense: the other marginal may itself be empirical, but it is part of the common random subgaussian class. Thus the self term $\mathcal{L}_c^\varepsilon(\hat\alpha_n,\hat\alpha_n)-\mathcal{L}_c^\varepsilon(\alpha,\alpha)$ is bounded at the same order as the cross term involving α, and similarly for β. Summing these three contributions gives the displayed estimate, up to changing C_d. □

Sample complexity of estimating OT maps. The preceding estimates concern transport values. In many learning tasks, however, one wants the map itself. The standard sample-based principle is to solve an empirical dual problem for potentials $(\mathrm{f}_{n,m}^\varepsilon,\mathrm{g}_{n,m}^\varepsilon)\in\mathbb{R}^n\times\mathbb{R}^m$, where $\varepsilon>0$ denotes the entropic temperature and $\varepsilon=0$ denotes the unregularized Kantorovich problem. The soft c-transform of Definition 8.21 then acts as an extrapolator from the samples to the full ambient space.

For the quadratic cost $c_2(x,y)=\|x-y\|^2/2$, let $X_i\sim\alpha$, $Y_j\sim\beta$, $\hat\alpha_n=\sum_i a_i\delta_{X_i}$, $\hat\beta_m=\sum_j b_j\delta_{Y_j}$, and let $\mathrm{P}_{n,m}^\varepsilon=(\mathrm{P}_{i,j}^\varepsilon)$ be the optimal empirical entropic coupling. If $\mathrm{g}_{n,m}^\varepsilon$ is the target-side dual potential, and if $\bar c(y,x):=c_2(x,y)$, define

$$u_{n,m}^\varepsilon(x):=(\mathrm{g}_{n,m}^\varepsilon)^{\bar c,\varepsilon}(x)=-\varepsilon\log\sum_j b_j\exp\left(\frac{\mathrm{g}_{n,m,j}^\varepsilon-c_2(x,Y_j)}{\varepsilon}\right),\qquad T_{n,m}^\varepsilon(x):=x-\nabla u_{n,m}^\varepsilon(x).$$

Equivalently, subtracting the quadratic marginal terms turns c_2 into the bilinear cost $c_{\mathrm{ip}}(x,y)=-\langle x,\,y\rangle$. With the sign convention used here, the associated soft convex potential is

$$\varphi_{n,m}^\varepsilon(x):=\frac12\|x\|^2-u_{n,m}^\varepsilon(x)=\varepsilon\log\sum_j b_j\exp\left(\frac{\langle x,\,Y_j\rangle+\mathrm{g}_{n,m,j}^\varepsilon-\frac12\|Y_j\|^2}{\varepsilon}\right),\qquad T_{n,m}^\varepsilon=\nabla\varphi_{n,m}^\varepsilon.$$

This is the precise form of the informal statement that empirical dual potentials are extended by a soft conjugation and then differentiated. It estimates the Brenier map $\alpha\to\beta$; the inverse map $\beta\to\alpha$ is obtained symmetrically from the source-side potential. Differentiating the log-sum-exp gives the explicit half-squared formula

$$T_{n,m}^\varepsilon(x)=\sum_j\omega_j^\varepsilon(x)Y_j,\qquad \omega_j^\varepsilon(x)=\frac{b_j\exp\Big((\mathrm{g}_{n,m,j}^\varepsilon-c_2(x,Y_j))/\varepsilon\Big)}{\sum_k b_k\exp\Big((\mathrm{g}_{n,m,k}^\varepsilon-c_2(x,Y_k))/\varepsilon\Big)}.\tag{10.5}$$

At a sampled source point X_i, the Sinkhorn normalization gives

$$\omega_j^\varepsilon(X_i)=\frac{\mathrm{P}_{i,j}^\varepsilon}{a_i},\qquad T_{n,m}^\varepsilon(X_i)=\frac{1}{a_i}\sum_j\mathrm{P}_{i,j}^\varepsilon Y_j.$$

Thus the extrapolated soft-transform map coincides on the empirical support with the barycentric projection (12.43) of the optimal entropic plan. At $\varepsilon=0$, the hard c-transform is generally nondifferentiable: away from Laguerre boundaries its gradient selects one target site, whereas on a boundary its subdifferential contains several sites. The barycentric projection of an unregularized optimal plan is still well defined on the empirical support, but it need not coincide with a single gradient selection when a row of the plan splits. The following consistency result is therefore stated directly for barycentric projections. It uses uniform empirical weights, as in the usual statistical setting; deterministic nonuniform weights are handled by replacing the grid-entropy estimate in the proof.

Proposition 10.12 (Consistency of empirical barycentric maps)**.** *Let $\Omega\subset\mathbb{R}^d$ be compact, let $\alpha,\beta\in\mathcal{P}_2(\Omega)$, and assume that α is absolutely continuous. Let φ be a Brenier potential from α to β for the cost $c_2(x,y)=\|x-y\|^2/2$, and choose a bounded Borel selection $T(x)\in\partial\varphi(x)$ on Ω. At every differentiability point of φ, this selection necessarily equals $\nabla\varphi(x)$, so it is a representative of the unique Brenier map. Let*

$$\hat\alpha_n=\frac1n\sum_{i=1}^n\delta_{X_i},\qquad \hat\beta_m=\frac1m\sum_{j=1}^m\delta_{Y_j}$$

satisfy $\hat{\alpha}_n \to \alpha$, $\hat{\beta}_m \to \beta$ in $\mathcal{W}_2$. Let $\varepsilon_{n,m} \geqslant 0$ satisfy $\varepsilon_{n,m} \to 0$ and $\varepsilon_{n,m} \log(\min\{n,m\}) \to 0$. For each n, m, let $\mathrm{P}^{\varepsilon}_{n,m}$ be an optimal coupling between $\hat{\alpha}_n$ and $\hat{\beta}_m$, with $\varepsilon = \varepsilon_{n,m}$: unregularized if $\varepsilon = 0$, entropic if $\varepsilon > 0$. Its barycentric projection, defined on the support of $\hat{\alpha}_n$ by

$$\bar{T}^{\varepsilon}_{n,m}(X_i) := n \sum_j \mathrm{P}^{\varepsilon}_{i,j} Y_j,$$

satisfies

$$\int \|\bar{T}^{\varepsilon}_{n,m}(x) - T(x)\|^2 \mathrm{d}\hat{\alpha}_n(x) = \frac{1}{n} \sum_i \|\bar{T}^{\varepsilon}_{n,m}(X_i) - T(X_i)\|^2 \longrightarrow 0$$

as $n, m \to +\infty$. If β is also absolutely continuous, the analogous backward barycentric projections converge in $L^2(\hat{\beta}_m)$ to the Brenier map $S : \beta \to \alpha$.

Proof. The assumptions imply $\mathcal{L}_{c_2}(\hat{\alpha}_n, \hat{\beta}_m) \to \mathcal{L}_{c_2}(\alpha, \beta)$. If $\varepsilon = 0$, $\mathrm{P}^{\varepsilon}_{n,m}$ is exactly optimal. If $\varepsilon > 0$, compare its entropic objective with an unregularized optimal empirical plan $Q_{n,m}$. Regarding $Q_{n,m}$ as the joint law of two uniform discrete indices, its relative entropy with respect to their product law is their mutual information. Hence

$$\mathrm{KL}(Q_{n,m} | \hat{\alpha}_n \otimes \hat{\beta}_m) \leqslant \min\{\log n, \log m\}.$$

It follows that

$$\int c_2 \, \mathrm{dP}^{\varepsilon}_{n,m} \leqslant \mathcal{L}_{c_2}(\hat{\alpha}_n, \hat{\beta}_m) + \varepsilon \log(\min\{n, m\}).$$

Hence the transport costs of $\mathrm{P}^{\varepsilon}_{n,m}$ converge to the population optimum. By compactness, every weak limit point of the plans $\mathrm{P}^{\varepsilon}_{n,m}$ has marginals α, β and optimal quadratic cost. Brenier uniqueness therefore forces the whole sequence to converge weakly to $(\mathrm{Id}, T)_\sharp \alpha$.

Disintegrate $\mathrm{P}^{\varepsilon}_{n,m}$ with respect to its first marginal and denote its barycentric projection by $\bar{T}^{\varepsilon}_{n,m}$. Jensen's inequality gives

$$\int \|\bar{T}^{\varepsilon}_{n,m}(x) - T(x)\|^2 \mathrm{d}\hat{\alpha}_n(x) \leqslant \iint \|y - T(x)\|^2 \mathrm{dP}^{\varepsilon}_{n,m}(x, y).$$

At every differentiability point of φ, the subdifferential is the singleton $\{\nabla\varphi(x)\}$. Since the target is compact, the relevant subgradients are bounded, and the closed-graph property of $\partial\varphi$ implies continuity of the selected gradient at such a point. Convex functions are differentiable Lebesgue-almost everywhere; because α is absolutely continuous, the bounded function $(x, y) \mapsto \|y - T(x)\|^2$ is continuous $(\mathrm{Id}, T)_\sharp \alpha$-almost everywhere. The portmanteau theorem for bounded functions whose discontinuity set has zero limiting mass gives

$$\iint \|y - T(x)\|^2 \mathrm{dP}^{\varepsilon}_{n,m}(x, y) \longrightarrow \int \|T(x) - T(x)\|^2 \mathrm{d}\alpha(x) = 0.$$

The proof for the backward map is identical when the optimal plan is also induced by the Brenier map from β to α. □

The preceding proposition is qualitative. Quantitative rates for maps are substantially more delicate than rates for values, because they require stability of Brenier potentials. Hütter and Rigollet identify the minimax scale for smooth transport maps, while Deb–Ghosal–Sen and Manole–Balakrishnan–Niles-Weed–Wasserman analyze barycentric and density plug-in constructions with suitable out-of-sample extensions [284, 177, 367]. Under their respective compact-support, density, curvature and smoothness hypotheses, the optimal squared $L^2(\alpha)$ risk for an s-smooth map, $s > 1$, has the representative form

$$\mathbb{E}\|\widehat{T}_n - T_0\|^2_{L^2(\alpha)} \lesssim \left(n^{-\frac{2s}{2(s-1)+d}} \vee n^{-1} \right) \quad \text{up to logarithmic factors.}$$

The exact estimator attaining this scale depends on the assumed model: it need not be the raw empirical coupling. The entropic barycentric estimator of Pooladian–Niles-Weed [440] instead emphasizes scalable computation. Under their assumptions (A1)-(A3), if $T_{\varepsilon,(n,n)}$ denotes the entropic barycentric map and $\bar{s} = s \wedge 3$, then for $\varepsilon \leqslant 1$

$$\mathbb{E}\|T_{\varepsilon,(n,n)} - T_0\|^2_{L^2(\alpha)} \lesssim \varepsilon^{-d/2} \log(n) n^{-1/2} + \varepsilon^{(\bar{s}+1)/2} + \varepsilon^2 \mathcal{I}_0(\alpha, \beta),$$

where $\mathcal{I}_0(\alpha,\beta)$ is the Fisher-information quantity appearing in Proposition 8.17. The first term is the available sampling bound for the entropic map, whereas the last two terms are regularization-bias terms that vanish only as $\varepsilon \to 0$. Choosing $\varepsilon \simeq n^{-1/(d+\bar{s}+1)}$ gives

$$\mathbb{E}\|T_{\varepsilon,(n,n)} - T_0\|^2_{L^2(\alpha)} \lesssim (1+\mathcal{I}_0(\alpha,\beta))n^{-\frac{\bar{s}+1}{2(d+\bar{s}+1)}}\log n.$$

This summarizes the main tradeoff: unregularized plug-in constructions combined with suitable smoothing and extension can attain the minimax exponent, whereas $\varepsilon > 0$ gives a simple Sinkhorn-computable soft map with an explicit, though not generally minimax-optimal, finite-sample guarantee.

Remark 10.13 (No free lunch when approximating exact OT). The parametric rate in Proposition 10.11 holds for fixed ε. If the goal is to approximate the unregularized OT value, one must also account for the regularization bias. In a typical bounded-cost finite-dimensional regime,

$$\left|\bar{\mathcal{L}}^{\varepsilon}_c(\alpha,\beta) - \mathcal{L}_c(\alpha,\beta)\right| \leqslant C\varepsilon, \qquad \mathbb{E}\left|\bar{\mathcal{L}}^{\varepsilon}_c(\hat{\alpha}_n,\hat{\beta}_n) - \bar{\mathcal{L}}^{\varepsilon}_c(\alpha,\beta)\right| \leqslant C_{d,\sigma}\varepsilon^{-p_d}n^{-1/2}, \qquad p_d := \left\lceil \frac{5d}{4}\right\rceil + 2,$$

where the second estimate is the small-ε form of the polynomial Mena–Niles-Weed bound. Balancing the two terms gives $\varepsilon \simeq n^{-1/(2(p_d+1))}$ and total error of order $n^{-1/(2(p_d+1))}$ under this conservative bound. Thus entropic smoothing improves the statistical behavior at fixed scale, and the refined analysis avoids an exponential $e^{C/\varepsilon}$ penalty, but approximating exact OT still forces a bias-variance tradeoff whose exponent deteriorates with dimension.

Sliced Wasserstein. For a direction $\theta \in \mathbb{S}^{d-1}$, let $P_\theta(x) = \langle \theta,\, x\rangle$ be the one-dimensional projection and let σ denote the normalized surface measure on the sphere. For $\alpha,\beta \in \mathcal{P}_p(\mathbb{R}^d)$, the sliced Wasserstein distance is

$$\mathrm{S}\mathcal{W}_p(\alpha,\beta)^p := \int_{\mathbb{S}^{d-1}} \mathcal{W}_p\left((P_\theta)_\sharp\alpha, (P_\theta)_\sharp\beta\right)^p \mathrm{d}\sigma(\theta).$$

This construction keeps genuine one-dimensional transport while avoiding a full high-dimensional matching at every evaluation. It is studied in detail in Section 11.2, including its metric properties, Monte-Carlo approximation and variants. Here we only record the statistical consequence: empirical fluctuations are inherited from one-dimensional distribution-function or quantile-process estimates rather than from ambient-dimensional matching. This is one of the main motivations for sliced objectives in generative modeling and statistics [396, 395, 368].

Theorem 10.14 (Dimension-free empirical rate for sliced Wasserstein)**.** *Let $p \geqslant 1$, let $\alpha,\beta \in \mathcal{P}_p(\mathbb{R}^d)$, and let $\hat{\alpha}_n, \hat{\beta}_m$ be independent empirical measures. Assume that all one-dimensional projections satisfy the uniform empirical bounds*

$$\sup_{\theta\in\mathbb{S}^{d-1}} \mathbb{E}\,\mathcal{W}_p\left((P_\theta)_\sharp\hat{\alpha}_n, (P_\theta)_\sharp\alpha\right)^p \leqslant \frac{A_\alpha}{n^{p/2}}, \qquad \sup_{\theta\in\mathbb{S}^{d-1}} \mathbb{E}\,\mathcal{W}_p\left((P_\theta)_\sharp\hat{\beta}_m, (P_\theta)_\sharp\beta\right)^p \leqslant \frac{A_\beta}{m^{p/2}}.$$

Then

$$\mathbb{E}\left|\mathrm{S}\mathcal{W}_p(\hat{\alpha}_n,\hat{\beta}_m) - \mathrm{S}\mathcal{W}_p(\alpha,\beta)\right| \leqslant \frac{A_\alpha^{1/p}}{\sqrt{n}} + \frac{A_\beta^{1/p}}{\sqrt{m}}.$$

In particular, for $p = 1$ and supports contained in a Euclidean ball of radius R, the assumption holds with a constant depending only on R. For $p = 2$, the same $n^{-1/2}$ rate for SW_2 holds whenever the projected laws satisfy a uniform one-dimensional quantile-process bound. A sufficient condition is that the projected laws have positive densities $h_{\alpha,\theta}$, $h_{\beta,\theta}$, with quantile functions $Q_{\alpha,\theta}$, $Q_{\beta,\theta}$, and obey

$$\sup_{\theta\in\mathbb{S}^{d-1}} \int_0^1 \frac{u(1-u)}{h_{\alpha,\theta}(Q_{\alpha,\theta}(u))^2} + \frac{u(1-u)}{h_{\beta,\theta}(Q_{\beta,\theta}(u))^2}\,\mathrm{d}u < +\infty.$$

This is the uniform projected version of the one-dimensional J_2 condition of Bobkov and Ledoux [68].

Proof. The triangle inequality for $\mathrm{S}\mathcal{W}_p$ gives

$$\left|\mathrm{S}\mathcal{W}_p(\hat{\alpha}_n,\hat{\beta}_m) - \mathrm{S}\mathcal{W}_p(\alpha,\beta)\right| \leqslant \mathrm{S}\mathcal{W}_p(\hat{\alpha}_n,\alpha) + \mathrm{S}\mathcal{W}_p(\hat{\beta}_m,\beta).$$

By the definition above, Jensen's inequality and the assumed one-dimensional bounds,

$$\mathbb{E}\,\mathrm{S}\mathcal{W}_p(\hat{\alpha}_n,\alpha) \leqslant \left(\int_{\mathbb{S}^{d-1}} \mathbb{E}\,\mathcal{W}_p\left((P_\theta)_\sharp\hat{\alpha}_n,(P_\theta)_\sharp\alpha\right)^p \mathrm{d}\sigma(\theta)\right)^{1/p} \leqslant \frac{A_\alpha^{1/p}}{\sqrt{n}},$$

and the same argument applies to β. For $p=1$, compact support gives the displayed one-dimensional empirical bound by the Dvoretzky–Kiefer–Wolfowitz inequality and the identity $\mathcal{W}_1(\alpha',\beta') = \int_{\mathbb{R}} |F_{\alpha'} - F_{\beta'}|\mathrm{d}s$. For $p=2$, the displayed J_2 condition gives $\mathbb{E}\,\mathcal{W}_2^2(\hat{\gamma}_n,\gamma) = O(n^{-1})$, uniformly over the projected laws [68]. □

> **Remark 10.15 (Directions are another sample budget).** The theorem concerns the statistical samples used to form $\hat{\alpha}_n$ and $\hat{\beta}_m$. In computation one also replaces the spherical average in the definition of $\mathrm{S}\mathcal{W}_p$ by an empirical average over L random directions. For bounded support, the resulting Monte-Carlo error for $\mathrm{S}\mathcal{W}_p^p$ is of order $L^{-1/2}$, independently of d, because it is just the average of bounded one-dimensional costs. This should not be read as saying that a small number of directions captures full $\mathcal{W}_2$ geometry in high dimension. Even the exact sliced distance is an averaged projected geometry, not a bi-Lipschitz surrogate for $\mathcal{W}_2$ uniformly in dimension; see Proposition 11.25 and the paragraph on intrinsic sliced length in Section 11.2. Moreover, if a discrepancy is visible only inside a spherical cap of angular radius δ, the probability that L random directions hit it is roughly $1-(1-\sigma(\mathrm{cap}_\delta))^L$, and $\sigma(\mathrm{cap}_\delta)$ scales like δ^{d-1} for small caps. For a fixed narrow aperture, it decays exponentially with d. Thus estimating the sliced objective is dimension-friendly, while using random slices to approximate a worst direction, a max-sliced distance, or a proxy meant to behave like full $\mathcal{W}_2$, reintroduces an angular-covering cost. This is the practical no-free-lunch behind sliced, max-sliced and subspace-sliced variants; see also Section 11.2.

10.3 Bias and Variance of OT

The previous section answered a coarse but essential question: how many samples are needed before an empirical OT quantity is accurate, typically up to universal constants. We now ask for a finer statistical description. Rather than only estimating the size of the error, we seek the first expansion of the plug-in estimator. Given empirical laws $\hat{\alpha}_n$ and $\hat{\beta}_m$, and a scalar OT value $\mathcal{V}$, this expansion separates the leading statistical bias

$$B_{n,m}(\mathcal{V}) := \mathbb{E}\mathcal{V}(\hat{\alpha}_n,\hat{\beta}_m) - \mathcal{V}(\alpha,\beta)$$

and the centered fluctuation

$$Z_{n,m}(\mathcal{V}) := \mathcal{V}(\hat{\alpha}_n,\hat{\beta}_m) - \mathbb{E}\mathcal{V}(\hat{\alpha}_n,\hat{\beta}_m),$$

for OT values such as $\mathcal{V}(\alpha,\beta) = \mathcal{W}_p^p(\alpha,\beta)$, $\mathcal{V}(\alpha,\beta) = \mathcal{L}_c(\alpha,\beta)$, or their entropic analogues. Proposition 10.5 describes the linear fluctuations of $\sqrt{n}(\hat{\alpha}_n - \alpha)$ against finitely many test functions. Here that empirical process is pushed through a nonlinear, often nonsmooth, transport value. The resulting law is governed by the local differentiability of the OT functional and by the geometry of its optimal dual potentials.

Literature map. The shape of the limit depends sharply on the analytic regularity of the transport value. For exact OT on a finite or countable space, the map from empirical weights to optimal cost is convex and piecewise affine. Sommerfeld and Munk [496], and the countable-space extensions of Tameling–Sommerfeld–Munk [515], show that the natural limit is a directional delta-method limit: often a support function of a Gaussian process over an optimal dual face, rather than a Gaussian random variable. For Euclidean costs, del Barrio–Loubes and collaborators obtain central-limit theorems for empirical transportation costs under regularity and uniqueness assumptions on the Kantorovich potentials [180, 179]. Entropic regularization makes the problem smoother. For fixed $\varepsilon > 0$, the value is differentiable with respect to the marginals, and the limiting variance is read directly from the entropic dual potentials. This is the point of view developed by Bigot–Cazelles–Papadakis, Klatt–Tameling–Munk, Hundrieser–Klatt–Munk, and Mena–Niles-Weed [56, 307, 283, 379]. More recent results clarify weak limits beyond smooth costs and second-order null limits for Sinkhorn divergences [246, 243]. These asymptotic statements do not replace the dimension-dependent bounds of Section 10.2; they explain what the leading random object is, what constants appear, and where the bias comes from.

Bias versus centered fluctuation. The decomposition

$$\mathcal{V}(\hat{\alpha}_n, \hat{\beta}_m) - \mathcal{V}(\alpha, \beta) = \underbrace{\mathbb{E}\mathcal{V}(\hat{\alpha}_n, \hat{\beta}_m) - \mathcal{V}(\alpha, \beta)}_{\text{bias}} + \underbrace{\mathcal{V}(\hat{\alpha}_n, \hat{\beta}_m) - \mathbb{E}\mathcal{V}(\hat{\alpha}_n, \hat{\beta}_m)}_{\text{centered fluctuation}}$$

is elementary, but in OT it is genuinely diagnostic because the two terms may live on different scales. For the self-distance $\mathcal{W}_p^p(\hat{\alpha}_n, \alpha)$, the population value is zero and the deterministic bias $\mathbb{E}\,\mathcal{W}_p^p(\hat{\alpha}_n, \alpha)$ is the leading term. On a regular d'-dimensional support, the high-dimensional matching scale is $n^{-p/d'}$ when $d' > 2p$, with a critical logarithmic correction when $d' = 2p$. Below this threshold the rate is controlled by empirical-process fluctuations and can be faster; in one dimension, smooth positive densities give the familiar quantile-process asymptotics for $\mathcal{W}_p^p$, of order $n^{-p/2}$ for $p > 1$, and in particular n^{-1} for $\mathcal{W}_2^2$. Thus p/d' is the high-dimensional matching exponent, not a universal rate formula [192, 220, 548]. By contrast, when $\mathcal{V}(\alpha, \beta)$ is nonzero and the OT value is differentiable at (α, β), the centered fluctuation is often governed by an ordinary $n^{-1/2}$ central limit theorem. A "slow OT statistic" can therefore have three distinct causes: a large empirical bias, a nonsmooth directional limit, or a regularization bias such as $\mathcal{L}_c^\varepsilon(\alpha, \beta) - \mathcal{L}_c(\alpha, \beta)$ that still has to be removed.

Proposition 10.16 (Finite-space bias and CLT for exact OT)**.** *Let* $\mathrm{a} \in \Sigma_n$ *and* $\mathrm{b} \in \Sigma_m$ *have positive entries, and let* $\mathrm{C} \in \mathbb{R}^{n \times m}$*. Recall the discrete OT value*

$$\mathrm{L_C}(\mathrm{a}, \mathrm{b}) \coloneqq \min_{\mathrm{P} \in \mathrm{U}(\mathrm{a}, \mathrm{b})} \langle \mathrm{C}, \mathrm{P} \rangle.$$

Let $\widehat{\mathrm{a}}_N$ *be the empirical histogram of* N *independent samples from* a*, while* b *is fixed. Denote by*

$$\mathcal{D}^\star(\mathrm{a}, \mathrm{b}) \coloneqq \operatorname{argmax}_{f_i + g_j \leqslant \mathrm{C}_{i,j}} \langle \mathrm{f}, \mathrm{a} \rangle + \langle \mathrm{g}, \mathrm{b} \rangle$$

the set of optimal dual vectors, modulo the gauge $(\mathrm{f}, \mathrm{g}) \mapsto (\mathrm{f} + \lambda \mathbb{1}, \mathrm{g} - \lambda \mathbb{1})$*. If* G_a *is a centered Gaussian vector with covariance*

$$\mathbb{E}\, G_\mathrm{a} G_\mathrm{a}^\top = \operatorname{diag}(\mathrm{a}) - \mathrm{a}\mathrm{a}^\top,$$

then

$$\sqrt{N}\big(\mathrm{L_C}(\widehat{\mathrm{a}}_N, \mathrm{b}) - \mathrm{L_C}(\mathrm{a}, \mathrm{b})\big) \quad \Longrightarrow \quad \sup_{(\mathrm{f}, \mathrm{g}) \in \mathcal{D}^\star(\mathrm{a}, \mathrm{b})} \langle \mathrm{f}, G_\mathrm{a} \rangle.$$

Moreover,

$$\sqrt{N}\big(\mathbb{E}\mathrm{L_C}(\widehat{\mathrm{a}}_N, \mathrm{b}) - \mathrm{L_C}(\mathrm{a}, \mathrm{b})\big) \longrightarrow \mathbb{E} \sup_{(\mathrm{f}, \mathrm{g}) \in \mathcal{D}^\star(\mathrm{a}, \mathrm{b})} \langle \mathrm{f}, G_\mathrm{a} \rangle,$$

and the rescaled variance converges to the variance of the same limit. If the source dual potential $\mathrm{f}^\star$ *is unique up to constants, the limit is Gaussian with variance*

$$\sigma_{\mathrm{OT}}^2 = \sum_i a_i \left((\mathrm{f}^\star)_i - \sum_k a_k (\mathrm{f}^\star)_k \right)^2.$$

In this differentiable case, the first-order bias vanishes:

$$\mathbb{E}\mathrm{L_C}(\widehat{\mathrm{a}}_N, \mathrm{b}) - \mathrm{L_C}(\mathrm{a}, \mathrm{b}) = o(N^{-1/2}).$$

Proof. The multinomial central limit theorem gives

$$\sqrt{N}(\widehat{\mathrm{a}}_N - \mathrm{a}) \Longrightarrow G_\mathrm{a}, \qquad G_\mathrm{a} \in \{\xi : \langle \mathbb{1}, \xi \rangle = 0\}.$$

The dual formulation writes the map $\mathrm{a}' \mapsto \mathrm{L_C}(\mathrm{a}', \mathrm{b})$ as the supremum, over the feasible polyhedron $f_i + g_j \leqslant \mathrm{C}_{i,j}$, of the affine functions $\mathrm{a}' \mapsto \langle \mathrm{f}, \mathrm{a}' \rangle + \langle \mathrm{g}, \mathrm{b} \rangle$. Its directional derivative at a, in any tangent direction h with $\langle \mathbb{1}, h \rangle = 0$, is therefore

$$D_\mathrm{a}\mathrm{L_C}(\mathrm{a}, \mathrm{b})[h] = \sup_{(\mathrm{f}, \mathrm{g}) \in \mathcal{D}^\star(\mathrm{a}, \mathrm{b})} \langle \mathrm{f}, h \rangle,$$

which is the finite-dimensional form of Danskin's theorem already used in Proposition 12.28. The directional delta method gives the displayed distributional limit. After fixing a gauge, the relevant dual

face is bounded, so the support functions above grow at most linearly in the multinomial fluctuation. Uniform integrability then yields convergence of the first two moments, hence the bias and variance statements. If $f^\star$ is unique modulo constants, the support function reduces to the linear form $\langle f^\star, G_a\rangle$. Its expectation is zero, and the covariance $\mathrm{diag}(a) - aa^\top$ gives exactly the displayed centered second moment of $f^\star$. □

The same argument gives the independent two-sample limit. If $N/(N+M) \to \lambda \in (0,1)$, then, with $r_{N,M} = \sqrt{NM/(N+M)}$,

$$r_{N,M}\big(L_C(\widehat{a}_N, \widehat{b}_M) - L_C(a,b)\big) \Longrightarrow \sup_{(f,g)\in\mathcal{D}^\star(a,b)} \left(\sqrt{1-\lambda}\,\langle f, G_a\rangle + \sqrt{\lambda}\,\langle g, G_b\rangle\right),$$

where G_a and G_b are independent multinomial Gaussian limits. When both dual potentials are unique, the first-order bias is negligible and the asymptotic variance of the unscaled estimator is

$$\frac{1}{N}\sum_i a_i(f_i^\star - \bar{f}^\star)^2 + \frac{1}{M}\sum_j b_j(g_j^\star - \bar{g}^\star)^2, \qquad \bar{f}^\star = \sum_i a_i f_i^\star, \quad \bar{g}^\star = \sum_j b_j g_j^\star.$$

If the optimal face contains several dual potentials, the limit is the supremum of a Gaussian process over that face and need not itself be Gaussian. The leading bias is then the expectation of this supremum divided by the square-root sample size. This nonsmoothness is the basic reason why inference for exact OT can remain delicate even on a finite space.

Proposition 10.17 (Finite-space bias and CLT for entropic OT)**.** *Fix $\varepsilon > 0$, positive histograms a, b, and a finite cost matrix C. Consider the finite-dimensional KL-normalized entropic value*

$$\mathcal{V}_{C,\varepsilon}(a,b) := \min_{P\in U(a,b)} \langle C, P\rangle + \varepsilon \mathrm{KL}(P|a\otimes b).$$

This is the discrete version of the value $\mathcal{V}_{c,\varepsilon}$ used in Proposition 12.29. Let $(f_\varepsilon, g_\varepsilon)$ be normalized entropic dual potentials for $\mathcal{V}_{C,\varepsilon}(a,b)$. If $\widehat{a}_N$ is the empirical histogram of N samples from a, then

$$\sqrt{N}\big(\mathcal{V}_{C,\varepsilon}(\widehat{a}_N, b) - \mathcal{V}_{C,\varepsilon}(a,b)\big) \quad \Longrightarrow \quad \mathcal{N}(0, \sigma^2_{\varepsilon,a}),$$

and

$$\mathbb{E}\mathcal{V}_{C,\varepsilon}(\widehat{a}_N, b) - \mathcal{V}_{C,\varepsilon}(a,b) = O(N^{-1}), \qquad N\ \mathrm{Var}\left(\mathcal{V}_{C,\varepsilon}(\widehat{a}_N, b)\right) \to \sigma^2_{\varepsilon,a},$$

where

$$\sigma^2_{\varepsilon,a} = \sum_i a_i \left((f_\varepsilon)_i - \sum_k a_k (f_\varepsilon)_k\right)^2.$$

For two independent empirical histograms, the first-order asymptotic variance is

$$\frac{\sigma^2_{\varepsilon,a}}{N} + \frac{\sigma^2_{\varepsilon,b}}{M}, \qquad \sigma^2_{\varepsilon,b} = \sum_j b_j \left((g_\varepsilon)_j - \sum_\ell b_\ell (g_\varepsilon)_\ell\right)^2.$$

Proof. For positive histograms and fixed $\varepsilon > 0$, the entropic problem has a unique optimizer, and the KL-normalized value is smooth on a neighborhood of (a,b) inside the relative interiors of the simplices. Proposition 12.29 gives

$$D_a \mathcal{V}_{C,\varepsilon}(a,b)[h] = \langle f_\varepsilon, h\rangle.$$

Applying the ordinary delta method to the multinomial CLT $\sqrt{N}(\widehat{a}_N - a) \Rightarrow G_a$ gives the one-sample limit $\langle f_\varepsilon, G_a\rangle$, whose variance is the displayed covariance formula. To justify the bias expansion, fix a relative-interior neighborhood U of a on which the Hessian is bounded. A multinomial concentration bound gives $\mathbb{P}(\widehat{a}_N \notin U) \leqslant Ce^{-cN}$. On U, Taylor's formula has a centered linear term and a quadratic remainder bounded by $C\|\widehat{a}_N - a\|^2$, whose expectation is $O(N^{-1})$. On the exponentially unlikely complement, the entropic value remains bounded by a constant depending only on C and ε; the natural definition on a boundary face simply discards zero-weight atoms. This proves the $O(N^{-1})$ bias. The same expansion, together with bounded moments of the multinomial fluctuation, gives convergence of the rescaled variance. The two-sample formula follows by applying the argument independently to the two marginals. □

For the entropy-only convention $\mathrm{L}_{\mathrm{C}}^{\varepsilon}$ used in Section 8.1, one has, for positive histograms,

$$\mathrm{L}_{\mathrm{C}}^{\varepsilon}(\mathrm{a},\mathrm{b}) = \mathcal{V}_{\mathrm{C},\varepsilon}(\mathrm{a},\mathrm{b}) - \varepsilon \mathrm{H}(\mathrm{a}) - \varepsilon \mathrm{H}(\mathrm{b}).$$

Thus the marginal entropy derivatives described after Proposition 12.29 must be added to the potentials: the derivative with respect to the source marginal is represented on the simplex tangent space by $\mathrm{f}_{\varepsilon} + \varepsilon \log \mathrm{a}$, up to an irrelevant additive constant. For the debiased Sinkhorn divergence $\bar{\mathcal{L}}_c^{\varepsilon}$, the derivative is instead the difference between cross and self entropic potentials. At the null $\alpha = \beta$, this first derivative can vanish for the debiased statistic, and second-order limits then become relevant [243]. This is a useful warning: the asymptotic law is determined not only by a rate, but also by the local geometry of the functional.

Three error terms when entropy estimates exact OT. If the target is the exact cost $\mathcal{L}_c(\alpha,\beta)$ but the statistic uses the entropic value $\mathcal{L}_c^{\varepsilon}$, three errors should be kept separate:

$$\mathcal{L}_c^{\varepsilon}(\hat{\alpha}_n,\hat{\beta}_m) - \mathcal{L}_c(\alpha,\beta) = \underbrace{\mathcal{L}_c^{\varepsilon}(\alpha,\beta) - \mathcal{L}_c(\alpha,\beta)}_{\text{regularization bias}} + \underbrace{B_{n,m}(\mathcal{L}_c^{\varepsilon})}_{\text{statistical bias}} + \underbrace{Z_{n,m}(\mathcal{L}_c^{\varepsilon})}_{\text{centered fluctuation}} .$$

The same decomposition applies to the debiased value $\bar{\mathcal{L}}_c^{\varepsilon}$ by replacing $\mathcal{L}_c^{\varepsilon}$ everywhere by $\bar{\mathcal{L}}_c^{\varepsilon}$. Although algebraic, this decomposition is statistically important and is usually read after choosing $\varepsilon = \varepsilon_n$. For fixed ε, the statistical bias is typically $O(n^{-1} + m^{-1})$ in finite dimension, and the centered fluctuation is $O_{\mathbb{P}}(n^{-1/2} + m^{-1/2})$; boundary events where an empirical finite histogram has a zero entry have exponentially small probability when the population weights are positive. The regularization bias disappears only when $\varepsilon \to 0$, whereas the variance constants usually deteriorate as $\varepsilon \downarrow 0$. This is the asymptotic form of the bias–variance tradeoff illustrated in Figure 10.2.

What changes in continuous spaces. The finite-dimensional formulas above are not merely toy models; they are the cleanest template for the general mechanism. Whenever an OT value is Hadamard differentiable and admits a unique sufficiently regular dual potential $f^{\star}$, the empirical-process CLT and the delta method give a first-order Gaussian limit with one-sample variance $\mathrm{Var}_{\alpha}(f^{\star}(X))$. When differentiability fails, as can happen for exact OT because the dual optimizer is not unique or because one studies the degenerate self-distance $\mathcal{L}_c(\alpha,\alpha) = 0$, the limit may be non-Gaussian or may occur at a slower nonparametric scale. Entropic regularization smooths the dual potentials and typically restores a conventional first-order CLT for fixed ε, even under weak assumptions on the cost. If the target is unregularized OT, however, this statistical expansion must still be balanced against the bias $\mathcal{L}_c^{\varepsilon}(\alpha,\beta) - \mathcal{L}_c(\alpha,\beta)$, or against its debiased analogue. Thus Section 10.2 gives dimension-dependent magnitudes, while the present section identifies the local asymptotic constants and covariance formulas needed for inference.

10.4 Sketching Sinkhorn in Linear Time

The statistical results above explain why entropic OT is attractive at fixed temperature, but they also expose a computational tension. Sampling the marginals replaces the input laws by empirical measures and reduces the problem to finitely many support points; subsampling or coarsening these supports is governed by the sample-complexity estimates of this chapter. Even after the supports are fixed, however, a Sinkhorn sweep costs $O(nm)$, because it repeatedly applies the Gibbs kernel to vectors. Kernel sketching attacks this second bottleneck: it keeps the sampled marginals fixed and samples, or otherwise compresses, the feature representation of the kernel used by Sinkhorn. With a rank R factorization, N_{it} scaling sweeps cost $O(N_{\mathrm{it}}(n+m)R)$, which is linear in the support sizes only when the required rank and iteration count are controlled. The approximation rank generally depends on the ambient or intrinsic dimension, the temperature, and the target accuracy.

PSD kernels and ordinary feature sketches. The RKHS/MMD section, Section 7.2, already used positive semidefinite kernels to define Hilbertian discrepancies between measures. Recall that a symmetric kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is positive semidefinite if, for every finite family $(x_i)_{i=1}^n$, the Gram matrix $(k(x_i,x_j))_{i,j}$ is positive semidefinite. Many scalable kernel methods start from an integral feature representation

$$k(x,y) = \int_Z \langle \varphi(x,z),\, \varphi(y,z) \rangle_{\mathbb{R}^p}\, d\xi(z), \qquad \varphi : \mathcal{X} \times Z \to \mathbb{R}^p,$$

where ξ is a latent probability measure. Drawing $z_1, \dots, z_r \sim \xi$ gives the Monte-Carlo sketch

$$\widetilde{k}_r(x, y) = \frac{1}{r} \sum_{\ell=1}^{r} \langle \varphi(x, z_\ell),\, \varphi(y, z_\ell) \rangle_{\mathbb{R}^p}.$$

This is the classical random-feature principle. For data points $(x_i)_{i=1}^n$, the dense Gram matrix $K = (k(x_i, x_j))_{i,j}$ is replaced by $\widetilde{K} = \Phi\Phi^\top$, where $\Phi \in \mathbb{R}^{n \times rp}$ is the sampled feature matrix with entries $\Phi_{i,(\ell,q)} = r^{-1/2}\varphi_q(x_i, z_\ell)$, $1 \leqslant \ell \leqslant r$, $1 \leqslant q \leqslant p$. The number r of sampled latent variables is thus a second sample size, distinct from the number of data points. At this stage no positivity of φ is assumed; signed or oscillatory features are exactly what make Fourier sketches useful.

For translation-invariant kernels on $\mathbb{R}^d$, Bochner's theorem gives the standard Fourier sketch. If $k(x, y) = \kappa(x - y)$ and $\kappa(t) = \int e^{\mathrm{i}\langle \omega,\, t \rangle}\, d\Lambda(\omega)$, with Λ a positive spectral measure, one may use real features $\varphi(x, \omega) = (\cos\langle \omega,\, x \rangle, \sin\langle \omega,\, x \rangle)$. For the Gaussian kernel $k_\sigma(x, y) = \exp(-\|x - y\|^2/(2\sigma^2))$, this corresponds to $\omega \sim \mathcal{N}(0, \sigma^{-2}\mathrm{Id})$. This random Fourier feature construction was introduced by Rahimi and Recht to accelerate large-scale kernel machines [447]; its statistical role for kernel ridge regression and related supervised methods is analyzed, for instance, in [458, 25].

Application to Sinkhorn kernels. Let

$$\alpha = \sum_{i=1}^{n} a_i \delta_{x_i}, \qquad \beta = \sum_{j=1}^{m} b_j \delta_{y_j}, \qquad K_{i,j} = k_\varepsilon(x_i, y_j) \coloneqq e^{-c(x_i, y_j)/\varepsilon}.$$

Assume here that the weights are positive and that $K_{i,j} > 0$. The Sinkhorn scaling equations are

$$u = \frac{a}{Kv}, \qquad v = \frac{b}{K^\top u},$$

where divisions are componentwise. The factorization used by Sinkhorn is rectangular, because the source points x_i and target points y_j need not coincide. Starting from the same sketch $k(x, y) \simeq r^{-1} \sum_{\ell=1}^{r} \langle \varphi(x, z_\ell),\, \varphi(y, z_\ell) \rangle_{\mathbb{R}^p}$, set $R = rp$ and flatten the pair (ℓ, q) into a single column index. Define

$$(\Phi_X)_{i,(\ell,q)} = r^{-1/2}\varphi_q(x_i, z_\ell), \qquad (\Phi_Y)_{j,(\ell,q)} = r^{-1/2}\varphi_q(y_j, z_\ell).$$

Then $\Phi_X \in \mathbb{R}^{n \times R}$, $\Phi_Y \in \mathbb{R}^{m \times R}$, and

$$\widetilde{K}_{i,j} = (\Phi_X \Phi_Y^\top)_{i,j} = \frac{1}{r} \sum_{\ell=1}^{r} \langle \varphi(x_i, z_\ell),\, \varphi(y_j, z_\ell) \rangle_{\mathbb{R}^p} \simeq K_{i,j}.$$

When the two supports are the same, one has $\Phi_X = \Phi_Y = \Phi$ and this reduces to the symmetric Gram approximation $\widetilde{K} = \Phi\Phi^\top$. In the rectangular case, if K is replaced by the rank-R factorization $\widetilde{K} = \Phi_X \Phi_Y^\top$, the two matrix-vector products are evaluated as

$$\widetilde{K}v = \Phi_X(\Phi_Y^\top v), \qquad \widetilde{K}^\top u = \Phi_Y(\Phi_X^\top u),$$

so each scaling sweep costs $O((n + m)R)$ operations and stores only $O((n + m)R)$ numbers. The difficulty is that Sinkhorn is not a generic kernel method: it divides by $\widetilde{K}v$ and, in dual variables, applies logarithms. Thus one needs $\widetilde{K}_{i,j} > 0$, preferably with a relative error on the entries, because $-\varepsilon \log K_{i,j}$ is the effective cost. This is a more stringent requirement than the usual spectral or Frobenius approximation of a PSD Gram matrix.

The next proposition is not a substitute for the sharp low-rank Sinkhorn theory. It is an elementary concentration lemma whose role is to isolate the two facts that matter for Sinkhorn: positivity of the approximate kernel and accuracy after taking logarithms. The full algorithmic statements rely on more refined Nyström or Gaussian-kernel approximations together with stability of Sinkhorn scaling [6, 5].

Proposition 10.18 (Finite positivity and logarithmic accuracy)**.** *Let $\varepsilon > 0$ and $K_{i,j} = \mathbb{E}h_{i,j}(Z) > 0$ for $1 \leqslant i \leqslant n$, $1 \leqslant j \leqslant m$, where Z has law ξ and $|h_{i,j}(Z)| \leqslant M$ almost surely. Set $c_{i,j} = -\varepsilon \log K_{i,j}$, and let*

$$\widetilde{K}_{i,j} = \frac{1}{r} \sum_{\ell=1}^{r} h_{i,j}(Z_\ell), \qquad Z_1, \dots, Z_r \sim \xi.$$

Then, for every $\eta > 0$,

$$\mathbb{P}\left(\max_{i,j}\left|\widetilde{K}_{i,j} - K_{i,j}\right| > \eta\right) \leqslant 2nm \exp\left(-\frac{r\eta^2}{2M^2}\right).$$

In particular, if $\kappa := \min_{i,j} K_{i,j} > 0$, $\eta < \kappa$, and the event above does not occur, then $\widetilde{K}$ is entrywise positive. If moreover $\eta \leqslant \kappa/2$, then the sketched cost $\widetilde{c}_{i,j} = -\varepsilon \log \widetilde{K}_{i,j}$ satisfies

$$\max_{i,j}\left|\widetilde{c}_{i,j} - c_{i,j}\right| \leqslant \frac{2\varepsilon\eta}{\kappa}.$$

If, more sharply, the relative variables $\zeta_{i,j}(Z) := h_{i,j}(Z)/K_{i,j}$ satisfy $|\zeta_{i,j}(Z)| \leqslant \psi$ almost surely, then for every $0 < \delta \leqslant 1/2$,

$$\mathbb{P}\left(\max_{i,j}\left|\frac{\widetilde{K}_{i,j}}{K_{i,j}} - 1\right| > \delta\right) \leqslant 2nm \exp\left(-\frac{r\delta^2}{2\psi^2}\right).$$

On the complementary event,

$$\widetilde{K}_{i,j} > 0 \quad \textit{and} \quad \max_{i,j}\left|\widetilde{c}_{i,j} - c_{i,j}\right| \leqslant 2\varepsilon\delta.$$

Proof. For a fixed pair (i, j), Hoeffding's inequality applied to bounded variables in $[-M, M]$ gives

$$\mathbb{P}\left(\left|\widetilde{K}_{i,j} - K_{i,j}\right| > \eta\right) \leqslant 2\exp\left(-\frac{r\eta^2}{2M^2}\right).$$

A union bound over the nm entries gives the first claim. If $\max_{i,j}\left|\widetilde{K}_{i,j} - K_{i,j}\right| \leqslant \eta < \kappa$, then $\widetilde{K}_{i,j} \geqslant \kappa - \eta > 0$. Writing $\widetilde{K}_{i,j} = K_{i,j}(1 + s_{i,j})$, one has $|s_{i,j}| \leqslant \eta/\kappa \leqslant 1/2$, hence $|\log(1 + s_{i,j})| \leqslant 2|s_{i,j}|$, which proves the absolute logarithmic bound. Applying the same Hoeffding argument to $\zeta_{i,j}(Z)$, whose expectation is 1, gives the relative estimate. On the corresponding event, $\widetilde{K}_{i,j} = K_{i,j}(1 + s_{i,j})$ with $|s_{i,j}| \leqslant \delta \leqslant 1/2$, so positivity and the bound $\varepsilon|\log(1 + s_{i,j})| \leqslant 2\varepsilon\delta$ follow. □

This proposition is deliberately elementary, but it pinpoints the obstruction. When ε is small or the data diameter is large, $\kappa = \min K_{i,j}$ can be exponentially small, so an ordinary signed sketch may need a very large r just to keep all entries positive. What Sinkhorn really needs is relative control of the kernel, equivalently additive control of the sketched costs. Positive features remove the most basic failure mode, negative entries in $\widetilde{K}$, but they do not by themselves provide relative concentration. In particular, the log-normal features introduced below are unbounded, so this Hoeffding proposition does not apply to them without truncation or a different concentration argument.

The available worst-case bounds retain an explicit dimension dependence. For the quadratic cost on points contained in a radius-R ball of $\mathbb{R}^d$, the Nyström analysis of Altschuler, Bach, Rudi and Niles-Weed [5] studies $K_{i,j} = \exp(-\eta\|x_i - x_j\|^2)$ and bounds the effective rank by a quantity of the form

$$r_*(X, \eta, \varepsilon_0) \lesssim \left(\eta R^2 + \log(n/\varepsilon_0)\right)^d$$

up to universal constants and harmless logarithms. Thus the resulting Sinkhorn computation is near-linear in n once this rank is fixed, but the required rank is polynomial in the inverse kernel bandwidth, with exponent d and only logarithmic dependence on n in this bound. When the same machinery is used to approximate unregularized quadratic OT to additive accuracy τ, one takes η of order $\tau^{-1}\log n$, so the complexity contains a factor $\tau^{-O(d)}$, often summarized informally as an ε^{-d}-type dependence on the target precision. The same work also explains why this can be much better on data with lower intrinsic dimension.

Positive features and total positivity. A positive feature sketch asks for more than positive semidefiniteness. The relevant matrix and kernel cones are the following.

Definition 10.19 (Doubly nonnegative and completely positive matrices and kernels)**.** In finite dimension, set

$$\mathrm{DNN}_n = \left\{A \in \mathbb{R}^{n\times n} \,;\, A \succeq 0,\ A_{i,j} \geqslant 0\right\}, \qquad \mathrm{CP}_n = \left\{BB^\top \,;\, B \in \mathbb{R}_+^{n\times q} \text{ for some } q \geqslant 1\right\}.$$

Matrices in these sets are called doubly nonnegative and completely positive, respectively. A kernel

is doubly positive, respectively completely positive, if every finite Gram matrix belongs to DNN_n, respectively CP_n. It is a positive-feature kernel if

$$k(x,y) = \int_Z \langle \varphi(x,z),\, \varphi(y,z)\rangle_{\mathbb{R}^p} \mathrm{d}\xi(z) \quad \text{with} \quad \varphi(x,z) \in \mathbb{R}^p_+.$$

One always has $\mathrm{CP}_n \subset \mathrm{DNN}_n$, with equality for $n \leqslant 4$ and strict inclusion for $n \geqslant 5$ [45, 88]. Checking membership in the completely positive cone is computationally hard: strong membership is NP-hard, and weak membership for the completely positive cone and its copositive dual is NP-hard [188]. This is why certifying positive sketchability is substantially harder than checking that a kernel is PSD and pointwise nonnegative.

For kernels, doubly positive means equivalently PSD and pointwise nonnegative. The book also uses *totally positive* for completely positive kernels admitting positive features. This terminology should not be confused with Karlin's classical total positivity, which concerns nonnegativity of all minors of an ordered kernel [297]; it is a completely-positive-type condition, studied for instance in [178].

The positive-feature construction of Scetbon and Cuturi [473] makes the reason for this definition direct. If

$$k(x,y) = \int_Z \langle \varphi(x,z),\, \varphi(y,z)\rangle_{\mathbb{R}^p}\, d\xi(z), \qquad \varphi(x,z) \in \mathbb{R}^p_+,$$

then every finite Gram matrix is a conic combination of rank-one matrices $vv^\top$ with $v \geqslant 0$, and hence is completely positive. Conversely, on a fixed finite set, $K \in \mathrm{CP}_n$ is exactly the existence of a nonnegative factorization $K = BB^\top$. Thus total positivity in this sense is precisely the kernel-level condition behind positive random features. For rectangular source–target matrices, one applies the same statement to the Gram matrix on the union of the sampled source and target points and then extracts its $n \times m$ cross block.

Proposition 10.20 (Basic closure of totally positive kernels)**.** *Nonnegative sums and pointwise products of totally positive kernels are totally positive. The same holds for measurable nonnegative mixtures whenever the resulting kernel is finite. In particular, positive autocorrelations*

$$k(x,y) = \int g(u-x)g(u-y)\, du, \qquad g \geqslant 0,$$

and all nonnegative mixtures of such kernels are totally positive.

Proof. For sums, concatenate the nonnegative features and rescale them by the square roots of the nonnegative coefficients. For products, tensor the nonnegative features:

$$\langle \varphi_1(x,z_1),\, \varphi_1(y,z_1)\rangle \langle \varphi_2(x,z_2),\, \varphi_2(y,z_2)\rangle = \langle \varphi_1(x,z_1)\otimes\varphi_2(x,z_2),\, \varphi_1(y,z_1)\otimes\varphi_2(y,z_2)\rangle.$$

The tensor feature remains entrywise nonnegative. A measurable mixture is obtained by adjoining the mixing parameter to the latent variable; equivalently, on each finite point set it is a limit of conic combinations in the closed completely positive cone. The autocorrelation formula is already a positive feature representation with latent variable u and scalar feature $g(u-x)$. □

The Gaussian Gibbs kernels used by Sinkhorn admit an especially simple positive feature representation, which is the construction exploited by Scetbon and Cuturi [473]. The same construction extends to the generalized Gaussian kernels $e^{-\|x-y\|^p/\varepsilon}$ for $0 < p \leqslant 2$, because radial stable kernels are mixtures of Gaussian kernels. This is the Schoenberg–Bernstein mechanism behind the positive definiteness of radial stable kernels [479, 43].

Proposition 10.21 (Generalized Gaussian positive features)**.** *Let $0 < p \leqslant 2$, $\varepsilon > 0$, and*

$$k_{p,\varepsilon}(x,y) = \exp\left(-\frac{\|x-y\|^p}{\varepsilon}\right), \qquad x,y \in \mathbb{R}^d.$$

Set $a = p/2 \in (0,1]$. Let S_a be the positive a-stable random variable normalized by

$$\mathbb{E}\left(e^{-tS_a}\right) = e^{-t^a}, \qquad t \geqslant 0,$$

and set $\Lambda = \varepsilon^{-1/a} S_a$. Denote by $\nu_{a,\varepsilon}$ the law of Λ. Then

$$\exp\left(-\frac{t^a}{\varepsilon}\right) = \int_0^{+\infty} e^{-\lambda t}\, d\nu_{a,\varepsilon}(\lambda), \qquad t \geqslant 0.$$

Equivalently, for a nonnegative random variable $\Lambda \sim \nu_{a,\varepsilon}$,

$$k_{p,\varepsilon}(x,y) = \mathbb{E}_\Lambda \exp\left(-\Lambda \|x-y\|^2\right).$$

Fix a centering point $x_0 \in \mathbb{R}^d$, let $Z \sim \mathcal{N}(0, \mathrm{Id}_d)$, independent of Λ, and set

$$q_x(\Lambda) = \sqrt{2\Lambda}\,(x - x_0), \qquad \varphi_p(x,\Lambda,Z) = \exp\left(\langle Z,\, q_x(\Lambda)\rangle - \|q_x(\Lambda)\|^2\right).$$

Then

$$k_{p,\varepsilon}(x,y) = \mathbb{E}_{\Lambda,Z}\left(\varphi_p(x,\Lambda,Z)\varphi_p(y,\Lambda,Z)\right).$$

Thus $k_{p,\varepsilon}$ is totally positive in the completely positive sense. In particular, if $(\Lambda_\ell, Z_\ell)_{\ell=1}^r$ are i.i.d. copies of (Λ, Z), then

$$\widetilde{k}_{p,\varepsilon}(x,y) = \sum_{\ell=1}^r \varphi_\ell(x)\varphi_\ell(y), \qquad \varphi_\ell(x) = r^{-1/2}\varphi_p(x,\Lambda_\ell,Z_\ell),$$

is an unbiased nonnegative feature sketch of $k_{p,\varepsilon}$.

For $p = 2$, the mixing law reduces to the Dirac mass $\nu_{1,\varepsilon} = \delta_{1/\varepsilon}$, which recovers the usual Gaussian feature formula.

Proof. The function $s \mapsto e^{-s^a}$ is completely monotone for $0 < a \leqslant 1$. Bernstein's theorem therefore gives the Laplace-transform representation of a positive a-stable random variable S_a. The scaling $\Lambda = \varepsilon^{-1/a} S_a$ gives

$$\mathbb{E} e^{-t\Lambda} = \exp\left(-\frac{t^a}{\varepsilon}\right),$$

which is the displayed representation by $\nu_{a,\varepsilon}$. Applying this identity with $t = \|x-y\|^2$ gives the Gaussian-mixture formula.

Conditionally on $\Lambda = \lambda$, the Gaussian moment-generating function gives

$$\begin{aligned}\mathbb{E}_Z\left(\varphi_p(x,\lambda,Z)\varphi_p(y,\lambda,Z)\right) &= \exp\left(-\|q_x(\lambda)\|^2 - \|q_y(\lambda)\|^2\right)\mathbb{E}_Z\left(\exp\langle Z,\, q_x(\lambda) + q_y(\lambda)\rangle\right)\\ &= \exp\left(-\frac{1}{2}\|q_x(\lambda) - q_y(\lambda)\|^2\right) = \exp\left(-\lambda\|x-y\|^2\right).\end{aligned}$$

Taking the expectation over Λ gives $k_{p,\varepsilon}$. Since φ_p is positive, this is a positive-feature representation; hence every finite Gram matrix is completely positive. The Monte-Carlo sketch is just the empirical average of this representation. □

Remark 10.22 (The range $0 < p \leqslant 2$). The restriction $0 < p \leqslant 2$ is essential. Schoenberg's theorem says that $x \mapsto e^{-\gamma\|x\|^p}$ is positive definite on every Euclidean space precisely in this range. For $p > 2$, the generalized Gaussian kernel is not positive definite in general: there are finite Euclidean point clouds for which the Gram matrix has a negative eigenvalue. Thus no universal positive-feature factorization of the above type should be expected. Constants in the cost, such as using $\|x-y\|^p/p$, are absorbed by rescaling ε.

For $p < 2$, the positive stable mixing variable has heavy tails, so unbiasedness does not imply a useful Monte-Carlo variance. Indeed, for any $x \neq x_0$, the diagonal estimator is $\varphi_p(x,\Lambda,Z)^2$, has mean $k_{p,\varepsilon}(x,x) = 1$, but

$$\mathbb{E}_Z \varphi_p(x,\Lambda,Z)^4 = \exp\left(8\Lambda\|x-x_0\|^2\right).$$

The last quantity has infinite expectation over a positive stable Λ when $p < 2$. Practical sketches therefore require variance reduction, truncation, deterministic quadrature, or a different positive representation. The $p = 2$ Gaussian case avoids the stable mixing tail because $\Lambda = 1/\varepsilon$ is deterministic.

Choosing x_0 near the data can reduce the dynamic range of the features without changing the kernel. More generally, Gaussian scale mixtures give many radial totally positive kernels, including standard covariance families used in Gaussian processes [449].

The gap between doubly positive and totally positive kernels is real. On the circle, consider

$$k_\lambda(x,y) = \lambda + \cos^2\left(\frac{x-y}{2}\right), \qquad \lambda > 0.$$

This kernel is PSD and strictly positive pointwise, since $\cos^2(t/2) = (1+\cos t)/2$ has nonnegative Fourier coefficients and λ adds a positive constant mode. For $\lambda = 1/20$, however, its Gram matrix on five equally spaced points is not completely positive. The Horn copositive matrix H satisfies $\langle H, B\rangle \geqslant 0$ for every completely positive matrix B, but the five-point Gram matrix K gives

$$\langle H, K\rangle = \frac{21}{4} - \frac{5\sqrt{5}}{2} < 0.$$

Thus $K \in \mathrm{DNN}_5 \setminus \mathrm{CP}_5$, and the kernel is doubly positive but not totally positive. Figure 10.3 displays this finite certificate.

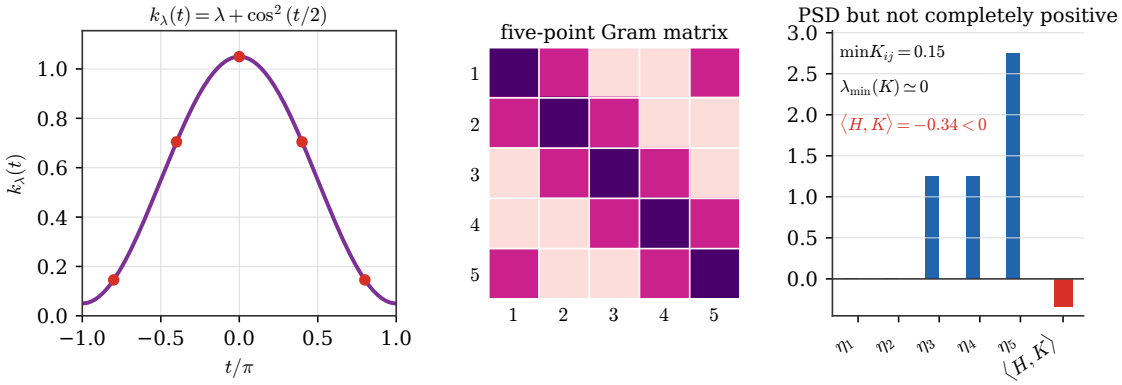


Figure 10.3: A doubly positive kernel that is not totally positive. The periodic kernel $k_\lambda(t) = \lambda + \cos^2(t/2)$, with $\lambda = 1/20$, is positive semidefinite and pointwise positive. On five equally spaced points, its Gram matrix is entrywise positive and PSD, but the Horn copositive certificate gives $\langle H, K\rangle < 0$, so this finite restriction is not completely positive and cannot admit a nonnegative Gram factorization.

Positive sketches for Sinkhorn. For the p-power cost $c_p(x,y) = \|x-y\|^p$, with $0 < p \leqslant 2$, the Sinkhorn Gibbs kernel is the generalized Gaussian kernel

$$k_{p,\varepsilon}(x,y) = \exp\left(-\frac{\|x-y\|^p}{\varepsilon}\right).$$

For $p \geqslant 1$, this is the usual p-Wasserstein power cost; for $0 < p < 1$, it should be read simply as a concave power transport cost. The factor $1/p$, often used in the definition of $\mathcal{W}_p^p$, only rescales ε. Proposition 10.21 gives a positive factorization of this kernel. With r draws $(\Lambda_\ell, Z_\ell) \sim \nu_{p/2,\varepsilon} \otimes \mathcal{N}(0, \mathrm{Id}_d)$, define the normalized features

$$\varphi_\ell(x) = r^{-1/2}\varphi_p(x, \Lambda_\ell, Z_\ell), \qquad 1 \leqslant \ell \leqslant r.$$

They satisfy

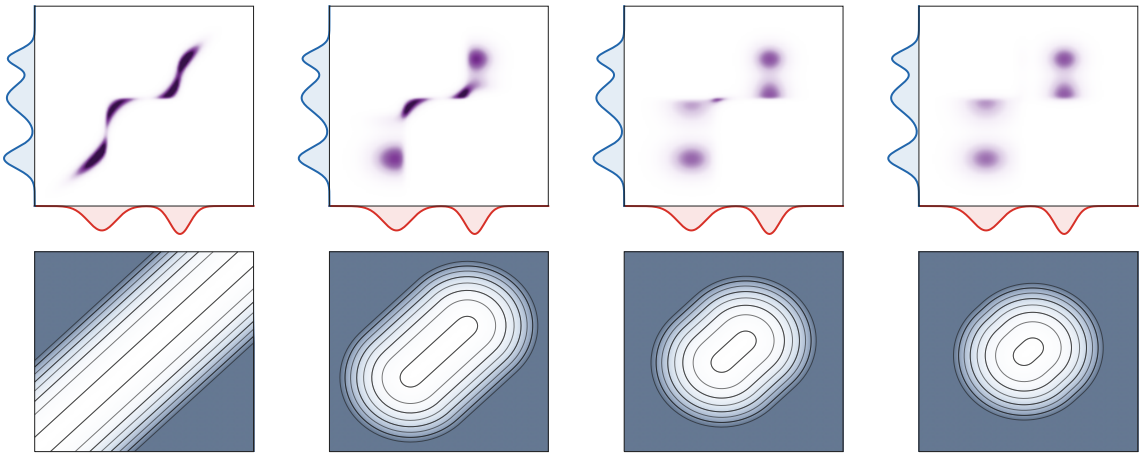
$$\widetilde{k}_{p,\varepsilon}(x,y) = \sum_{\ell=1}^{r} \varphi_\ell(x)\varphi_\ell(y) \geqslant 0, \qquad \mathbb{E}[\widetilde{k}_{p,\varepsilon}(x,y)] = k_{p,\varepsilon}(x,y).$$

The corresponding matrices $(\Phi_X)_{i,\ell} = \varphi_\ell(x_i)$ and $(\Phi_Y)_{j,\ell} = \varphi_\ell(y_j)$ give a positive factorization $\widetilde{K} = \Phi_X\Phi_Y^\top$. Sinkhorn can therefore be run safely on the sketched kernel, and the resulting plan is represented as

$$\widetilde{P} = \mathrm{diag}(u)\Phi_X\Phi_Y^\top \mathrm{diag}(v) = LR^\top, \qquad L = \mathrm{diag}(u)\Phi_X, \quad R = \mathrm{diag}(v)\Phi_Y.$$

The output is the exact entropic coupling for the sketched effective cost $\widetilde{c}_{p,\varepsilon} = -\varepsilon \log \widetilde{k}_{p,\varepsilon}$, and it approximates the original $\mathcal{W}_p$-type entropic problem only insofar as this logarithmic kernel approximation is accurate. This kernel-sketching route is complementary to worst-case near-linear Sinkhorn analyses [7], low-rank Gaussian-kernel approximations for quadratic transport [6], and factored-coupling models in which the coupling itself is constrained to have low rank [474].

Figure 10.4 shows the object that is changed by this approximation. The displayed figure uses the $p = 2$ Gaussian Gibbs kernel, so the exact effective cost is $c(x, y) = |x - y|^2$. A sufficiently rich positive-feature sketch produces both a visually close sketched plan and a faithful approximation of this quadratic cost in logarithmic scale. Coarser sketches still enforce the prescribed marginals after Sinkhorn scaling, but their effective costs become rounded wells rather than parallel quadratic level sets, which explains the loss of geometric concentration in the coupling.



exact dense kernel positive sketch, $r = 40$ positive sketch, $r = 10$ positive sketch, $r = 3$

Figure 10.4: Positive-feature sketches of a Sinkhorn plan. The top row shows entropic couplings between one-dimensional Gaussian mixtures, with red and blue marginal curves around each coupling image. The bottom row shows the matching effective costs $\widetilde{c}_\varepsilon(x, y) = -\varepsilon \log \widetilde{k}_\varepsilon(x, y)$, rendered as images with black level sets. The left column is the dense kernel, whose effective cost is $|x - y|^2$. The other columns solve Sinkhorn with positive features for this Gaussian special case of ranks 40, 10, and 3. The rank-40 sketch remains close to the dense plan and to the quadratic cost, while lower ranks preserve the marginals but blur the transport geometry because the logarithmic kernel approximation has become too coarse. The displayed sketch uses Gaussian quadrature for reproducibility; random positive features use the same nonnegative factorization identity.

Connection with linear attention. The same algebra appears in transformer attention, studied later from a continuous-depth transport viewpoint in Section 16.4. A softmax attention matrix has entries proportional to $\exp(\langle q_i, k_j \rangle)$; linear-attention methods replace this positive kernel by features $\Phi(q_i)^\top \Phi(k_j)$, so that attention can be applied as $\Phi(Q)(\Phi(K)^\top V)$ rather than by forming the full $n \times n$ matrix [299, 144, 545, 554]. Sinkhorn sketching is the transport analogue of this idea: replace the Gibbs matrix by a positive feature factorization, keep only the scaling vectors and feature factors, and control the approximation in the logarithmic scale relevant to entropic potentials. The analogy is useful because both settings ask for scalable positive kernel algebra. The transport case is more constrained, however, since the approximate kernel is reused inside nonlinear normalizations until the prescribed marginals are reached for the sketched problem.

CHAPTER 11

Generalized Wasserstein Distances

This chapter keeps the idea of comparing measures while changing the geometry of the comparison. The constructions below relax mass conservation, average lower-dimensional projections, quotient nuisance symmetries, linearize transport around a reference measure, replace the trace cost by spectral gauges, or constrain motion to conditional fibers. They are useful when standard $\mathcal{W}_p$ is too rigid or too expensive, but each modification also changes which metric, geodesic, or stability properties survive.

11.1 Unbalanced OT

Unbalanced OT allows mass creation and destruction by penalizing marginal mismatch. It is essential when histograms are not normalized, when observations contain outliers, or when only part of the source should match the target [340, 141, 143].

Relaxed formulation. The relaxed coupling problem replaces exact marginal constraints by divergence penalties.

Definition 11.1 (Unbalanced optimal transport)**.** For nonnegative measures $(\alpha,\beta)\in\mathcal{M}_+(\mathcal{X})\times\mathcal{M}_+(\mathcal{Y})$, a nonnegative measurable cost c, and proper lower-semicontinuous convex entropy functions $\psi_s:[0,+\infty)\to[0,+\infty]$ satisfying $\psi_s(1)=0$, define

$$\mathrm{UW}_c(\alpha,\beta):=\inf_{\pi\in\mathcal{M}_+(\mathcal{X}\times\mathcal{Y})}\int_{\mathcal{X}\times\mathcal{Y}}c(x,y)\mathrm{d}\pi(x,y)+\mathcal{D}_{\psi_1}(\pi_1|\alpha)+\mathcal{D}_{\psi_2}(\pi_2|\beta), \tag{11.1}$$

Exact conservation $(\pi_1,\pi_2)=(\alpha,\beta)$ is replaced by a cost for changing the marginals. Writing $\psi_s=\tau\bar{\psi}_s$ exposes the relaxation scale:

$$\mathrm{UW}_{c,\tau}(\alpha,\beta)=\inf_{\pi\geqslant 0}\int c\mathrm{d}\pi+\tau\mathcal{D}_{\bar{\psi}_1}(\pi_1|\alpha)+\tau\mathcal{D}_{\bar{\psi}_2}(\pi_2|\beta).$$

As τ increases, marginal mismatch becomes expensive and the relaxed value approaches balanced OT whenever the hard marginal constraints are feasible. Small τ makes creation and destruction cheap; after division by τ, the diagonal zero-transport part reveals the pure divergence geometry.

Marginal relaxation changes which measures are compared, but not the transport problem once those modified marginals are fixed. To state this precisely, extend the Kantorovich value (3.7) by positive homogeneity: for finite measures $(\widetilde{\alpha},\widetilde{\beta})$ with common mass M,

$$\mathcal{L}_c(\widetilde{\alpha},\widetilde{\beta}):=\begin{cases}M\mathcal{L}_c(\widetilde{\alpha}/M,\widetilde{\beta}/M), & M>0,\\ 0, & M=0.\end{cases} \tag{11.2}$$

Proposition 11.2 (Balanced transport between relaxed marginals)**.** *Under the hypotheses of Definition 11.1,*

$$\mathrm{UW}_c(\alpha,\beta)=\inf_{\substack{\widetilde{\alpha}\in\mathcal{M}_+(\mathcal{X}),\ \widetilde{\beta}\in\mathcal{M}_+(\mathcal{Y})\\ \widetilde{\alpha}(\mathcal{X})=\widetilde{\beta}(\mathcal{Y})}}\left\{\mathcal{L}_c(\widetilde{\alpha},\widetilde{\beta})+\mathcal{D}_{\psi_1}(\widetilde{\alpha}|\alpha)+\mathcal{D}_{\psi_2}(\widetilde{\beta}|\beta)\right\}. \tag{11.3}$$

A plan π^* solves the unbalanced problem if and only if its marginals (π_1^*,π_2^*) solve (11.3) and π^* is an optimal balanced coupling between them.

Proof. Every plan π has equal-mass marginals $(\widetilde{\alpha},\widetilde{\beta})=(\pi_1,\pi_2)$. For fixed marginals, the divergence terms are constant and minimizing the remaining transport term gives (11.2). Grouping the plans in (11.1) by their marginals proves (11.3). The same argument shows that an optimizer must be transport-optimal conditionally on its marginals and that these marginals minimize the outer problem. Conversely, an optimal balanced coupling between any minimizing pair in (11.3) attains the unbalanced value. □

KL mass–shape separation. Proposition 11.2 separates transport from marginal relaxation for every φ-divergence. KL penalties allow one further simplification: the common mass of the relaxed marginals separates from their normalized shapes. A scalar optimization then determines how much mass is transported, while a probability-level transport problem determines how it is arranged.

Let $A = \alpha(\mathcal{X}) > 0$, $B = \beta(\mathcal{Y}) > 0$, and write the normalized input shapes as $\widehat{\alpha} = \alpha/A$ and $\widehat{\beta} = \beta/B$.

Proposition 11.3 (Exact mass–shape separation for KL relaxation)**.** *Assume $\psi_1 = \tau_0\varphi_{\mathrm{KL}}$ and $\psi_2 = \tau_1\varphi_{\mathrm{KL}}$, where φ_{KL} is defined in (7.9) and $\tau_0, \tau_1 > 0$. Set $T = \tau_0 + \tau_1$, $\theta_i = \tau_i/T$, and*

$$\mathcal{E}_{c,\tau_0,\tau_1}(\widehat{\alpha},\widehat{\beta}) := \inf_{\substack{\widetilde{\alpha}\in\mathcal{P}(\mathcal{X})\\ \widetilde{\beta}\in\mathcal{P}(\mathcal{Y})}} \left\{ \mathcal{L}_c(\widetilde{\alpha},\widetilde{\beta}) + \tau_0\,\mathrm{KL}(\widetilde{\alpha}|\widehat{\alpha}) + \tau_1\,\mathrm{KL}(\widetilde{\beta}|\widehat{\beta}) \right\}. \tag{11.4}$$

Then, with $e^{-\infty} = 0$,

$$\mathrm{UW}_c(\alpha,\beta) = \tau_0 A + \tau_1 B - T A^{\theta_0} B^{\theta_1} \exp\left(-\frac{\mathcal{E}_{c,\tau_0,\tau_1}(\widehat{\alpha},\widehat{\beta})}{T}\right). \tag{11.5}$$

If $(\widetilde{\alpha}^*, \widetilde{\beta}^*)$ attains the finite shape envelope and $\bar{\pi}^*$ is an optimal probability coupling between these marginals, then $\pi^* = M^*\bar{\pi}^*$ is optimal, where

$$M^* = A^{\theta_0} B^{\theta_1} \exp\left(-\frac{\mathcal{E}_{c,\tau_0,\tau_1}(\widehat{\alpha},\widehat{\beta})}{T}\right) \tag{11.6}$$

Proof. Write every nonzero plan as $\pi = M\bar{\pi}$, where $M = \pi(\mathcal{X}\times\mathcal{Y}) > 0$ and $\bar{\pi}$ is a probability coupling with marginals $(\widetilde{\alpha}, \widetilde{\beta})$. The KL homogeneity identity

$$\mathrm{KL}(M\widetilde{\alpha}|A\widehat{\alpha}) = M\,\mathrm{KL}(\widetilde{\alpha}|\widehat{\alpha}) + M\log(M/A) - M + A \tag{11.7}$$

holds in the extended-value sense, and similarly for the second marginal. Minimizing first over $\bar{\pi}$ and its normalized marginals therefore reduces the objective to

$$M\mathcal{E}_{c,\tau_0,\tau_1}(\widehat{\alpha},\widehat{\beta}) + \tau_0\big(M\log(M/A) - M + A\big) + \tau_1\big(M\log(M/B) - M + B\big).$$

When the shape value is finite, this strictly convex function of M has the unique minimizer (11.6); substitution gives (11.5). If the shape value is infinite, every nonzero plan has infinite objective and the same formula returns the value $\tau_0 A + \tau_1 B$ of the zero plan. □

The two immediate numerical displays make the penalty roles explicit. Figure 11.1 fixes a KL marginal penalty and varies τ: the transported marginals, shown in violet, are allowed to differ from the prescribed red and blue marginals, and the gaps are precisely the created or destroyed mass.

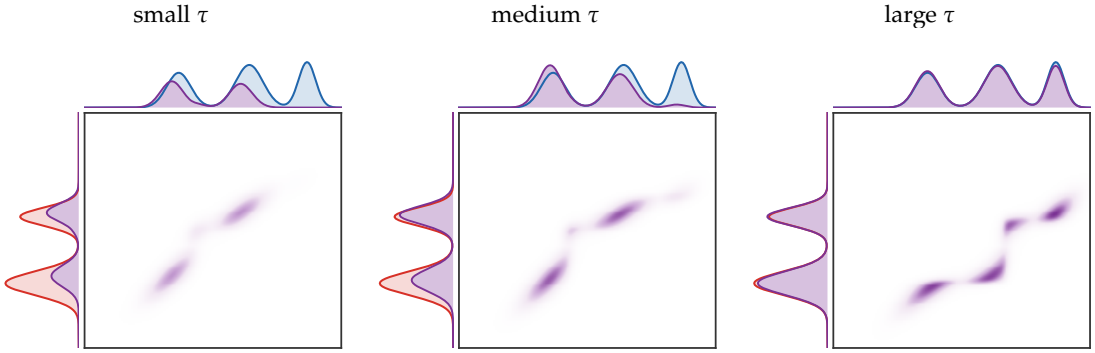


Figure 11.1: KL unbalanced OT on one-dimensional Gaussian-mixture densities. The central matrix is the transported coupling. On the left, the red curve is the prescribed source marginal and the violet curve is the transported source marginal; the red gap is destroyed mass. On the top, the blue curve is the prescribed target marginal and the violet curve is the transported target marginal; the blue gap is created mass. Increasing τ makes marginal mismatch more expensive, so more mass is transported, including toward the far right target mode.

The entropy used in the marginal relaxation also changes the qualitative behavior. A KL penalty leads to smooth multiplicative rescaling. The reverse-KL, or Burg, penalty blows up when a transported

marginal vanishes where the prescribed marginal is positive, so it discourages complete deletion of small modes. Total variation has a linear kink and behaves closer to partial transport: mass is either kept active or created and destroyed at nearly constant marginal price.

Example 11.4 (Application to proliferating and dying cell populations). In single-cell time courses, the observed populations can grow, die or change composition between two sampling times. A balanced coupling would force every unit of mass at time t to reappear at time $t + \Delta t$, which is too rigid biologically. The relaxed value

$$\inf_{\pi \geqslant 0} \int c \mathrm{d}\pi + \tau_1 \mathcal{D}(\pi_1 | \alpha_t) + \tau_2 \mathcal{D}(\pi_2 | \alpha_{t+\Delta t})$$

where π_1, π_2 are the marginals of π, interprets missing marginal mass as death or undersampling, and excess target marginal mass as growth or birth. Neural semi-coupling and entropic dynamic variants use this idea to model proliferating or dying cell populations from unpaired measurements [356, 308]. The dynamic counterpart, where creation and transport are both part of the tangent dynamics, is discussed in Section 15.7.

Figure 11.2 now keeps the geometry, entropic plan regularization and relaxation strength fixed, and changes only the marginal divergence. This isolates the effect of the penalty: KL gives smooth rescaling, Burg discourages complete deletion of prescribed modes, while total variation produces sharper active-mass selection.

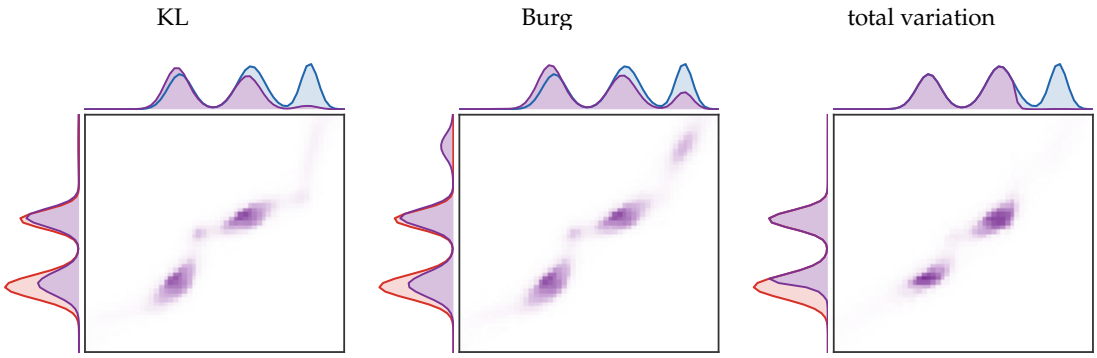


Figure 11.2: Effect of the marginal divergence in unbalanced entropic OT. The geometric cost, entropic plan regularization ε, and relaxation strength τ are fixed; only the marginal penalty changes. KL allows smooth mass variation, Burg keeps transported marginals from vanishing on prescribed modes, and total variation gives a sharper active-mass selection. The side plots use the same convention as Figure 11.1.

These figures should be read as pictures of a single relaxed plan π: the same nonnegative measure determines both the transported coupling and its two relaxed marginals. The small-τ result below formalizes the opposite regime, where transport becomes negligible compared with local mass variation.

Proposition 11.5 (Small-transport-scale limit for marginal penalties)**.** *Assume that α, β are finite measures on a compact metric space $\mathcal{X}$, that c is continuous, $c \geqslant 0$, and $c(x, y) = 0$ if and only if $x = y$. Assume also that the marginal divergences are nonnegative, weak-* lower semicontinuous, and have weak-* compact sublevel sets on $\mathcal{M}_+(\mathcal{X})$. Then*

$$\lim_{\tau \downarrow 0} \frac{1}{\tau} \mathrm{UW}_{c,\tau}(\alpha, \beta) = \inf_{\chi \in \mathcal{M}_+(\mathcal{X})} \mathcal{D}_{\bar{\psi}_1}(\chi | \alpha) + \mathcal{D}_{\bar{\psi}_2}(\chi | \beta).$$

The right-hand side is the infimal gluing divergence obtained by matching the two measures through a common zero-transport marginal χ. In the dominated case, if $\alpha = a\lambda$, $\beta = b\lambda$, and the minimizing common marginal is absolutely continuous, $\chi = r\lambda$, this divergence decouples pointwise as

$$\int \mathfrak{m}_{\bar{\psi}_1, \bar{\psi}_2}(a(x), b(x)) \mathrm{d}\lambda(x), \qquad \mathfrak{m}_{\bar{\psi}_1, \bar{\psi}_2}(a, b) := \inf_{r \geqslant 0} a\, \bar{\psi}_1(r/a) + b\, \bar{\psi}_2(r/b),$$

with the usual recession conventions when $a = 0$ or $b = 0$. For superlinear entropies, and in particular for KL, finite energy forces this dominated form. Thus, when $\bar{\psi}_1 = \bar{\psi}_2$ is the KL entropy,

$$\inf_{\chi \in \mathcal{M}_+(\mathcal{X})} \mathrm{KL}(\chi | \alpha) + \mathrm{KL}(\chi | \beta) = \int (\sqrt{a} - \sqrt{b})^2 \mathrm{d}\lambda.$$

Thus the KL marginal relaxation contains the squared Hellinger distance as its local mass-variation limit.

Proof. For the upper bound, restrict to diagonal plans $\pi = (\mathrm{Id}, \mathrm{Id})_\sharp \chi$, whose transport cost is zero and whose two marginals are both χ. This gives the desired upper bound after optimizing over χ.

For the lower bound, let $\tau_n \downarrow 0$ and let π_n be almost minimizing plans with bounded scaled values $\tau_n^{-1} \mathrm{UW}_{c,\tau_n}(\alpha, \beta)$. Since the divergences are nonnegative, $\int c \mathrm{d}\pi_n = O(\tau_n)$, hence $\int c \mathrm{d}\pi_n \to 0$. The bounded scaled values also put the two marginals in compact divergence sublevel sets. Since a coupling has the same total mass as each marginal, the couplings are tight on $\mathcal{X} \times \mathcal{X}$. Up to subsequences, $\pi_n \rightharpoonup \pi_0$. Lower semicontinuity of the transport cost yields $\int c \mathrm{d}\pi_0 = 0$, so π_0 is concentrated on the diagonal. Its two marginals are therefore equal to a common measure χ. Lower semicontinuity of the marginal divergences gives

$$\liminf_n \frac{1}{\tau_n} \mathrm{UW}_{c,\tau_n}(\alpha, \beta) \geqslant \mathcal{D}_{\bar{\psi}_1}(\chi|\alpha) + \mathcal{D}_{\bar{\psi}_2}(\chi|\beta),$$

and optimizing over χ gives the lower bound.

In the dominated case, writing $\chi = r\lambda$ gives

$$\mathcal{D}_{\bar{\psi}_1}(\chi|\alpha) + \mathcal{D}_{\bar{\psi}_2}(\chi|\beta) = \int a\, \bar{\psi}_1(r/a) + b\, \bar{\psi}_2(r/b) \mathrm{d}\lambda,$$

so the minimization over χ decouples into the scalar envelope $\mathfrak{m}_{\bar{\psi}_1,\bar{\psi}_2}$. For KL, no singular part is admissible when α and β are dominated by λ. The pointwise objective is $r \log(r/a) - r + a + r \log(r/b) - r + b$. Its optimality condition is $\log(r/a) + \log(r/b) = 0$, hence $r = \sqrt{ab}$, and the minimum is $a + b - 2\sqrt{ab} = (\sqrt{a} - \sqrt{b})^2$. □

Proposition 11.6 (Dual of unbalanced optimal transport)**.** *Under the usual Fenchel–Rockafellar qualification assumptions, one has equality between* (11.1) *and*

$$\mathrm{UW}_c(\alpha, \beta) = \sup_{f \oplus g \leqslant c} \; - \mathcal{D}^*_{\psi_1}(-f|\alpha) - \mathcal{D}^*_{\psi_2}(-g|\beta).$$

Proof. Use the variational formula (7.13) for the dual of a divergence and introduce the marginal variables through continuous potentials:

$$\inf_{\pi \geqslant 0} \sup_{f,g} \int c \mathrm{d}\pi + \int -f \mathrm{d}\pi_1 + \int -g \mathrm{d}\pi_2 - \mathcal{D}^*_{\psi_1}(-f|\alpha) - \mathcal{D}^*_{\psi_2}(-g|\beta).$$

Exchanging the infimum and the supremum gives

$$\sup_{f,g} -\mathcal{D}^*_{\psi_1}(-f|\alpha) - \mathcal{D}^*_{\psi_2}(-g|\beta) + \inf_{\pi \geqslant 0} \int \big(c - (f \oplus g)\big) \mathrm{d}\pi.$$

The last infimum is 0 when $f \oplus g \leqslant c$ and $-\infty$ otherwise, which gives the displayed dual. □

Reverse and homogeneous formulations. The Liero–Mielke–Savaré formulation rewrites marginal penalties as a local transport cost and then homogenizes it. Assuming first that the reference measures and transported marginals have mutually absolutely continuous parts, one can factor the objective as

$$\begin{aligned} &\int c(x,y) \mathrm{d}\pi(x,y) + \mathcal{D}_{\psi_1}(\pi_1|\alpha) + \mathcal{D}_{\psi_2}(\pi_2|\beta) \\ &\quad = \int \left(c(x,y) + \psi_1\left(\frac{\mathrm{d}\pi_1}{\mathrm{d}\alpha}(x)\right) \frac{\mathrm{d}\alpha}{\mathrm{d}\pi_1}(x) + \psi_2\left(\frac{\mathrm{d}\pi_2}{\mathrm{d}\beta}(y)\right) \frac{\mathrm{d}\beta}{\mathrm{d}\pi_2}(y) \right) \mathrm{d}\pi(x,y). \end{aligned}$$

This motivates the local reverse cost

$$L_c(r,s) \coloneqq c + r\psi_1(1/r) + s\psi_2(1/s), \tag{11.8}$$

with the usual recession convention at $r = 0$ or $s = 0$. If $\alpha = F\pi_1 + \alpha^\perp$ and $\beta = G\pi_2 + \beta^\perp$ are the Lebesgue decompositions of the reference marginals with respect to the transported marginals, then the reverse formulation reads

$$\mathrm{UW}_c(\alpha, \beta) = \inf_{\pi \geqslant 0} \int L_{c(x,y)}(F(x), G(y)) \mathrm{d}\pi(x,y) + \psi_1(0)\alpha^\perp(\mathcal{X}) + \psi_2(0)\beta^\perp(\mathcal{Y}).$$

The homogeneous formulation is obtained by taking the marginal perspective of L_c,

$$H_c(r,s) := \inf_{\theta>0} \theta L_c(r/\theta, s/\theta), \tag{11.9}$$

which is positively 1-homogeneous. It defines

$$\mathrm{HW}_c(\alpha,\beta) = \inf_{\pi\geqslant 0} \int H_{c(x,y)}(F(x),G(y))\mathrm{d}\pi(x,y) + \psi_1(0)\alpha^\perp(\mathcal{X}) + \psi_2(0)\beta^\perp(\mathcal{Y}). \tag{11.10}$$

For the cone construction, and also for the proof of equivalence, it is useful to make explicit the semi-coupling form of the homogeneous value [340]. Here one uses a base measure λ on $\mathcal{X}\times\mathcal{Y}$ and two nonnegative weights u, v attached to each transported pair:

$$\mathrm{HW}_c(\alpha,\beta) = \inf_{\substack{\lambda\in\mathcal{M}_+(\mathcal{X}\times\mathcal{Y}),\ u,v\geqslant 0 \\ (\mathrm{p}_1)_\sharp(u\lambda)=\alpha,\ (\mathrm{p}_2)_\sharp(v\lambda)=\beta}} \int H_{c(x,y)}\big(u(x,y),v(x,y)\big)\mathrm{d}\lambda(x,y). \tag{11.11}$$

The constraints mean, for instance, that $\int \varphi(x)u(x,y)\mathrm{d}\lambda(x,y) = \int \varphi \mathrm{d}\alpha$ for every bounded Borel test function φ on $\mathcal{X}$, and similarly for β. The cases $u = 0$ or $v = 0$ encode the recession terms $\psi_1(0)$ and $\psi_2(0)$: they correspond to mass that is created or destroyed rather than transported.

Proposition 11.7 (Homogenization does not change the unbalanced cost)**.** *Assume that the entropy functions are proper, lower semicontinuous and convex, with the usual recession extensions at zero. Then*

$$\mathrm{HW}_c(\alpha,\beta) = \mathrm{UW}_c(\alpha,\beta).$$

Proof. Since $H_c(r,s) \leqslant L_c(r,s)$ by choosing the scale equal to one, every competitor in the reverse formulation gives a homogeneous competitor of no larger cost. Hence $\mathrm{HW}_c \leqslant \mathrm{UW}_c$.

For the converse, use the semi-coupling representation (11.11) and fix a feasible triple (λ, u, v). For a positive measurable scale $\theta(x,y)$, set $\tilde\pi = \theta\lambda$. Disintegrate λ with respect to its first spatial marginal. If $\bar u(x)$ and $\bar\theta(x)$ denote the corresponding conditional averages of u and θ, then $\alpha = \bar u\,(\mathrm{p}_1)_\sharp\lambda$ and $\tilde\pi_1 = \bar\theta\,(\mathrm{p}_1)_\sharp\lambda$. Convexity of the perspective $(a,m)\mapsto a\psi_1(m/a)$ and conditional Jensen give

$$\mathcal{D}_{\psi_1}(\tilde\pi_1|\alpha) \leqslant \int u\,\psi_1(\theta/u)\mathrm{d}\lambda.$$

The same argument on the second marginal yields the analogous bound with (v,ψ_2). Therefore

$$\int c\mathrm{d}\tilde\pi + \mathcal{D}_{\psi_1}(\tilde\pi_1|\alpha) + \mathcal{D}_{\psi_2}(\tilde\pi_2|\beta) \leqslant \int \big[c\theta + u\psi_1(\theta/u) + v\psi_2(\theta/v)\big]\mathrm{d}\lambda.$$

The pointwise infimum of the integrand over $\theta > 0$ is precisely $H_c(u,v)$. A measurable η-minimizing choice of θ exists because this is a lower-semicontinuous normal integrand; the recession conventions cover $u = 0$ or $v = 0$. Letting $\eta\downarrow 0$ and then minimizing over (λ,u,v) gives $\mathrm{UW}_c \leqslant \mathrm{HW}_c$. □

Conic lifting. Assume now that $\mathcal{X} = \mathcal{Y}$ and $\psi_1 = \psi_2 = \psi$. The homogeneous formulation naturally lifts mass to a radial coordinate.

Definition 11.8 (Cone lift)**.** The cone over $\mathcal{X}$ is $\mathfrak{C}[\mathcal{X}] := (\mathcal{X}\times\mathbb{R}_+)/\!\sim$, where all $(x,0)$ form one apex. For $p \geqslant 1$, its ground cost is

$$\Delta((x,r),(y,s)) := H_{c(x,y)}(r^p, s^p)^{1/p}.$$

Several classical unbalanced geometries are obtained by choosing ψ, c and p so that Δ is a distance on the cone:

- $\mathcal{D}_\psi = \mathrm{KL}$, $p = 2$, and $c(x,y) = -\log\cos^2(d(x,y)\wedge\pi/2)$ give the Hellinger–Kantorovich or Wasserstein–Fisher–Rao cone metric

$$\Delta((x,r),(y,s))^2 = r^2 + s^2 - 2rs\cos(d(x,y)\wedge\pi/2).$$

- $\mathcal{D}_\psi = \mathrm{KL}$, $p = 2$, and $c(x,y) = d(x,y)^2$ give the Gaussian Hellinger formula

$$\Delta((x,r),(y,s))^2 = r^2 + s^2 - 2rse^{-d(x,y)^2/2}.$$

  It is a cone metric whenever the Gaussian kernel $k(x,y) = e^{-d(x,y)^2/2}$ is positive definite. This holds, for example, when $(\mathcal{X}, d)$ is a subset of a Hilbert space: if $k(x,y) = \langle \Phi(x), \Phi(y)\rangle$ with $\|\Phi(x)\| = 1$, then the displayed quantity is $\|r\Phi(x) - s\Phi(y)\|^2$. For a general metric space, positive definiteness of this kernel is an additional hypothesis.

- $\mathcal{D}_\psi = \mathrm{TV}$, $p = 1$, and $c(x,y) = d(x,y)$ give the partial-transport cone cost

$$\Delta((x,r),(y,s)) = r + s - (r \wedge s)(2 - d(x,y))_+.$$

For a finite measure η on the cone, define its weighted base projection $\mathsf{P}_p\eta \in \mathcal{M}_+(\mathcal{X})$ by

$$\int_{\mathcal{X}} \varphi(x)\mathrm{d}(\mathsf{P}_p\eta)(x) = \int_{\mathfrak{C}[\mathcal{X}]} \varphi(x) r^p \mathrm{d}\eta(x,r)$$

for every bounded Borel φ. This is well defined at the apex because the weight r^p vanishes there. The corresponding cone value is

$$\mathrm{CW}(\alpha,\beta) = \inf_{\gamma \in \mathcal{M}_+(\mathfrak{C}[\mathcal{X}]^2)} \left\{ \int \Delta((x,r),(y,s))^p \mathrm{d}\gamma \;;\; \mathsf{P}_p\gamma_1 = \alpha, \quad \mathsf{P}_p\gamma_2 = \beta \right\}.$$

The cosine cone gives the intrinsic metric most commonly used for simultaneous transport and mass variation.

Definition 11.9 (Wasserstein–Fisher–Rao distance)**.** For $p = 2$ and

$$\Delta((x,r),(y,s))^2 = r^2 + s^2 - 2rs\cos(d(x,y) \wedge \pi/2),$$

the Wasserstein–Fisher–Rao, or Hellinger–Kantorovich, distance is $\mathrm{WFR}(\alpha,\beta) := \mathrm{CW}(\alpha,\beta)^{1/2}$.

In the following theorem, CW denotes the p-action value on the cone. Thus, when the cone cost comes from a distance Δ, the actual metric is $\mathrm{CW}^{1/p}$.

Theorem 11.10 (Cone formulation of unbalanced OT)**.** *Assume that $\mathcal{X}$ is a compact metric space, that $(r,s) \mapsto H_{c(x,y)}(r,s)$ is convex for every (x,y), and that $(x,y,r,s) \mapsto H_{c(x,y)}(r,s)$ is lower semicontinuous. Then*

$$\mathrm{UW}_c(\alpha,\beta) = \mathrm{HW}_c(\alpha,\beta) = \mathrm{CW}(\alpha,\beta).$$

If, in addition, Δ is a continuous distance on $\mathfrak{C}[\mathcal{X}]$, then $\mathrm{CW}^{1/p}$ is a distance between nonnegative finite measures.

Proof. The equality UW = HW is Proposition 11.7. It remains to compare the semi-coupling representation (11.11) with the cone value.

Let (λ, u, v) be feasible in (11.11). Define the cone coupling

$$\gamma = \big((x,y) \mapsto ((x, u(x,y)^{1/p}), (y, v(x,y)^{1/p}))\big)_\sharp \lambda.$$

Then $\mathsf{P}_p\gamma_1 = \alpha$ and $\mathsf{P}_p\gamma_2 = \beta$ by the weighted marginal constraints. Moreover

$$\int \Delta((x,r),(y,s))^p \mathrm{d}\gamma = \int H_{c(x,y)}(u(x,y), v(x,y))\mathrm{d}\lambda(x,y).$$

Taking the infimum gives $\mathrm{CW} \leqslant \mathrm{HW}$.

Conversely, let γ be feasible for CW. Choose arbitrary representatives of cone points at the apex. This does not affect the weighted constraints, and it does not change the cost because, by the recession convention, $H_{c(x,y)}(0,s)$ and $H_{c(x,y)}(r,0)$ do not depend on the arbitrary spatial representative attached

to the zero radius. Let λ be the projection of γ onto the spatial variables (x, y), and disintegrate $\gamma = \gamma_{x,y} \mathrm{d}\lambda(x, y)$ with respect to this projection. Set

$$u(x, y) = \int r^p \mathrm{d}\gamma_{x,y}(r, s), \qquad v(x, y) = \int s^p \mathrm{d}\gamma_{x,y}(r, s).$$

The cone marginal constraints imply $(\mathrm{p}_1)_\sharp(u\lambda) = \alpha$ and $(\mathrm{p}_2)_\sharp(v\lambda) = \beta$. Since H_c is convex in its two arguments, as a perspective of the convex reverse cost L_c, Jensen's inequality gives, for λ-a.e. (x, y),

$$H_{c(x,y)}(u(x, y), v(x, y)) \leqslant \int H_{c(x,y)}(r^p, s^p) \mathrm{d}\gamma_{x,y}(r, s).$$

After integration and using $\Delta((x, r), (y, s))^p = H_{c(x,y)}(r^p, s^p)$, this yields HW $\leqslant$ CW. Hence UW = HW = CW.

Assume now that Δ is a distance. The nontrivial point in the triangle inequality is that two cone plans sharing the same weighted base projection need not have the same ordinary cone marginal. The radial homogeneity removes this obstruction. Indeed, common radial rescaling

$$((x, r), (y, s)) \mapsto ((x, r/\vartheta), (y, s/\vartheta)), \qquad \gamma \mapsto \vartheta^p \gamma,$$

preserves both weighted projections and the p-action. After adding mass at the cone apex, the normalization lemma for homogeneous lifts therefore lets one rescale two almost optimal plans for (α, β) and (β, ζ) so that their ordinary middle cone marginal is the same lift

$$\eta_\beta = M\delta_{\mathfrak{o}} + (x \mapsto (x, 1))_\sharp \beta.$$

Here $M < +\infty$ is chosen large enough for both plans; additional mass at $(\mathfrak{o}, \mathfrak{o})$ changes neither the weighted projections nor the action. The ordinary gluing lemma now produces a joint cone plan (Z_0, Z_1, Z_2). Since Δ is a metric, Minkowski's inequality gives

$$\left(\int \Delta(Z_0, Z_2)^p\right)^{1/p} \leqslant \left(\int \Delta(Z_0, Z_1)^p\right)^{1/p} + \left(\int \Delta(Z_1, Z_2)^p\right)^{1/p}.$$

Taking infima proves the triangle inequality. This normalization-and-gluing argument is the metric engine of the homogeneous cone construction [340].

For completeness, normalize each non-apex pair by taking $\vartheta = (r^p + s^p)^{1/p}$. The transformed radii then satisfy $r^p + s^p = 1$, while the transformed plan has total mass $\alpha(\mathcal{X}) + \beta(\mathcal{X})$; apex mass can again be fixed without changing the problem. Thus minimizing plans may be restricted to a weakly compact family with bounded radii. Compactness of $\mathcal{X}$ and lower semicontinuity of the cone cost give an optimal plan. If $\mathrm{CW}(\alpha, \beta) = 0$, such a plan is concentrated on the diagonal of the cone. There $x = y$ whenever the common radius is positive and $r = s$, so its two weighted base projections coincide and $\alpha = \beta$. Nonnegativity and symmetry are immediate, completing the metric proof. □

Entropic KL relaxation. A generic entropic regularization of unbalanced OT reads

$$\inf_{\pi \in \mathcal{M}_+(\mathcal{X} \times \mathcal{Y})} \int c \mathrm{d}\pi + \mathcal{D}_{\psi_1}(\pi_1|\alpha) + \mathcal{D}_{\psi_2}(\pi_2|\beta) + \varepsilon \mathcal{D}_\varphi(\pi|\alpha \otimes \beta). \tag{11.12}$$

Its dual is

$$\sup_{f,g} -\mathcal{D}^*_{\psi_1}(-f|\alpha) - \mathcal{D}^*_{\psi_2}(-g|\beta) - \varepsilon \mathcal{D}^*_\varphi\left(\frac{f \oplus g - c}{\varepsilon}\middle|\alpha \otimes \beta\right).$$

For $\mathcal{D}_\varphi = \mathrm{KL}$, the primal-dual relation is $\mathrm{d}\pi = e^{(f \oplus g - c)/\varepsilon} \mathrm{d}\alpha \mathrm{d}\beta$. If, in addition, $\mathcal{D}_{\psi_1} = \mathcal{D}_{\psi_2} = \tau\,\mathrm{KL}$, the dual specializes to

$$\sup_{f,g} -\tau \int (e^{-f/\tau} - 1) \mathrm{d}\alpha - \tau \int (e^{-g/\tau} - 1) \mathrm{d}\beta - \varepsilon \iint \left(e^{(f \oplus g - c)/\varepsilon} - 1\right) \mathrm{d}\alpha \mathrm{d}\beta,$$

and coordinate maximization gives the damped soft-transform updates

$$f \leftarrow \omega\, g^{\bar{c},\varepsilon}, \qquad g \leftarrow \omega\, f^{c,\varepsilon}, \qquad \omega := \frac{\tau}{\tau + \varepsilon},$$

where the soft c-transforms are those of Definition 8.21. Equivalently,

$$f(x) = -\frac{\tau\varepsilon}{\tau+\varepsilon}\log\int_{\mathcal{Y}}\exp\left(\frac{g(y)-c(x,y)}{\varepsilon}\right)\mathrm{d}\beta(y),$$
$$g(y) = -\frac{\tau\varepsilon}{\tau+\varepsilon}\log\int_{\mathcal{X}}\exp\left(\frac{f(x)-c(x,y)}{\varepsilon}\right)\mathrm{d}\alpha(x).$$

The second assignment uses the freshly updated value of f, as in a Gauss–Seidel block-coordinate ascent. In the discrete case, with $K_{i,j} = e^{-\mathrm{C}_{i,j}/\varepsilon}\mathrm{a}_i\mathrm{b}_j$, this gives the generalized Sinkhorn scaling

$$u_i \leftarrow \left(\frac{\mathrm{a}_i}{(Kv)_i}\right)^{\omega}, \qquad v_j \leftarrow \left(\frac{\mathrm{b}_j}{(K^\top u)_j}\right)^{\omega}, \qquad \mathrm{P} = \mathrm{diag}(u)K\,\mathrm{diag}(v).$$

The exponent $\omega < 1$ is the visible difference with balanced Sinkhorn: marginal corrections are damped because violating the marginals is allowed.

Metric contraction of the damped updates. For balanced entropic OT, potentials are only defined up to the gauge $(f,g) \sim (f+\lambda, g-\lambda)$, so Hilbert's projective metric is natural. The KL marginal penalties break this invariance: under the usual finite-dual assumptions, the dual potentials are unique up to α- and β-null sets, and the natural distance is the ordinary L^∞ distance on potentials. In scaling variables $u = e^{f/\varepsilon}$ and $v = e^{g/\varepsilon}$, this is the Thompson metric on the product positive cone,

$$d_T\big((u,v),(u',v')\big) := \max\{\|\log u - \log u'\|_\infty, \|\log v - \log v'\|_\infty\},$$

up to the factor ε on potentials [335].

Proposition 11.11 (Linear contraction of unbalanced Sinkhorn)**.** *Assume that the damped soft-transform map sends bounded potentials to bounded potentials; for instance, this holds when c is bounded. Let $\omega = \tau/(\tau+\varepsilon) < 1$, and define*

$$d_\infty((f,g),(f',g')) := \max\{\|f-f'\|_\infty, \|g-g'\|_\infty\}.$$

The damped soft-transform map

$$\mathcal{T}_{\mathrm{GS}}(f,g) := \big(\tilde f, \tilde g\big), \qquad \tilde f = \omega\, g^{\bar c,\varepsilon}, \qquad \tilde g = \omega\, \tilde f^{c,\varepsilon},$$

is a contraction:

$$d_\infty\big(\mathcal{T}_{\mathrm{GS}}(f,g), \mathcal{T}_{\mathrm{GS}}(f',g')\big) \leqslant \omega\, d_\infty((f,g),(f',g')).$$

Consequently, on the bounded-potential space under consideration, the unbalanced Sinkhorn potentials have a unique fixed point $(f^\star, g^\star)$ and the iterates satisfy

$$d_\infty((f^{(\ell)},g^{(\ell)}),(f^\star,g^\star)) \leqslant \omega^\ell\, d_\infty((f^{(0)},g^{(0)}),(f^\star,g^\star)).$$

Equivalently, for $u^{(\ell)} = e^{f^{(\ell)}/\varepsilon}$ and $v^{(\ell)} = e^{g^{(\ell)}/\varepsilon}$,

$$d_T((u^{(\ell)},v^{(\ell)}),(u^\star,v^\star)) \leqslant \omega^\ell\, d_T((u^{(0)},v^{(0)}),(u^\star,v^\star)).$$

Proof. The soft c-transform is 1-Lipschitz in L^∞. Indeed, if $\|g-g'\|_\infty \leqslant \delta$, then $g-\delta \leqslant g' \leqslant g+\delta$. Since $g \mapsto g^{\bar c,\varepsilon}$ is order reversing and satisfies $(g+\lambda)^{\bar c,\varepsilon} = g^{\bar c,\varepsilon} - \lambda$, this implies

$$g^{\bar c,\varepsilon} - \delta \leqslant (g')^{\bar c,\varepsilon} \leqslant g^{\bar c,\varepsilon} + \delta.$$

Hence $\|g^{\bar c,\varepsilon} - (g')^{\bar c,\varepsilon}\|_\infty \leqslant \|g-g'\|_\infty$, and the same argument applies to $f \mapsto f^{c,\varepsilon}$.

Let $(\tilde f, \tilde g) = \mathcal{T}_{\mathrm{GS}}(f,g)$ and $(\tilde f', \tilde g') = \mathcal{T}_{\mathrm{GS}}(f',g')$. With $D = d_\infty((f,g),(f',g'))$, the previous Lipschitz estimate gives

$$\|\tilde f - \tilde f'\|_\infty \leqslant \omega\|g-g'\|_\infty \leqslant \omega D,$$

and then

$$\|\tilde g - \tilde g'\|_\infty \leqslant \omega\|\tilde f - \tilde f'\|_\infty \leqslant \omega^2 D \leqslant \omega D.$$

This proves the contraction. The Banach fixed-point theorem gives existence, uniqueness and the displayed linear rate. □

Algorithm 11.1: Unbalanced Sinkhorn scaling

Input: Positive weights a, b, cost matrix C, entropic scale $\varepsilon > 0$, KL strength $\tau > 0$, tolerance tol, iteration budget $L \geqslant 1$.
Output: Unbalanced entropic coupling P.
Initialize: Set $K_{i,j} = e^{-C_{i,j}/\varepsilon} a_i b_j$, $\omega = \frac{\tau}{\tau+\varepsilon}$, $u^{(0)} = \mathbb{1}_n$, $v^{(0)} = \mathbb{1}_m$, $\eta^{(0)} = +\infty$, $\ell = 0$.
While $\eta^{(\ell)} >$ tol and $\ell < L$ **do**:
  Set $\ell \leftarrow \ell + 1$.
  $u^{(\ell)} = \left(\frac{a}{Kv^{(\ell-1)}}\right)^{\omega}$, $v^{(\ell)} = \left(\frac{b}{K^\top u^{(\ell)}}\right)^{\omega}$.
  Set $\eta^{(\ell)} = \varepsilon \max\{\|\log u^{(\ell)} - \log u^{(\ell-1)}\|_\infty, \|\log v^{(\ell)} - \log v^{(\ell-1)}\|_\infty\}$.
Return $\mathrm{P}^{(\ell)} = \mathrm{diag}(u^{(\ell)}) K \,\mathrm{diag}(v^{(\ell)})$.

Gaussian KL-unbalanced transport. Gaussian inputs provide one of the few continuous unbalanced problems with an explicit endpoint cost. The terminology nevertheless requires care. With a quadratic ground cost, KL marginal penalties and no entropy on the coupling, the static problem below is commonly called the *squared Gaussian Hellinger–Kantorovich endpoint cost*. The Wasserstein–Fisher–Rao distance is instead the intrinsic length metric represented by the cosine cone cost in Section 14.5; see Proposition 14.26 and [340]. The graph-supported zero-coupling-entropy formulas below follow [557]; the nondegenerate positive-coupling-entropy counterpart recalled afterward is from [291].

Let

$$\alpha = r_0 \mathcal{N}(m_0, \Sigma_0), \qquad \beta = r_1 \mathcal{N}(m_1, \Sigma_1), \qquad r_0, r_1 > 0, \quad \Sigma_0, \Sigma_1 > 0,$$

where r_i is the total mass, $m_i \in \mathbb{R}^d$ is the mean and Σ_i is the covariance. For a common marginal-relaxation strength $\tau > 0$, define

$$\mathrm{GHK}^2_\tau(\alpha, \beta) := \inf_{\pi \in \mathcal{M}_+(\mathbb{R}^d \times \mathbb{R}^d)} \left\{ \int \|x - y\|^2 \mathrm{d}\pi + \tau\, \mathrm{KL}(\pi_1|\alpha) + \tau\, \mathrm{KL}(\pi_2|\beta) \right\}. \tag{11.13}$$

Write $\bar\alpha = \mathcal{N}(m_0, \Sigma_0)$ and $\bar\beta = \mathcal{N}(m_1, \Sigma_1)$, and denote by

$$\mathcal{E}_\tau := \mathcal{E}_{c_2,\tau,\tau}(\bar\alpha, \bar\beta), \qquad c_2(x, y) = \|x - y\|^2, \tag{11.14}$$

the normalized KL shape envelope from (11.4).

The first step is independent of the closed-form matrix calculation: among all normalized marginals, Gaussian ones are optimal. For $P, \Sigma \in \mathbb{S}^d_{++}$, introduce the LogDet divergence

$$D_{\mathrm{ld}}(P|\Sigma) := \mathrm{tr}(\Sigma^{-1}P) - \log\det(\Sigma^{-1}P) - d, \tag{11.15}$$

and set

$$G_\tau = \Sigma_0 + \Sigma_1 + \frac{\tau}{2}\mathrm{Id}, \tag{11.16}$$

$$C_\tau(\Sigma_0, \Sigma_1) := \min_{P,Q \in \mathbb{S}^d_{++}} \left\{ \mathcal{B}^2(P, Q) + \frac{\tau}{2} D_{\mathrm{ld}}(P|\Sigma_0) + \frac{\tau}{2} D_{\mathrm{ld}}(Q|\Sigma_1) \right\}. \tag{11.17}$$

Proposition 11.12 (Gaussian reduction of the KL shape envelope). *Let $\Delta m = m_0 - m_1$. Then*

$$\mathcal{E}_\tau = \frac{\tau}{2}\|\Delta m\|^2_{G_\tau^{-1}} + C_\tau(\Sigma_0, \Sigma_1). \tag{11.18}$$

The unique minimizing normalized marginals are Gaussian. Their means are

$$u_\tau = m_0 - \Sigma_0 G_\tau^{-1} \Delta m, \qquad v_\tau = m_1 + \Sigma_1 G_\tau^{-1} \Delta m, \tag{11.19}$$

and their covariances minimize (11.17).

Proof. Consider finite-objective competitors $\widetilde\alpha, \widetilde\beta$ in (11.14), with respective means and covariances (u, P) and (v, Q). Finite relative entropy with respect to a nondegenerate Gaussian gives finite second moments and $P, Q \succ 0$. Gelbrich's inequality, Theorem 16.30, lowers the transport cost after replacing the competitors by $\mathcal{N}(u, P)$ and $\mathcal{N}(v, Q)$. Moment matching also gives the information-projection identity

$$\mathrm{KL}(\widetilde\alpha|\bar\alpha) = \mathrm{KL}\big(\widetilde\alpha|\mathcal{N}(u, P)\big) + \mathrm{KL}\big(\mathcal{N}(u, P)|\bar\alpha\big),$$

and similarly for the second marginal. Hence Gaussian replacement cannot increase the objective.

The Gaussian formulas for $\mathcal{W}_2^2$ and KL separate the mean and covariance variables. The first-order equations for the mean part are

$$u - v + \frac{\tau}{2}\Sigma_0^{-1}(u - m_0) = 0, \qquad v - u + \frac{\tau}{2}\Sigma_1^{-1}(v - m_1) = 0.$$

Solving them gives (11.19); substitution gives the first term in (11.18), while the covariance part is (11.17). The resulting finite-dimensional objective is coercive in the means and covariances and diverges at the boundary of $\mathbb{S}^d_{++}$, so it attains its minimum. Finally, Proposition 3.32 gives joint convexity of the transport term, and each KL term is strictly convex in its first argument. Thus the minimizing pair is unique. □

The covariance envelope is a strictly convex optimization problem, despite its nonlinear matrix square roots. Its first-order equations reduce to one positive-definite Riccati equation.

Proposition 11.13 (Adjusted covariance pair)**.** *Set $\eta = 2/\tau$, $R_i = \Sigma_i^{-1} + \eta\mathrm{Id}$, and*

$$L_\tau = R_1^{-1/2}\left(R_1^{1/2} R_0 R_1^{1/2}\right)^{1/2} R_1^{-1/2}. \tag{11.20}$$

Then $L_\tau R_1 L_\tau = R_0$, and the unique minimizer of (11.17) is

$$P_\tau = \left[\Sigma_0^{-1} + \eta(\mathrm{Id} - L_\tau)\right]^{-1}, \qquad Q_\tau = L_\tau P_\tau L_\tau. \tag{11.21}$$

Its value is

$$C_\tau(\Sigma_0, \Sigma_1) = \frac{\tau}{2}\log\frac{\det\Sigma_0 \det\Sigma_1}{\det P_\tau \det Q_\tau}. \tag{11.22}$$

Proof. For $P, Q \succ 0$, let L be the Gaussian optimal map from P to Q, so $Q = LPL$. The Frobenius gradients of the squared Bures term are $\mathrm{Id} - L$ and $\mathrm{Id} - L^{-1}$ [54]; hence stationarity of (11.17) reads

$$P^{-1} = \Sigma_0^{-1} + \eta(\mathrm{Id} - L), \qquad Q^{-1} = \Sigma_1^{-1} + \eta(\mathrm{Id} - L^{-1}). \tag{11.23}$$

Substituting $Q^{-1} = L^{-1}P^{-1}L^{-1}$ gives $LR_1L = R_0$, whose unique positive-definite solution is (11.20). Back-substitution gives (11.21). Positivity follows directly from

$$2\eta^{-1}P_\tau^{-1} = (\mathrm{Id} - L_\tau)^2 + \eta^{-1}\Sigma_0^{-1} + \eta^{-1}L_\tau\Sigma_1^{-1}L_\tau \succ 0.$$

The Bures term is jointly convex by Proposition 2.59, while each LogDet term is strictly convex and coercive on $\mathbb{S}^d_{++}$; the stationary pair is therefore the unique global minimizer. Multiplying the two equations in (11.23) by P and Q, respectively, and taking traces gives

$$\mathcal{B}^2(P, Q) + \frac{\tau}{2}\left[\mathrm{tr}(\Sigma_0^{-1}P) + \mathrm{tr}(\Sigma_1^{-1}Q) - 2d\right] = 0.$$

At $(P, Q) = (P_\tau, Q_\tau)$, this identity cancels the Bures term against the trace parts of the two LogDet divergences. The remaining logarithms give (11.22). □

The preceding convex covariance problem has a factor interpretation that directly extends the ordinary Bures–Procrustes formula. For a full-row-rank matrix $Z \in \mathbb{R}^{d\times d'}$, define

$$\Psi(Z) := \|Z\|_{\mathrm{F}}^2 - \log\det(ZZ^\top) - d. \tag{11.24}$$

Proposition 11.14 (Soft Procrustes formula)**.** *Let $S_i = \Sigma_i^{1/2}$ and $d' \geqslant d$. Then*

$$C_\tau(\Sigma_0, \Sigma_1) = \min_{\substack{Z_0, Z_1 \in \mathbb{R}^{d\times d'} \\ \mathrm{rank}(Z_0)=\mathrm{rank}(Z_1)=d}} \left\{\|S_0Z_0 - S_1Z_1\|_{\mathrm{F}}^2 + \frac{\tau}{2}\Psi(Z_0) + \frac{\tau}{2}\Psi(Z_1)\right\}. \tag{11.25}$$

Moreover, $\Psi(Z) \geqslant 0$, with equality if and only if $ZZ^\top = \mathrm{Id}$. Thus the LogDet terms soften the row-Stiefel constraints in the ordinary Bures–Procrustes problem.

Proof. Insert the free-factor Bures formula of Proposition 2.58 into (11.17). Every full-row-rank factor of $P_i > 0$ can be written $F_i = S_i Z_i$, and cyclicity of the trace together with similarity invariance of the determinant gives

$$D_{\mathrm{ld}}(F_i F_i^\top | \Sigma_i) = \operatorname{tr}(Z_i Z_i^\top) - \log\det(Z_i Z_i^\top) - d = \Psi(Z_i).$$

This proves (11.25). If $s_1, \dots, s_d$ are the singular values of Z, then $\Psi(Z) = \sum_j (s_j^2 - 2\log s_j - 1)$; each summand is nonnegative and vanishes exactly at $s_j = 1$. □

The covariance problem (11.17) is convex, but its factorized form (11.25) is generally not convex in (Z_0, Z_1). It is an exact redundant parameterization, invariant under the simultaneous right action $(Z_0, Z_1) \mapsto (Z_0 R, Z_1 R)$ for $R \in \mathrm{O}(d')$.

Remark 11.15 (Hard Bures limit). Formula (11.25) is the penalized counterpart of Proposition 2.58. Indeed, row-Stiefel competitors give $C_\tau \leqslant \mathcal{B}^2(\Sigma_0, \Sigma_1)$. Therefore any minimizing factors satisfy $\Psi(Z_i) \to 0$ as $\tau \to +\infty$. Coercivity of Ψ and lower semicontinuity then reduce the limiting problem to the hard row-Stiefel constraints, yielding

$$C_\tau(\Sigma_0, \Sigma_1) \longrightarrow \mathcal{B}^2(\Sigma_0, \Sigma_1).$$

Together with Proposition 11.12, this recovers the Gaussian $\mathcal{W}_2^2$ formula (2.19). In particular, when $r_0 = r_1 = r$, the endpoint value (11.13) converges to $r\,\mathcal{W}_2^2(\bar\alpha, \bar\beta)$.

We can now state the endpoint formula in terms of the adjusted Gaussian marginals. Define

$$\mathfrak{a}_\tau := \exp\left(-\frac{1}{4}\|m_0 - m_1\|^2_{G_\tau^{-1}}\right)\left(\frac{\det P_\tau \det Q_\tau}{\det\Sigma_0 \det\Sigma_1}\right)^{1/4}. \tag{11.26}$$

Proposition 11.16 (Gaussian Hellinger–Kantorovich endpoint)**.** *With the preceding notation,*

$$\mathrm{GHK}^2_\tau(\alpha, \beta) = \tau\left(r_0 + r_1 - 2\sqrt{r_0 r_1}\,\mathfrak{a}_\tau\right). \tag{11.27}$$

The optimal transported mass is $M_\tau = \sqrt{r_0 r_1}\,\mathfrak{a}_\tau$. More precisely, the unique optimal plan is

$$\begin{aligned} \pi_\tau &= M_\tau(\mathrm{Id}, T_\tau)_\sharp \mathcal{N}(u_\tau, P_\tau), \\ T_\tau(x) &= v_\tau + L_\tau(x - u_\tau). \end{aligned} \tag{11.28}$$

Proof. Propositions 11.12 and 11.13 give

$$\exp\left(-\frac{\mathcal{E}_\tau}{2\tau}\right) = \mathfrak{a}_\tau.$$

The equal-penalty specialization of the KL mass–shape formula (11.5) therefore gives (11.27) and the mass M_τ. The optimal normalized coupling is the Gaussian $\mathcal{W}_2$ graph coupling between the adjusted marginals from Propositions 11.12 and 11.13, which is precisely (11.28). The scalar mass minimization and the normalized marginals are unique; since the adjusted source is nondegenerate, Brenier uniqueness then gives uniqueness of the full plan. □

Remark 11.17 (Failure of Gaussian closure for WFR). Gaussian optimal-plan closure is established at zero coupling entropy in [557] and at positive coupling entropy in [291]. In the first case, the optimizer is a scaled Gaussian Wasserstein coupling between two adjusted Gaussian marginals; in the second, it is a non-degenerate scaled Gaussian on the product space. Consequently, linearly interpolating a pair drawn from either plan produces Gaussian one-time marginals. This interpolation should not be identified with an ambient WFR geodesic, which is governed by the cosine cone geometry of Proposition 14.26 and need not remain Gaussian.

The failure of geodesic closure is already explicit at the pure Hellinger endpoint of WFR. Let ρ_i be the density of $\mathcal{N}(m_i, \Sigma_i)$ for $i \in \{0, 1\}$, and set

$$\mathfrak{h}_{0,1} = \int \sqrt{\rho_0 \rho_1}, \qquad \Sigma_{0,1} = 2(\Sigma_0^{-1} + \Sigma_1^{-1})^{-1}, \qquad m_{0,1} = (\Sigma_0^{-1} + \Sigma_1^{-1})^{-1}(\Sigma_0^{-1} m_0 + \Sigma_1^{-1} m_1).$$

Completing the square gives $\sqrt{\rho_0\rho_1} = \mathfrak{h}_{0,1}\rho_{0,1}$, where $\rho_{0,1}$ is the density of $\mathcal{N}(m_{0,1}, \Sigma_{0,1})$. The Hellinger

geodesic is therefore

$$\rho_t = \big((1-t)\sqrt{\rho_0} + t\sqrt{\rho_1}\big)^2 = (1-t)^2\rho_0 + 2t(1-t)\mathfrak{b}_{0,1}\rho_{0,1} + t^2\rho_1. \tag{11.29}$$

Except in degenerate cases, every interior point is thus an unnormalized mixture of three distinct Gaussian densities, not a single Gaussian.
The same obstruction applies to barycenters. If ρ_s are Gaussian densities and $\lambda_s > 0$ sum to one, their squared-Hellinger barycenter satisfies

$$\rho^\star = \left(\sum_{s=1}^{S} \lambda_s \sqrt{\rho_s}\right)^2 = \sum_{s,t=1}^{S} \lambda_s\lambda_t \sqrt{\rho_s\rho_t}. \tag{11.30}$$

Every cross-term is proportional to a Gaussian density, so this is generically a mixture with $S(S+1)/2$ distinct Gaussian components. Consequently, the Gaussian manifold is neither geodesically nor barycentrically closed for the Hellinger endpoint, and no Gaussian-closure property holds for WFR in general.

The strictly entropic counterpart of (11.13) also has a closed form. Specializing (11.12) to $c(x,y) = \|x-y\|^2$, $\mathcal{D}_{\psi_1} = \mathcal{D}_{\psi_2} = \tau\,\mathrm{KL}$, and $\mathcal{D}_\varphi = \mathrm{KL}$ gives

$$\mathrm{UW}^{\mathrm{KL}}_{\tau,\varepsilon}(\alpha,\beta) \coloneqq \inf_{\pi\in\mathcal{M}_+(\mathbb{R}^d\times\mathbb{R}^d)} \left\{ \int \|x-y\|^2 \mathrm{d}\pi + \tau\,\mathrm{KL}(\pi_1|\alpha) + \tau\,\mathrm{KL}(\pi_2|\beta) + \varepsilon\,\mathrm{KL}(\pi|\alpha\otimes\beta) \right\}. \tag{11.31}$$

Write $\varepsilon = 2\sigma^2 > 0$ and introduce

$$\omega = \frac{\tau}{\tau+2\sigma^2}, \qquad \ell = \sigma^2 + \frac{\tau}{2}, \qquad X = \Sigma_0 + \Sigma_1 + \ell\mathrm{Id},$$
$$\widetilde{\Sigma}_i = \frac{\tau}{2}\left(\mathrm{Id} - \ell(\Sigma_i + \ell\mathrm{Id})^{-1}\right), \qquad C = \left(\frac{1}{\omega}\widetilde{\Sigma}_0\widetilde{\Sigma}_1 + \frac{\sigma^4}{4}\mathrm{Id}\right)^{1/2} - \frac{\sigma^2}{2}\mathrm{Id}.$$

Define the transported mass

$$M_{\varepsilon,\tau} = \sigma^{\frac{d\sigma^2}{\tau+\sigma^2}} \left[r_0 r_1 \det(C) \left(\frac{\det(\widetilde{\Sigma}_0\widetilde{\Sigma}_1)^\omega}{\det(\Sigma_0\Sigma_1)} \right)^{1/2} \right]^{\frac{1}{1+\omega}} \times \frac{\exp\left(-\frac{\|m_0-m_1\|^2_{X^{-1}}}{2(1+\omega)}\right)}{\sqrt{\det\left(C - \frac{2}{\tau}\widetilde{\Sigma}_0\widetilde{\Sigma}_1\right)}}. \tag{11.32}$$

Proposition 11.18 (Entropic KL–UOT between unnormalized Gaussians)**.** *For the preceding parameters, the unique optimizer of* (11.31) *is a scaled Gaussian plan of total mass $M_{\varepsilon,\tau}$, and the optimal value is*

$$\mathrm{UW}^{\mathrm{KL}}_{\tau,\varepsilon}(\alpha,\beta) = \tau(r_0+r_1) + \varepsilon r_0 r_1 - (\varepsilon+2\tau)M_{\varepsilon,\tau}. \tag{11.33}$$

Proof. The KL primal–dual relation below (11.12) turns quadratic dual potentials into the exponential of a quadratic form, hence into a scaled Gaussian plan. The damped soft transforms preserve this quadratic class. Solving their matrix fixed point gives the matrices X, $\widetilde{\Sigma}_i$ and C above; evaluating the Gaussian normalizing constant gives (11.32). Finally, integrating the two marginal optimality relations and the plan optimality relation cancels all moment terms and leaves

$$\tau(r_0+r_1) + 2\sigma^2 r_0 r_1 - 2(\sigma^2+\tau)M_{\varepsilon,\tau},$$

which is (11.33) because $\varepsilon = 2\sigma^2$ [291]. □

Partial optimal transport. Total variation gives a sharp active-mass selection and connects unbalanced OT with the classical partial-transport problem. The latter fixes in advance the amount of transported mass.

Definition 11.19 (Partial optimal transport)**.** For $0 \leqslant m \leqslant \min\{\alpha(\mathcal{X}), \beta(\mathcal{Y})\}$, define

$$\mathrm{POT}_m(\alpha,\beta) \coloneqq \inf_{\substack{\pi\in\mathcal{M}_+(\mathcal{X}\times\mathcal{Y})\\ \pi_1\leqslant\alpha,\ \pi_2\leqslant\beta\\ \pi(\mathcal{X}\times\mathcal{Y})=m}} \int c\,\mathrm{d}\pi.$$

Writing $A = \alpha(\mathcal{X})$ and $B = \beta(\mathcal{Y})$, for $\lambda \geqslant 0$ its TV-penalized form is

$$\mathrm{POT}_\lambda^{\mathrm{TV}}(\alpha,\beta) \coloneqq \inf_{\substack{\pi\in\mathcal{M}_+(\mathcal{X}\times\mathcal{Y})\\ \pi_1\leqslant\alpha,\ \pi_2\leqslant\beta}} \left\{ \int c\,\mathrm{d}\pi + \lambda\big(A + B - 2\pi(\mathcal{X}\times\mathcal{Y})\big) \right\}.$$

Thus only a submeasure of α is transported onto a submeasure of β; the remaining mass is left unmatched. The corresponding Lagrangian form is obtained by adding total-variation penalties. For a price $\lambda > 0$ for discarding or creating one unit of mass, consider

$$\inf_{\pi\in\mathcal{M}_+(\mathcal{X}\times\mathcal{Y})} \int c\,\mathrm{d}\pi + \lambda\|\alpha - \pi_1\|_{\mathrm{TV}} + \lambda\|\beta - \pi_2\|_{\mathrm{TV}}.$$

Assume $c \geqslant 0$. Allowing transported marginals larger than the available marginals does not improve the value: excess transported mass can be trimmed, which decreases the transport cost and cannot increase the sum of the two total-variation penalties. Hence an optimal plan may be chosen with $\pi_1 \leqslant \alpha$ and $\pi_2 \leqslant \beta$. If $m = \pi(\mathcal{X}\times\mathcal{Y})$, the penalty then reduces to

$$\lambda\big(\alpha(\mathcal{X}) - m\big) + \lambda\big(\beta(\mathcal{Y}) - m\big),$$

so the precise relation is a one-dimensional Lagrange duality in the transported mass.

The trimmed expression in Definition 11.19 makes the relation between penalized and prescribed transported mass explicit.

Proposition 11.20 (TV penalization selects fixed-mass partial OT)**.** *Assume that $\mathcal{X}, \mathcal{Y}$ are compact metric spaces, that c is continuous and nonnegative, and set $A = \alpha(\mathcal{X})$, $B = \beta(\mathcal{Y})$, $M = \min(A, B)$. For $\lambda \geqslant 0$, Then*

$$\mathrm{POT}_\lambda^{\mathrm{TV}}(\alpha,\beta) = \lambda(A+B) + \inf_{0\leqslant m\leqslant M} \{\mathrm{POT}_m(\alpha,\beta) - 2\lambda m\}.$$

If π_λ minimizes $\mathrm{POT}_\lambda^{\mathrm{TV}}$ and $m_\lambda = \pi_\lambda(\mathcal{X}\times\mathcal{Y})$, then m_λ minimizes $m \mapsto \mathrm{POT}_m(\alpha,\beta) - 2\lambda m$ and π_λ minimizes $\mathrm{POT}_{m_\lambda}(\alpha,\beta)$. Conversely, if 2λ is a subgradient at m of the convex function $m \mapsto \mathrm{POT}_m(\alpha,\beta)$ extended by $+\infty$ outside $[0, M]$, then every minimizer of $\mathrm{POT}_m(\alpha,\beta)$ is a minimizer of $\mathrm{POT}_\lambda^{\mathrm{TV}}(\alpha,\beta)$. In particular, every prescribed mass $m \in [0, M]$ is selected by at least one multiplier $\lambda \geqslant 0$, possibly together with a whole interval of masses when the value function has a flat exposed face.

Proof. Write $V(m) = \mathrm{POT}_m(\alpha,\beta)$. Compactness and continuity give an optimizer for every $m \in [0, M]$: the admissible subcouplings form a weakly compact set and the cost is continuous. The function V is convex on $[0, M]$: if π_0, π_1 are admissible competitors of masses m_0, m_1, then $(1-t)\pi_0 + t\pi_1$ is admissible for mass $(1-t)m_0 + tm_1$ and has the corresponding convex-combination cost. It is also nondecreasing because any competitor of mass $m_1 > m_0$ can be rescaled by m_0/m_1 to give a competitor of mass m_0 with no larger cost. Finally, V is Lipschitz. If $0 \leqslant m_0 < m_1 \leqslant M$ and π_0 is optimal at mass m_0, both residual measures $\alpha - (\pi_0)_1$ and $\beta - (\pi_0)_2$ have mass at least $m_1 - m_0$. Choose submeasures of each residual with this common mass and couple them. Adding the resulting plan to π_0 gives an admissible plan of mass m_1 at an additional cost bounded by $\|c\|_\infty(m_1 - m_0)$.

For any admissible π in the definition of $\mathrm{POT}_\lambda^{\mathrm{TV}}$, let $m = \pi(\mathcal{X}\times\mathcal{Y})$. The plan is admissible for $V(m)$, hence its objective is at least $V(m) + \lambda(A + B - 2m)$. Taking the infimum over π gives the lower bound by the displayed right-hand side. Conversely, for each m and each minimizing sequence for $V(m)$, the same plans are admissible in $\mathrm{POT}_\lambda^{\mathrm{TV}}$ and have objective tending to $V(m) + \lambda(A + B - 2m)$. Taking the infimum over m gives the reverse inequality.

The optimizer statement follows from the same inequalities. If π_λ minimizes the penalized problem with mass m_λ, then equality in the lower bound forces both $V(m_\lambda) - 2\lambda m_\lambda$ to be minimal over $[0, M]$ and π_λ to be optimal at fixed mass m_λ. Conversely, $2\lambda \in \partial V(m)$ means

$$V(m') - 2\lambda m' \geqslant V(m) - 2\lambda m \quad \text{for all } m' \in [0, M].$$

Thus any minimizer of $V(m)$ attains the penalized infimum. Because the extended function is proper, convex, lower semicontinuous and Lipschitz on its effective interval, its subdifferential is nonempty at every $m \in [0, M]$ after adding the normal cone of the interval. Monotonicity of V ensures that a subgradient can be chosen nonnegative, hence equal to 2λ for some $\lambda \geqslant 0$. □

Proposition 11.21 (Fixed total mass reduces partial OT to balanced OT)**.** *Let $(\mathcal{X}, d)$ be a metric space, let $p \geqslant 1$, set $c = d^p$, and fix $m > 0$. On the class of nonnegative measures with total mass exactly m and finite pth moment,*

$$\mathcal{M}_m^p(\mathcal{X}) := \left\{ \alpha \in \mathcal{M}_+(\mathcal{X}) \,;\, \alpha(\mathcal{X}) = m,\ \int d(x_0, x)^p \mathrm{d}\alpha(x) < +\infty \right\}$$

for one, hence any, base point x_0, the quantity

$$d_m(\alpha, \beta) := \mathrm{POT}_m(\alpha, \beta)^{1/p}$$

is a distance. More precisely, if $\bar{\alpha} = \alpha/m$ and $\bar{\beta} = \beta/m$, then

$$d_m(\alpha, \beta) = m^{1/p}\, \mathcal{W}_p(\bar{\alpha}, \bar{\beta}).$$

If instead POT_m is considered on a larger class of measures with mass at least m, then it is not a distance in general for $m > 0$: different measures can share a common submeasure of mass m, giving zero partial cost.

Proof. If $\alpha(\mathcal{X}) = \beta(\mathcal{X}) = m$ and π is admissible in $\mathrm{POT}_m(\alpha, \beta)$, then $\pi_1 \leqslant \alpha$ and both measures have mass m, hence $\pi_1 = \alpha$. Similarly $\pi_2 = \beta$. Thus the admissible plans are exactly the ordinary couplings between α and β. Dividing such a coupling by m gives a probability coupling between $\bar{\alpha}$ and $\bar{\beta}$, and conversely multiplying any coupling of $\bar{\alpha}, \bar{\beta}$ by m gives a coupling of α, β. Therefore

$$\mathrm{POT}_m(\alpha, \beta) = m\, \mathcal{W}_p(\bar{\alpha}, \bar{\beta})^p.$$

The metric properties of d_m follow from those of $\mathcal{W}_p$ in Proposition 4.7.

For the last claim, take three distinct points x, y, z in a metric space and a number $r > 0$. The measures $\alpha = m\delta_x + r\delta_y$ and $\beta = m\delta_x + r\delta_z$ are different but the partial plan $m\delta_{(x,x)}$ is admissible and has zero cost. Hence $\mathrm{POT}_m(\alpha, \beta) = 0$ although $\alpha \neq \beta$. □

Thus the TV-penalized formulation is the Lagrangian envelope of constrained partial OT: increasing λ penalizes unmatched mass more strongly and therefore selects larger transported masses, while small λ makes deletion and creation cheaper. The constrained theory, including active regions and free boundaries, was developed by Caffarelli–McCann and Figalli [96, 214]. Modern computational and learning applications include partial Wasserstein and partial Gromov–Wasserstein variants, for instance in Chapel–Alaya–Gasso [123]. Figure 11.3 shows this active-region mechanism for two one-dimensional two-Gaussian mixtures, with the source mixture shifted to the left and the target mixture shifted to the right. As the prescribed transported mass decreases, the plan keeps only the least costly facing portions of the two marginals.

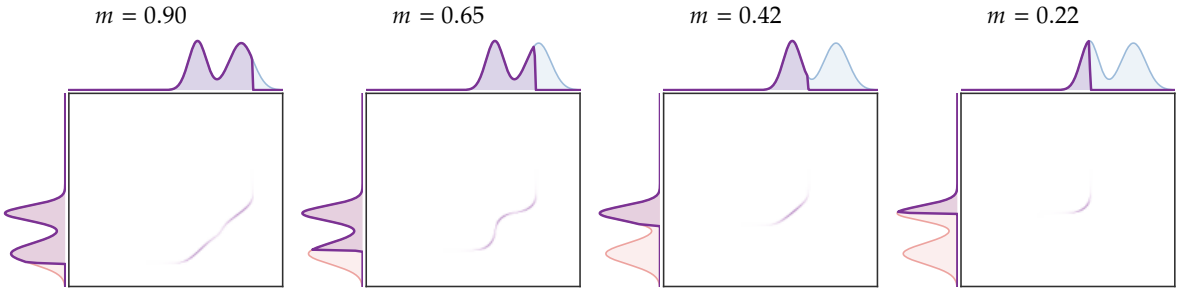


Figure 11.3: Partial optimal transport with prescribed transported mass between two shifted two-Gaussian mixtures. The central image is the optimal subcoupling, with contrast normalized independently in each panel to keep the low-mass plans readable. The pale red and blue side curves are the original source and target densities, while the violet curves are the active truncated marginals π_1 and π_2. As m decreases from left to right, only the closest facing portions remain matched and the remaining mass is left unmatched.

The same mechanism becomes a geometric active-region selection in higher dimension. Figure 11.4 shows a two-dimensional shape example: the partial plan selects the closest pieces of the two supports, while leaving distant regions unmatched.

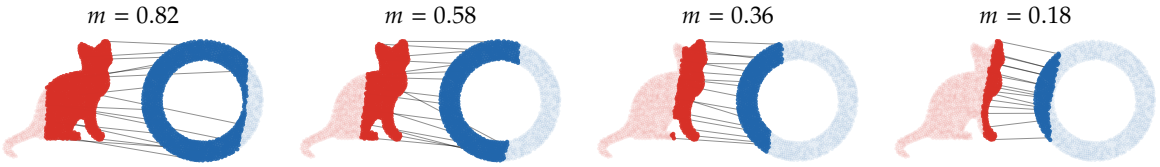


Figure 11.4: Two-dimensional partial optimal transport between a red cat-shaped indicator measure and a blue annulus indicator measure. Both supports are sampled by farthest-point sampling. In each panel, saturated points are the active source and target marginals of the optimal partial plan, while pale points are available mass left unmatched. Thin black segments show a farthest-spaced subsample of transported source locations connected to the dominant target selected by the plan. As the prescribed mass decreases, the active regions contract to the nearest compatible pieces of the two shapes.

11.2 Sliced Wasserstein Distances

Sliced Wasserstein distances replace one high-dimensional comparison by an average of explicit one-dimensional optimal transport problems.

One-dimensional projections. Sliced Wasserstein trades exact high-dimensional geometry for many one-dimensional projections. The idea was originally proposed by Marc Bernot, and the first published work using this construction for Wasserstein barycenters and texture mixing is due to Rabin, Peyré, Delon and Bernot [444]. It is cheap, differentiable after sorting, and often effective in imaging and learning. For measures on $\mathbb{R}^d$ and $\theta \in \mathbb{S}^{d-1}$, let $P_\theta(x) = \langle \theta, x \rangle$ be the projection on direction θ.

Spherical averaging. The projected measures $(P_\theta)_\sharp \alpha$ and $(P_\theta)_\sharp \beta$ live on the real line, where Wasserstein distances are computed by sorting or quantiles. Averaging these one-dimensional discrepancies over all directions gives the sliced distance.

Definition 11.22 (Sliced Wasserstein distance)**.** Let $p \geqslant 1$, let $\alpha, \beta \in \mathcal{P}_p(\mathbb{R}^d)$, and let σ be the uniform probability measure on the sphere $\mathbb{S}^{d-1}$. The sliced p-Wasserstein distance is

$$\mathrm{SW}_p(\alpha,\beta)^p := \int_{\mathbb{S}^{d-1}} \mathcal{W}_p\left((P_\theta)_\sharp \alpha, (P_\theta)_\sharp \beta\right)^p \mathrm{d}\sigma(\theta). \tag{11.34}$$

Since each projected problem can be solved by sorting or quantiles, SW_p is much cheaper to approximate numerically than high-dimensional OT. It metrizes the same weak-plus-moment topology as $\mathcal{W}_p$, but its geometry is not bi-Lipschitz equivalent to $\mathcal{W}_p$ in high dimension [396].

The family of all projected laws is itself a transform of measures.

Definition 11.23 (Measure-valued Radon transform)**.** For $\alpha \in \mathcal{P}(\mathbb{R}^d)$, define

$$\mathfrak{R}\alpha(\theta) := (P_\theta)_\sharp \alpha, \qquad \theta \in \mathbb{S}^{d-1}.$$

Remark 11.24 (Radon viewpoint). Slicing can also be understood as applying Definition 11.23 before measuring discrepancy. Thus $\mathfrak{R}\alpha$ is a collection of one-dimensional projected measures indexed by directions. If $\alpha = \rho(x)\mathrm{d}x$ has a density, then $\mathfrak{R}\alpha(\theta)$ has density given by the classical Radon transform

$$R\rho(\theta, s) = \int_{\{x:\langle\theta,x\rangle=s\}} \rho(x)\, \mathrm{d}\mathcal{H}^{d-1}(x).$$

To make the metric meaning of "pull-back" explicit, equip Radon fields $h : \mathbb{S}^{d-1} \to \mathcal{P}_p(\mathbb{R})$ with

$$d_{\mathrm{Rad},p}(h_0, h_1) := \left(\int_{\mathbb{S}^{d-1}} \mathcal{W}_p\left(h_0(\theta), h_1(\theta)\right)^p \mathrm{d}\sigma(\theta)\right)^{1/p}.$$

For any map $F : E \to (Y, d_Y)$, the pull-back of the metric d_Y through F is the function on $E \times E$ defined

by

$$(F^* d_Y)(x, x') := d_Y\big(F(x), F(x')\big).$$

Consequently, (11.34) reads

$$\mathrm{SW}_p(\alpha, \beta) = d_{\mathrm{Rad},p}(\mathfrak{R}\alpha, \mathfrak{R}\beta) = (\mathfrak{R}^* d_{\mathrm{Rad},p})(\alpha, \beta).$$

This is not the adjoint relation between the pull-back $T^\sharp g = g \circ T$ of a test function and the push-forward $T_\sharp$ of a measure discussed in Remark 2.11, although both uses amount to precomposing an object by a map. Here one precomposes both arguments of a metric; each coordinate $\mathfrak{R}\alpha(\theta) = (P_\theta)_\sharp \alpha$ is itself a push-forward. A metric pull-back is generally only a pseudometric when F is not injective. In the present setting, the full all-direction transform is injective on probability measures by the Cramér–Wold theorem [162], which is why sliced Wasserstein separates measures. The warning is numerical and variational: finitely sampled Radon data are not injective, and independently constructed one-dimensional Radon data need not satisfy the range conditions of an actual image. This is precisely the issue behind sliced and Radon barycenter reconstructions discussed in Section 12.1 and illustrated in Figure 12.3.

Figure 11.5 turns this Radon viewpoint into a concrete comparison: each planar density produces a family of one-dimensional projected laws, which can then be compared by ordinary one-dimensional Wasserstein distances.

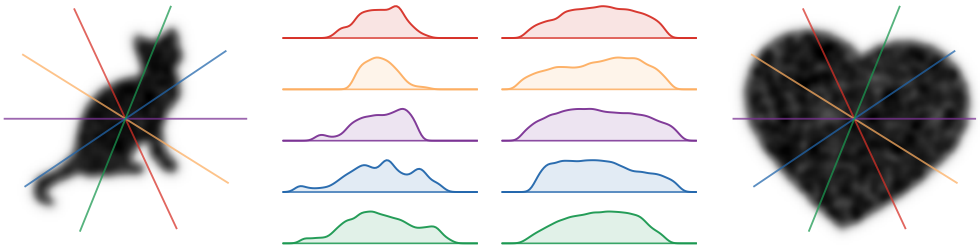


Figure 11.5: Sliced Wasserstein projections between two planar densities. The source and target are rendered as smoothed black-and-white density images obtained from dense farthest-point samples of two silhouettes. Five fixed directions are drawn on both densities. For each direction, the middle panels show smoothed one-dimensional density estimates of the projected measures $(P_\theta)_\sharp \alpha$ and $(P_\theta)_\sharp \beta$. Sliced OT averages one-dimensional Wasserstein discrepancies over many such directions, replacing a difficult planar comparison by a collection of sorted one-dimensional comparisons.

Proposition 11.25 (Metric properties of sliced Wasserstein)**.** *For $p \geqslant 1$, SW_p is a distance on $\mathcal{P}_p(\mathbb{R}^d)$. Moreover, SW_p metrizes weak convergence together with convergence of the pth moment. Finally,*

$$\mathrm{SW}_p(\alpha, \beta)^p \leqslant \kappa_{d,p}\, \mathcal{W}_p(\alpha, \beta)^p, \qquad \kappa_{d,p} := \int_{\mathbb{S}^{d-1}} |\theta_1|^p \mathrm{d}\sigma(\theta) = \frac{\Gamma(d/2)\Gamma((p+1)/2)}{\sqrt{\pi}\,\Gamma((d+p)/2)}.$$

In particular, $\kappa_{d,2} = 1/d$. Conversely, if $d \geqslant 2$, $R > 0$, and $\alpha, \beta \in \mathcal{P}_p(\mathbb{R}^d)$ are supported in $B_R := \{x \in \mathbb{R}^d \,;\, \|x\| \leqslant R\}$, then there is a constant $C_{d,p}$ depending only on d and p such that Bonnotte's historical estimate reads

$$\mathcal{W}_p(\alpha, \beta)^p \leqslant C_{d,p} R^{p - \frac{1}{d+1}}\, \mathrm{SW}_p(\alpha, \beta)^{\frac{1}{d+1}}.$$

The sharp reverse estimate for $p = 1$ also gives, for every $p \geqslant 1$, constants $\widetilde{C}_{d,p}, \widetilde{c}_{d,p} > 0$ such that

$$\mathcal{W}_p(\alpha, \beta) \leqslant \widetilde{C}_{d,p} R^{1 - \frac{1}{pd}}\, \mathrm{SW}_p(\alpha, \beta)^{\frac{1}{pd}}, \qquad \text{equivalently} \qquad \mathrm{SW}_p(\alpha, \beta) \geqslant \widetilde{c}_{d,p} R \left(\frac{\mathcal{W}_p(\alpha, \beta)}{R}\right)^{pd}.$$

For $p = 1$, the exponent $1/d$ in this last comparison is sharp.

Proof. Non-negativity and symmetry follow from the one-dimensional Wasserstein distance. For the triangle inequality, apply the triangle inequality of $\mathcal{W}_p$ for each direction θ and then Minkowski's inequality in $L^p(\mathbb{S}^{d-1})$.

If $\mathrm{SW}_p(\alpha,\beta)=0$, then $(P_\theta)_\sharp\alpha=(P_\theta)_\sharp\beta$ for almost every direction. By continuity of characteristic functions this holds for all directions, and the Cramér–Wold theorem implies $\alpha=\beta$. This proves separation.

Let π be any coupling between α and β. Projecting π gives an admissible coupling between the corresponding one-dimensional projections, hence

$$\mathcal{W}_p\left((P_\theta)_\sharp\alpha,(P_\theta)_\sharp\beta\right)^p\leqslant\int_{\mathbb{R}^d\times\mathbb{R}^d}|\langle\theta,\,x-y\rangle|^p\,\mathrm{d}\pi(x,y).$$

Rotational invariance of σ gives

$$\int_{\mathbb{S}^{d-1}}|\langle\theta,\,z\rangle|^p\,\mathrm{d}\sigma(\theta)=\kappa_{d,p}\|z\|^p.$$

Integrating the first inequality in θ and optimizing over π proves the stated direct comparison. The beta-integral formula for the first coordinate of a uniform point on the sphere gives the displayed value of $\kappa_{d,p}$; in particular, $\kappa_{d,2}=1/d$.

To prove the topological statement, first suppose that $\mathcal{W}_p(\alpha_n,\alpha)\to0$. The inequality $\mathrm{SW}_p\leqslant\mathcal{W}_p$ immediately gives $\mathrm{SW}_p(\alpha_n,\alpha)\to0$. Conversely, suppose that $\mathrm{SW}_p(\alpha_n,\alpha)\to0$. The spherical identity

$$\int_{\mathbb{S}^{d-1}}\int_{\mathbb{R}^d}|\langle\theta,x\rangle|^p\,\mathrm{d}\alpha_n(x)\mathrm{d}\sigma(\theta)=\kappa_{d,p}\int_{\mathbb{R}^d}\|x\|^p\,\mathrm{d}\alpha_n(x),\qquad\kappa_{d,p}>0,$$

and the triangle inequality in $L^p(\mathbb{S}^{d-1}\times(0,1))$ show that the pth moments of (α_n) are uniformly bounded. Hence the sequence is tight. From every subsequence one can extract a further subsequence for which the projected $\mathcal{W}_p$ distances converge to zero for σ-a.e. θ. Every weak limit therefore has the same one-dimensional projections as α, so Cramér–Wold identifies it with α. Thus $\alpha_n\rightharpoonup\alpha$. Finally, let $Q_{n,\theta}$ and Q_θ be the quantile functions of the projected measures. The one-dimensional quantile formula gives

$$\mathrm{SW}_p(\alpha_n,\alpha)=\|Q_{n,\theta}-Q_\theta\|_{L^p(\mathbb{S}^{d-1}\times(0,1))}.$$

Hence the corresponding L^p norms converge. By the spherical identity, their pth powers are respectively $\kappa_{d,p}\int\|x\|^p\mathrm{d}\alpha_n(x)$ and $\kappa_{d,p}\int\|x\|^p\mathrm{d}\alpha(x)$. Thus the pth moments converge. Weak convergence plus convergence of these moments is equivalent to convergence in $\mathcal{W}_p$.

The reverse estimates on compact sets are deeper and use the Radon structure of slices. Bonnotte's Lemma 5.1.4 [77] proves

$$\mathcal{W}_1(\alpha,\beta)\leqslant C_dR^{\frac{d}{d+1}}\,\mathrm{SW}_1(\alpha,\beta)^{\frac{1}{d+1}}$$

by mollifying a Kantorovich–Rubinstein test function, representing the regularized test through one-dimensional projections, bounding the resulting action by SW_1, and optimizing the smoothing scale. On B_R,

$$\mathcal{W}_p^p\leqslant(2R)^{p-1}\,\mathcal{W}_1,\qquad\mathrm{SW}_1\leqslant\mathrm{SW}_p,$$

where the first inequality follows by evaluating the p-cost on a $\mathcal{W}_1$-optimal coupling, and the second follows from the monotonicity of one-dimensional Wasserstein distances and Hölder's inequality in θ. These inequalities give Bonnotte's stated general-p estimate. The sharper $p=1$ comparison theorem of Carlier, Figalli, Mérigot and Wang [111] replaces $1/(d+1)$ by the optimal exponent $1/d$. Combining it with the same two bounded-support inequalities gives

$$\mathcal{W}_p^p\leqslant C_{d,p}R^{p-\frac{1}{d}}\,\mathrm{SW}_p^{\frac{1}{d}},$$

which is equivalent to the final two-sided formulation in the statement. This supplies a lower bound on SW_p in terms of $\mathcal{W}_p$ for every p, but the asserted sharpness concerns only $p=1$. □

The preceding estimates compare finite displacements. The infinitesimal comparison is clearest along smooth Brenier perturbations: the usual Wasserstein distance sees the full $L^2(\alpha)$ norm of the displacement, whereas the sliced distance sees only its one-dimensional projections. The same result also isolates the different behavior of finite atomic curves.

Proposition 11.26 (First-order comparison: strictness and atomic equality)**.** *Let $\alpha \in \mathcal{P}_2(\mathbb{R}^d)$ be absolutely continuous, and let $\varphi \in C^2(\mathbb{R}^d)$ be such that $\nabla\varphi \in L^2(\alpha;\mathbb{R}^d)$ and $\nabla^2\varphi$ is bounded. For $|t|$ small enough, set*

$$T_t(x) \coloneqq x + t\nabla\varphi(x), \qquad \alpha_t \coloneqq (T_t)_\sharp\alpha.$$

Then T_t is the quadratic Brenier map from α to α_t for $|t|$ small enough, and

$$\mathcal{W}_2(\alpha,\alpha_t)^2 = t^2\int_{\mathbb{R}^d}\|\nabla\varphi(x)\|^2\mathrm{d}\alpha(x).$$

Moreover,

$$\mathrm{SW}_2(\alpha,\alpha_t)^2 \leqslant \frac{t^2}{d}\int_{\mathbb{R}^d}\|\nabla\varphi(x)\|^2\mathrm{d}\alpha(x).$$

The upper bound can be strict at order t^2. Indeed, let $d \geqslant 2$, let $\alpha = \mathcal{N}(0,\mathrm{Id})$, and take $\varphi(x) = \frac{1}{2}x^\top Ax$ for a symmetric matrix A that is not a scalar multiple of Id*. Then*

$$\lim_{t\to0}\frac{\mathrm{SW}_2(\alpha,\alpha_t)^2}{t^2} = \frac{\mathrm{tr}(A)^2+2\,\mathrm{tr}(A^2)}{d(d+2)} < \frac{\mathrm{tr}(A^2)}{d} = \frac{1}{d}\lim_{t\to0}\frac{\mathcal{W}_2(\alpha,\alpha_t)^2}{t^2}.$$

Separately, equality holds at first order along every finite atomic curve with fixed weights. More precisely, let $a_i > 0$ satisfy $\sum_i a_i = 1$, let $x_1,\dots,x_n$ be pairwise distinct, and let $v_i \in \mathbb{R}^d$. If $\eta = \sum_{i=1}^n a_i\delta_{x_i}$ and $\eta_t = \sum_{i=1}^n a_i\delta_{x_i+tv_i}$, then

$$\lim_{t\to0}\frac{\mathrm{SW}_2(\eta,\eta_t)^2}{t^2} = \frac{1}{d}\sum_{i=1}^n a_i\|v_i\|^2 = \frac{1}{d}\lim_{t\to0}\frac{\mathcal{W}_2(\eta,\eta_t)^2}{t^2}.$$

Proof. The map T_t is the gradient of

$$\psi_t(x) \coloneqq \frac{1}{2}\|x\|^2 + t\varphi(x).$$

Since $\nabla^2\varphi$ is bounded, $\nabla^2\psi_t = \mathrm{Id} + t\nabla^2\varphi$ is positive semidefinite for $|t|$ small enough. Hence ψ_t is convex and $T_t = \nabla\psi_t$ is a Brenier map. Its graph is cyclically monotone, so the Brenier optimality criterion gives the optimality of $(\mathrm{Id}, T_t)_\sharp\alpha$, and

$$\mathcal{W}_2(\alpha,\alpha_t)^2 = \int\|x - T_t(x)\|^2\mathrm{d}\alpha(x) = t^2\int\|\nabla\varphi(x)\|^2\mathrm{d}\alpha(x).$$

For a fixed direction θ, the coupling $\big(P_\theta, P_\theta\circ T_t\big)_\sharp\alpha$ is admissible between $(P_\theta)_\sharp\alpha$ and $(P_\theta)_\sharp\alpha_t$. Therefore

$$\mathcal{W}_2\big((P_\theta)_\sharp\alpha,(P_\theta)_\sharp\alpha_t\big)^2 \leqslant t^2\int|\langle\theta,\,\nabla\varphi(x)\rangle|^2\,\mathrm{d}\alpha(x).$$

Integrating over the sphere and using

$$\int_{\mathbb{S}^{d-1}}|\langle\theta,\,w\rangle|^2\,\mathrm{d}\sigma(\theta) = \frac{1}{d}\|w\|^2$$

gives the sliced estimate.

For the Gaussian example, $T_t = \mathrm{Id} + tA$ and, for every $\theta \in \mathbb{S}^{d-1}$, the projected source and target are centered one-dimensional Gaussians with standard deviations 1 and $\|(\mathrm{Id}+tA)\theta\|$. Consequently,

$$\lim_{t\to0}\frac{1}{t^2}\,\mathcal{W}_2\big((P_\theta)_\sharp\alpha,(P_\theta)_\sharp\alpha_t\big)^2 = (\theta^\top A\theta)^2.$$

The spherical fourth-moment identity gives

$$\int_{\mathbb{S}^{d-1}}(\theta^\top A\theta)^2\mathrm{d}\sigma(\theta) = \frac{\mathrm{tr}(A)^2+2\,\mathrm{tr}(A^2)}{d(d+2)}.$$

On the other hand, the Brenier formula above gives $\mathcal{W}_2(\alpha,\alpha_t)^2 = t^2\,\mathrm{tr}(A^2)$. The strict inequality follows from $\mathrm{tr}(A)^2 < d\,\mathrm{tr}(A^2)$ when A is symmetric and non-scalar.

Finally, consider the atomic curve. For all sufficiently small t, the diagonal coupling between x_i and $x_i + tv_i$ is optimal for $\mathcal{W}_2$. Indeed, writing $c_{i,j}(t) = \|x_i - x_j - tv_j\|^2$, the discrete dual potentials $f_i = 0$ and $g_j = c_{j,j}(t)$ are feasible for small t, because $c_{j,j}(t) < c_{i,j}(t)$ whenever $i \neq j$, and they certify the diagonal plan. Thus

$$\mathcal{W}_2(\eta, \eta_t)^2 = t^2 \sum_{i=1}^{n} a_i \|v_i\|^2.$$

For every θ outside the finite union of great spheres $\{\langle \theta,\, x_i - x_j\rangle = 0\}$, the projected atoms are distinct and their order is unchanged for sufficiently small t. The one-dimensional monotone coupling then pairs each projected atom with its perturbation, so

$$\lim_{t\to 0} \frac{1}{t^2}\, \mathcal{W}_2\left((P_\theta)_\sharp \eta, (P_\theta)_\sharp \eta_t\right)^2 = \sum_{i=1}^{n} a_i\, |\langle \theta,\, v_i\rangle|^2 .$$

The diagonal projected coupling provides the same expression as an integrable upper bound for every t. Dominated convergence and the spherical second-moment identity therefore give the asserted sliced limit. □

Intrinsic sliced length. In dimension one, $\mathrm{SW}_2 = \mathcal{W}_2$. The geometry changes qualitatively for $d \geqslant 2$: SW_2 is not a length metric [415]. Its intrinsic, or path, metric is

$$\ell_{\mathrm{SW}_2}(\alpha, \beta) := \inf_{\substack{\gamma_0 = \alpha,\ \gamma_1 = \beta \\ \gamma\ \mathrm{SW}_2\text{-absolutely continuous}}} \int_0^1 |\dot\gamma_t|_{\mathrm{SW}_2}\, \mathrm{d}t, \tag{11.35}$$

where $|\dot\gamma_t|_{\mathrm{SW}_2}$ is the metric derivative. Park and Slepčev prove that the infimum is attained, so $(\mathcal{P}_2(\mathbb{R}^d), \ell_{\mathrm{SW}_2})$ is a geodesic space. Every path length dominates the distance between its endpoints, while Proposition 11.25 bounds the SW_2-length of a $\mathcal{W}_2$-geodesic. Consequently,

$$\mathrm{SW}_2 \leqslant \ell_{\mathrm{SW}_2} \leqslant d^{-1/2}\, \mathcal{W}_2 .$$

Neither inequality identifies the intrinsic geometry with a rescaled Wasserstein geometry. Proposition 11.26 shows that the sliced metric derivative equals $d^{-1/2}$ times the Wasserstein metric derivative along finite atomic curves with fixed weights. By contrast, the Gaussian deformation in the same proposition is strictly slower for SW_2. The sliced speed along $s \mapsto (\mathrm{Id} + sA)_\sharp \mathcal{N}(0, \mathrm{Id})$ depends continuously on s, so this strictness persists on a short interval. Integrating along that $\mathcal{W}_2$-geodesic gives, for every sufficiently small $t \neq 0$,

$$\ell_{\mathrm{SW}_2}(\alpha, \alpha_t) < d^{-1/2}\, \mathcal{W}_2(\alpha, \alpha_t).$$

There is, more strongly, no reverse comparison by a constant depending only on the dimension. Example 3.2 of Park and Slepčev considers two nearby parallel line segments moving in opposite directions. Most projected motions cancel after slicing: for every $\delta > 0$, the construction produces a compactly supported, atomless curve $(\alpha_t^\delta)_t$ such that, for small $t \geqslant 0$,

$$\left|\dot\alpha_t^\delta\right|_{\mathcal{W}_2} = 1, \qquad \left|\dot\alpha_t^\delta\right|_{\mathrm{SW}_2} \leqslant C_d(t + \delta).$$

Since the first identity gives $\mathcal{W}_2(\alpha_0^\delta, \alpha_h^\delta) = h + o(h)$, while the second gives $\ell_{\mathrm{SW}_2}(\alpha_0^\delta, \alpha_h^\delta) \leqslant C_d(\delta h + h^2/2)$, one obtains

$$\limsup_{h\downarrow 0} \frac{\ell_{\mathrm{SW}_2}(\alpha_0^\delta, \alpha_h^\delta)}{\mathcal{W}_2(\alpha_0^\delta, \alpha_h^\delta)} \leqslant C_d\delta.$$

Letting $\delta \to 0$ yields

$$\inf_{\substack{\alpha,\beta\in\mathcal{P}_2(\mathbb{R}^d) \\ \alpha\neq\beta}} \frac{\ell_{\mathrm{SW}_2}(\alpha, \beta)}{\mathcal{W}_2(\alpha, \beta)} = 0. \tag{11.36}$$

Thus ℓ_{SW_2} and $\mathcal{W}_2$ are not globally bi-Lipschitz equivalent, even after inserting the factor $d^{-1/2}$.

The atomic equality is nevertheless stable in its natural regime. If η is a fixed finite atomic measure with positive masses and separated atoms, Theorem 5.5 of Park and Slepčev gives

$$\frac{\mathrm{SW}_2(\eta,\beta)}{\mathcal{W}_2(\eta,\beta)} \longrightarrow \frac{1}{\sqrt{d}}, \qquad \frac{\ell_{\mathrm{SW}_2}(\eta,\beta)}{\mathcal{W}_2(\eta,\beta)} \longrightarrow \frac{1}{\sqrt{d}} \quad \text{as } \mathcal{W}_\infty(\eta,\beta)\to 0,$$

along $\beta \neq \eta$; see (4.5). By contrast, under suitable common-support and density bounds in the diffuse regime, their Theorem 5.2 shows that both SW_2 and ℓ_{SW_2} are locally equivalent to the negative Sobolev norm $\dot{H}^{-(d+1)/2}$, whereas the infinitesimal geometry of $\mathcal{W}_2$ around a positive density is of $\dot{H}^{-1}$ type. These results settle the comparison: the intrinsic sliced metric agrees asymptotically with $d^{-1/2}\mathcal{W}_2$ near finite atomic measures, but it is a genuinely different metric globally and around diffuse measures.

> **Remark 11.27 (Hilbert embedding for SW_2).** In one dimension, $\mathcal{W}_2$ is the $L^2(0,1)$ distance between quantile functions. Hence
>
> $$\mathrm{SW}_2(\alpha,\beta)^2 = \int_{\mathbb{S}^{d-1}} \int_0^1 \left|F^{-1}_{\theta,\alpha}(u) - F^{-1}_{\theta,\beta}(u)\right|^2 \mathrm{d}u\,\mathrm{d}\sigma(\theta),$$
>
> where $F^{-1}_{\theta,\alpha}$ is the quantile of $(P_\theta)_\sharp\alpha$. Thus SW_2 is a Hilbertian distance after embedding each measure into its field of projected quantiles. Consequently, $\exp(-\gamma\,\mathrm{SW}_2^2)$ is a positive definite kernel on probability measures for every $\gamma > 0$. The compact reverse estimates in Proposition 11.25 explain why sliced distances metrize the same topology on bounded sets, while the dimension-dependent powers prevent a dimension-free bi-Lipschitz comparison with $\mathcal{W}_p$.

Sliced Wasserstein between Gaussians. Gaussian projections remain Gaussian, so quadratic sliced transport reduces to the one-dimensional Gaussian formula in every direction. This gives an exact angular representation and makes its relation with the Bures covariance geometry explicit.

Proposition 11.28 (Sliced Wasserstein between Gaussians). *Let $\alpha = \mathcal{N}(m_\alpha,\Sigma_\alpha)$ and $\beta = \mathcal{N}(m_\beta,\Sigma_\beta)$ on $\mathbb{R}^d$, with positive-semidefinite covariance matrices. Then*

$$\mathrm{SW}_2(\alpha,\beta)^2 = \frac{\|m_\alpha - m_\beta\|^2}{d} + \int_{\mathbb{S}^{d-1}} \left(\sqrt{\theta^\top\Sigma_\alpha\theta} - \sqrt{\theta^\top\Sigma_\beta\theta}\right)^2 \mathrm{d}\sigma(\theta). \tag{11.37}$$

It satisfies

$$\mathrm{SW}_2(\alpha,\beta)^2 \leqslant \frac{1}{d}\,\mathcal{W}_2(\alpha,\beta)^2.$$

Proof. For every $\theta \in \mathbb{S}^{d-1}$,

$$(P_\theta)_\sharp\alpha = \mathcal{N}\big(\theta^\top m_\alpha, \theta^\top\Sigma_\alpha\theta\big),$$

and similarly for β. The one-dimensional Gaussian formula therefore gives the integrand in (11.37). Expanding the square and using $\int_{\mathbb{S}^{d-1}} \theta\theta^\top\,\mathrm{d}\sigma(\theta) = \mathrm{Id}/d$ gives

$$\mathrm{SW}_2(\alpha,\beta)^2 = \frac{\|m_\alpha - m_\beta\|^2 + \mathrm{tr}(\Sigma_\alpha) + \mathrm{tr}(\Sigma_\beta)}{d} - 2\int_{\mathbb{S}^{d-1}} \sqrt{(\theta^\top\Sigma_\alpha\theta)(\theta^\top\Sigma_\beta\theta)}\,\mathrm{d}\sigma(\theta).$$

First suppose that both covariances are positive definite and let A be the Bures transport matrix (2.18), so that $A\Sigma_\alpha A = \Sigma_\beta$. Cauchy–Schwarz for the inner product $\langle x, y\rangle_{\Sigma_\alpha} = x^\top\Sigma_\alpha y$ gives, for every θ,

$$\theta^\top\Sigma_\alpha A\theta \leqslant \sqrt{(\theta^\top\Sigma_\alpha\theta)(\theta^\top A\Sigma_\alpha A\theta)} = \sqrt{(\theta^\top\Sigma_\alpha\theta)(\theta^\top\Sigma_\beta\theta)}.$$

Integrating and using the spherical second-moment identity yields

$$\int_{\mathbb{S}^{d-1}} \sqrt{(\theta^\top\Sigma_\alpha\theta)(\theta^\top\Sigma_\beta\theta)}\mathrm{d}\sigma(\theta) \geqslant \frac{\mathrm{tr}(\Sigma_\alpha A)}{d} = \frac{1}{d}\,\mathrm{tr}\Big[\big(\Sigma_\alpha^{1/2}\Sigma_\beta\Sigma_\alpha^{1/2}\big)^{1/2}\Big].$$

The general positive-semidefinite case follows by continuity. Combining this bound with (11.37) and the Bures formula (2.19) proves the comparison. □

Formula (11.37) is exact, but for generic anisotropic covariances its spherical integral has no Bures-like matrix-square-root simplification and is evaluated by angular quadrature or Monte Carlo. The comparison in the proposition is the Gaussian specialization of Proposition 11.25. Its normalization matters: for $d > 1$ and $\alpha \neq \beta$, SW_2 is necessarily strictly smaller than $\mathcal{W}_2$, whereas the two distances coincide in dimension one.

Remark 11.29 (Proportional and rank-one covariances). If $\Sigma_\alpha = s^2\Sigma$ and $\Sigma_\beta = t^2\Sigma$, then

$$\mathrm{SW}_2(\alpha,\beta)^2 = \frac{1}{d}\mathcal{W}_2(\alpha,\beta)^2 = \frac{\|m_\alpha - m_\beta\|^2 + (s-t)^2\,\mathrm{tr}(\Sigma)}{d}.$$

This includes two isotropic covariances. If both covariances are positive definite, these proportional pairs are exactly the equality cases; one isotropic covariance alone does not suffice.
For the rank-one covariances $\Sigma_\alpha = a^2uu^\top$ and $\Sigma_\beta = b^2vv^\top$, set $\chi = |\langle u, v\rangle|$. Formula (11.37) reduces to

$$\mathrm{SW}_2(\alpha,\beta)^2 = \frac{1}{d}\left[\|m_\alpha - m_\beta\|^2 + a^2 + b^2 - \frac{4ab}{\pi}\left(\sqrt{1-\chi^2} + \chi\arcsin(\chi)\right)\right].$$

Thus aligned directions attain the upper bound, whereas centered, orthogonal and equal-scale covariances satisfy $\mathrm{SW}_2 = \sqrt{1-2/\pi}\,\mathcal{W}_2/\sqrt{d}$. The associated Gaussian sliced-Wasserstein flow is revisited in Proposition 16.23.

L^q-sliced, max-sliced and subspace variants. The exponent used to aggregate projection directions need not coincide with the transport exponent. A second exponent q controls this aggregation: finite q pools information from many directions, whereas $q = \infty$ retains only the most discriminating one. Independently, replacing lines by k-dimensional subspaces preserves more correlations at the price of solving higher-dimensional projected OT problems. These two choices fit into one family.

Definition 11.30 (L^q-sliced and subspace-sliced Wasserstein)**.** Let $p \in [1,\infty)$, $q \in [1,\infty]$, and $1 \leqslant k \leqslant d$. Denote by

$$\mathrm{St}(d,k) := \left\{U \in \mathbb{R}^{d\times k}\ ;\ U^\top U = \mathrm{Id}_k\right\}$$

the Stiefel manifold, equipped with its normalized invariant probability measure $\sigma_{d,k}$. For $q < \infty$, define

$$\mathrm{SW}_{p,q,k}(\alpha,\beta) := \left(\int_{\mathrm{St}(d,k)} \mathcal{W}_p\left((U^\top)_\sharp\alpha, (U^\top)_\sharp\beta\right)^q \mathrm{d}\sigma_{d,k}(U)\right)^{1/q}, \tag{11.38}$$

and extend the definition to $q = \infty$ by

$$\mathrm{SW}_{p,\infty,k}(\alpha,\beta) := \sup_{U\in\mathrm{St}(d,k)} \mathcal{W}_p\left((U^\top)_\sharp\alpha, (U^\top)_\sharp\beta\right).$$

The supremum is attained because the Stiefel manifold is compact and the projected Wasserstein distance depends continuously on U. Moreover, the integrand is unchanged under $U \mapsto UQ$ for $Q \in \mathrm{O}(k)$, because $Q^\top$ is an isometry of $\mathbb{R}^k$. Thus the construction depends only on the subspace spanned by the columns of U: the Stiefel manifold is a convenient parametrization of the corresponding Grassmann manifold. We use the abbreviations

$$\mathrm{SW}_{p,q} := \mathrm{SW}_{p,q,1}, \qquad \mathrm{SW}_p = \mathrm{SW}_{p,p}, \qquad \mathrm{MaxSW}_{p,k} := \mathrm{SW}_{p,\infty,k}, \qquad \mathrm{MaxSW}_p = \mathrm{SW}_{p,\infty}.$$

Thus $k = 1$ gives line-based slicing, $q = \infty$ gives max-slicing, and $k = d$ gives $\mathrm{SW}_{p,q,d} = \mathcal{W}_p$ for every q. Quadratic max-slicing is useful when only a small set of directions carries the discrepancy, for instance in generative modeling [186]. For $p = 2$, the k-dimensional max-sliced case is the projection-robust Wasserstein distance, which is related to the subspace-robust distances studied later in Section 11.5 [419].

Metricity and the direct comparison with $\mathcal{W}_p$ carry over from ordinary slicing. The topological statement requires slightly more care when $q < p$: although an abstract L^q norm need not control an L^p norm, projected Wasserstein profiles have enough continuity and moment control to rule out concentration in a vanishing set of subspaces. The next proposition records these facts and separates the roles of p, q, and k.

Proposition 11.31 (Basic bounds for sliced variants)**.** *Let $p \in [1,\infty)$, $1 \leqslant q \leqslant r \leqslant \infty$, and $1 \leqslant k \leqslant \ell \leqslant d$. Each $\mathrm{SW}_{p,q,k}$ is a finite distance on $\mathcal{P}_p(\mathbb{R}^d)$ and metrizes the same topology as $\mathcal{W}_p$. For every $\alpha,\beta \in \mathcal{P}_p(\mathbb{R}^d)$,*

$$\mathrm{SW}_{p,q,k}(\alpha,\beta) \leqslant \mathrm{SW}_{p,r,k}(\alpha,\beta) \leqslant \mathcal{W}_p(\alpha,\beta), \tag{11.39}$$

$$\mathrm{SW}_{p,q,k}(\alpha,\beta) \leqslant \mathrm{SW}_{p,q,\ell}(\alpha,\beta) \leqslant \mathcal{W}_p(\alpha,\beta). \tag{11.40}$$

For a unit vector $e \in \mathbb{S}^{d-1}$, set

$$\kappa_{d,k,p} := \int_{\mathrm{St}(d,k)} \|U^\top e\|^p \mathrm{d}\sigma_{d,k}(U) = \frac{\Gamma(d/2)\Gamma((k+p)/2)}{\Gamma(k/2)\Gamma((d+p)/2)}.$$

This constant is independent of e, satisfies $\kappa_{d,d,p} = 1$ and $\kappa_{d,1,p} = \kappa_{d,p}$ from Proposition 11.25, and gives the sharper bounds

$$\mathrm{SW}_{p,q,k}(\alpha,\beta) \leqslant \begin{cases} \kappa_{d,k,p}^{1/p}\, \mathcal{W}_p(\alpha,\beta), & 1 \leqslant q \leqslant p, \\ \kappa_{d,k,p}^{1/q}\, \mathcal{W}_p(\alpha,\beta), & p \leqslant q < \infty, \\ \mathcal{W}_p(\alpha,\beta), & q = \infty. \end{cases} \tag{11.41}$$

Moreover,

$$\mathrm{SW}_p(\alpha,\beta) \leqslant \mathrm{SW}_{p,p,k}(\alpha,\beta) \leqslant \kappa_{d,k,p}^{1/p}\, \mathcal{W}_p(\alpha,\beta). \tag{11.42}$$

In particular, $\kappa_{d,k,2} = k/d$ and $\mathrm{SW}_{2,2,k}^2 \leqslant (k/d)\mathcal{W}_2^2$. If $q \geqslant p$, the compact-support reverse estimates in Proposition 11.25 remain valid with $\mathrm{SW}_{p,q,k}$ in place of SW_p.

Proof. Write $D_U(\alpha,\beta) := \mathcal{W}_p((U^\top)_\sharp\alpha, (U^\top)_\sharp\beta)$. Non-negativity and symmetry are inherited from $\mathcal{W}_p$. Minkowski's inequality in $L^q(\sigma_{d,k})$, or the supremum triangle inequality when $q = \infty$, proves the triangle inequality. The reverse triangle inequality for $\mathcal{W}_p$ shows that $U \mapsto D_U(\alpha,\beta)$ is continuous. Indeed, for $U_n \to U$,

$$\mathcal{W}_p\left((U_n{}^\top)_\sharp\alpha, (U^\top)_\sharp\alpha\right) \leqslant \|U_n - U\|_{\mathrm{op}} \left(\int \|x\|^p \mathrm{d}\alpha(x)\right)^{1/p},$$

and likewise for β. If $\mathrm{SW}_{p,q,k}(\alpha,\beta) = 0$, then $D_U = 0$ for all U: this is immediate for $q = \infty$, while for finite q it follows from continuity and the full support of $\sigma_{d,k}$. For any $\theta \in \mathbb{S}^{d-1}$, choose a frame U whose first column is θ. Equality of the two k-dimensional projected measures then implies equality of their first-coordinate marginals, hence $(P_\theta)_\sharp\alpha = (P_\theta)_\sharp\beta$. The Cramér–Wold theorem gives $\alpha = \beta$, proving separation.

Every $U^\top$ is 1-Lipschitz, so $D_U \leqslant \mathcal{W}_p(\alpha,\beta)$. Monotonicity of L^q norms on a probability space proves (11.39). To compare projection dimensions, let $V \in \mathrm{St}(d,\ell)$ and $R \in \mathrm{St}(\ell,k)$. Since $(VR)^\top = R^\top V^\top$ and $R^\top$ is 1-Lipschitz, $D_{VR}(\alpha,\beta) \leqslant D_V(\alpha,\beta)$. Under the product of the invariant measures, VR is uniformly distributed on $\mathrm{St}(d,k)$. Integration proves (11.40) for finite q; for $q = \infty$, extend each k-frame to an ℓ-frame and use the same contraction.

For the sharper constants, let $\pi \in \Gamma(\alpha,\beta)$. Then

$$D_U(\alpha,\beta)^p \leqslant \int \|U^\top(x-y)\|^p \mathrm{d}\pi(x,y).$$

Integration in U, followed by minimization in π, gives $\|D_\cdot\|_{L^p(\sigma_{d,k})}^p \leqslant \kappa_{d,k,p}\, \mathcal{W}_p(\alpha,\beta)^p$. If $q \leqslant p$, use $\|D_\cdot\|_{L^q} \leqslant \|D_\cdot\|_{L^p}$. If $p \leqslant q < \infty$, use $D_U \leqslant \mathcal{W}_p$ to obtain

$$\int D_U^q \mathrm{d}\sigma_{d,k}(U) \leqslant \mathcal{W}_p(\alpha,\beta)^{q-p} \int D_U^p \mathrm{d}\sigma_{d,k}(U).$$

This proves (11.41). The beta-integral for the pth power of the length of a random k-dimensional projection gives the displayed formula for $\kappa_{d,k,p}$.

Finally, let $\omega \in \mathbb{S}^{k-1}$. Since $P_{U\omega} = P_\omega \circ U^\top$, projection contraction gives

$$\mathcal{W}_p\left((P_{U\omega})_\sharp\alpha, (P_{U\omega})_\sharp\beta\right) \leqslant D_U(\alpha,\beta).$$

Averaging its pth power first over ω and then over U makes $U\omega$ uniform on $\mathbb{S}^{d-1}$ and proves the first inequality in (11.42). For $q \geqslant p$, (11.39) gives $\mathrm{SW}_p \leqslant \mathrm{SW}_{p,p,k} \leqslant \mathrm{SW}_{p,q,k}$, which transfers the stated reverse estimates.

It remains to prove the topological assertion. For $\eta \in \mathcal{P}_p(\mathbb{R}^d)$, set $M_p(\eta) := (\int \|x\|^p \mathrm{d}\eta(x))^{1/p}$. Since the only coupling between $(U^\top)_\sharp\eta$ and δ_0 sends every point to the origin,

$$D_U(\eta,\delta_0)^p = \int \|U^\top x\|^p \mathrm{d}\eta(x), \qquad \int D_U(\eta,\delta_0)^p \mathrm{d}\sigma_{d,k}(U) = \kappa_{d,k,p} M_p(\eta)^p.$$

Also $D_U(\eta, \delta_0) \leqslant M_p(\eta)$. Consequently,

$$\mathrm{SW}_{p,q,k}(\eta, \delta_0) \geqslant \begin{cases} \kappa_{d,k,p}^{1/q} M_p(\eta), & 1 \leqslant q \leqslant p, \\ \kappa_{d,k,p}^{1/p} M_p(\eta), & p \leqslant q \leqslant \infty. \end{cases} \tag{11.43}$$

Indeed, the second line follows from monotonicity of L^q norms. For $q \leqslant p$ and $M_p(\eta) > 0$, the pointwise bound $D_U \leqslant M_p(\eta)$ gives $D_U^q \geqslant M_p(\eta)^{q-p} D_U^p$; the zero-moment case is immediate.

Suppose now that $\mathrm{SW}_{p,q,k}(\alpha_n, \alpha) \to 0$. The triangle inequality and (11.43) imply $\sup_n M_p(\alpha_n) < \infty$. Write $D_{n,U} := D_U(\alpha_n, \alpha)$. The reverse triangle inequality and the coupling of each point with itself give

$$|D_{n,U} - D_{n,V}| \leqslant \|U - V\|_{\mathrm{op}} \big(M_p(\alpha_n) + M_p(\alpha)\big),$$

so $(D_{n,\cdot})_n$ is equicontinuous on the compact Stiefel manifold. For $q < \infty$, equicontinuity and the full support of $\sigma_{d,k}$ upgrade $\|D_{n,\cdot}\|_{L^q} \to 0$ to $\|D_{n,\cdot}\|_{L^\infty} \to 0$; for $q = \infty$ this is immediate. Given $\theta \in \mathbb{S}^{d-1}$, choose a frame U whose first column is θ. Projection contraction then yields

$$\mathcal{W}_p\left((P_\theta)_\sharp \alpha_n, (P_\theta)_\sharp \alpha\right) \leqslant D_{n,U},$$

uniformly in θ. Hence $\mathrm{SW}_p(\alpha_n, \alpha) \to 0$, and Proposition 11.25 gives $\mathcal{W}_p(\alpha_n, \alpha) \to 0$. The converse follows directly from $\mathrm{SW}_{p,q,k} \leqslant \mathcal{W}_p$. □

Min-SW lifted transport plans. The preceding constructions compare projected measures. Min-SW uses a projection differently: it lifts the one-dimensional monotone coupling back to the ambient space and retains the lift with the smallest quadratic cost. Consider equal-weight empirical measures $\alpha = n^{-1} \sum_i \delta_{x_i}$ and $\beta = n^{-1} \sum_i \delta_{y_i}$. For a direction θ along which both projected point families have distinct coordinates, let $\sigma_\theta, \tau_\theta \in \mathfrak{S}_n$ be their sorting permutations,

$$\langle x_{\sigma_\theta(1)}, \theta\rangle < \cdots < \langle x_{\sigma_\theta(n)}, \theta\rangle, \qquad \langle y_{\tau_\theta(1)}, \theta\rangle < \cdots < \langle y_{\tau_\theta(n)}, \theta\rangle.$$

Lifting the one-dimensional monotone matching gives

$$\pi_\theta = \frac{1}{n} \sum_{i=1}^n \delta_{(x_{\sigma_\theta(i)}, y_{\tau_\theta(i)})} \in \mathcal{U}(\alpha, \beta). \tag{11.44}$$

The cost of (11.44) is an upper bound on $\mathcal{W}_2(\alpha, \beta)^2$, and Min-SW minimizes this upper bound over θ. This inexpensive feasible-plan construction was introduced by Mahey, Chapel, Gasso, Bonet and Courty [364].

Projected ties make the extension beyond this generic empirical setting less immediate. At a tie, the sorting permutations are not unique. Breaking ties by labels is not invariant under a relabeling of the atoms, while a lexicographic rule depends on an auxiliary coordinate system; both choices can be discontinuous under perturbations. For a non-discrete measure, a projection fiber may moreover carry an entire conditional distribution, so there are no labels to sort. Tanguy, Chapel and Delon [516] avoid arbitrary tie breaking by optimizing over all compatible lifts. Set

$$\alpha_\theta = (P_\theta)_\sharp \alpha, \qquad \beta_\theta = (P_\theta)_\sharp \beta,$$

and let

$$\varpi_\theta = (q_{\alpha_\theta}, q_{\beta_\theta})_\sharp \mathrm{Leb}_{[0,1]}$$

be their canonical monotone coupling, as in Theorem 3.39. Define

Definition 11.32 (Min-SW discrepancy and lifted plans)**.**

$$C_\theta(\alpha, \beta) := \left\{\pi \in \mathcal{U}(\alpha, \beta) \ : \ (P_\theta, P_\theta)_\sharp \pi = \varpi_\theta\right\}. \tag{11.45}$$

The Min-SW discrepancy is

$$\text{Min-SW}_2(\alpha, \beta)^2 := \min_{\theta \in \mathbb{S}^{d-1}} \min_{\pi \in C_\theta(\alpha,\beta)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - y\|^2 \mathrm{d}\pi(x, y). \tag{11.46}$$

This set is nonempty. Indeed, disintegrate

$$\alpha(\mathrm{d}x) = \int_{\mathbb{R}} \alpha^s(\mathrm{d}x)\mathrm{d}\alpha_\theta(s), \qquad \beta(\mathrm{d}y) = \int_{\mathbb{R}} \beta^t(\mathrm{d}y)\mathrm{d}\beta_\theta(t)$$

along P_θ. The conditionally independent lift

$$\pi^0_\theta(\mathrm{d}x,\mathrm{d}y) = \int_{\mathbb{R}^2} \alpha^s(\mathrm{d}x)\beta^t(\mathrm{d}y)\mathrm{d}\omega_\theta(s,t) \tag{11.47}$$

belongs to $C_\theta(\alpha,\beta)$. This proves that the representation-invariant measure-level discrepancy in Definition 11.32 is well posed. Both minima are attained for $\alpha,\beta \in \mathcal{P}_2(\mathbb{R}^d)$ [516]. Unlike the particular lift (11.47), the inner minimization may correlate the conditional laws on $P_\theta^{-1}(s)\times P_\theta^{-1}(t)$; it chooses the least costly transverse coupling compatible with the prescribed projected coupling. When the empirical projections have no ties, $C_\theta(\alpha,\beta)=\{\pi_\theta\}$ and (11.46) reduces to (11.44).

Proposition 11.33 (Min-SW bounds, exactness and metric status)**.** *For $\alpha,\beta\in\mathcal{P}_2(\mathbb{R}^d)$,*

$$\mathcal{W}_2(\alpha,\beta) \leqslant \text{Min-SW}_2(\alpha,\beta) \leqslant \inf_{z\in\mathbb{R}^d}\left(2\int \|x-z\|^2\mathrm{d}\alpha(x) + 2\int \|y-z\|^2\mathrm{d}\beta(y)\right)^{1/2}. \tag{11.48}$$

If both supports lie in a compact set K, the last term can be replaced by $\operatorname{diam}(K)$. Moreover, $\text{Min-SW}_2(\alpha,\beta) = \mathcal{W}_2(\alpha,\beta)$ if and only if some quadratic optimal plan $\pi^\star\in\mathcal{U}(\alpha,\beta)$ satisfies $(P_\theta,P_\theta)_\sharp\pi^\star = \omega_\theta$ for some θ. Thus, for equal-weight point clouds, equality holds whenever an optimal assignment is induced by sorting both clouds along one direction. In particular, it holds for two uniform n-point measures when $d\geqslant 2n-1$ and the concatenated family $x_1,\dots,x_n,y_1,\dots,y_n$ is in general position [516]. Here general position means that, for each $1\leqslant k\leqslant d$, no subset of $k+2$ points from this family lies in a k-dimensional affine subspace.

The function Min-SW_2 is non-negative, symmetric and separates measures, but it does not satisfy the triangle inequality in general. It is therefore a separating symmetric discrepancy, not a distance.

Proof. Every $\pi\in C_\theta(\alpha,\beta)$ is an admissible coupling, which proves the first inequality in (11.48). For every $z\in\mathbb{R}^d$, the pointwise bound $\|x-y\|^2\leqslant 2\|x-z\|^2+2\|y-z\|^2$ proves the second one after minimizing over z, while $\|x-y\|\leqslant\operatorname{diam}(K)$ gives the compact-support version. Equality with $\mathcal{W}_2$ holds precisely when the constrained set for some direction contains a quadratic optimal plan. Under the stated dimension and general-position hypotheses, every ordering of the concatenated $2n$-point family is induced by some linear projection. One may therefore choose an ordering that realizes an optimal assignment; this is the projection-order theorem proved in [516].

Symmetry follows by transposing the constrained plans, and separation follows from the lower bound by $\mathcal{W}_2$. To see that the triangle inequality can fail, let $\alpha_X,\alpha_Y,\alpha_Z$ be the uniform laws on the rows of

$$X=\begin{pmatrix}-4&-4\\0&3\\1&-2\\4&-4\end{pmatrix}, \qquad Z=\begin{pmatrix}-3&2\\2&2\\0&0\\2&-4\end{pmatrix}, \qquad Y=\frac{X+Z}{2}.$$

A direct enumeration of the 4! assignments shows that the identity assignment is quadratically optimal for both (X,Y) and (Y,Z). It is induced by sorting along directions proportional to $(2,-1)$ and $(3,1)$, respectively. Hence

$$\text{Min-SW}_2(\alpha_X,\alpha_Y)^2 = \text{Min-SW}_2(\alpha_Y,\alpha_Z)^2 = \frac{51}{16}.$$

For (X,Z), the projected orders change only across the ten lines orthogonal to pairwise differences of rows of X or Z. Since the cost is linear in the plan, it suffices to enumerate the resulting angular cells and, on their boundaries, the compatible extreme permutations. This gives the nine distinct squared costs $\frac14\{63,67,73,79,91,95,105,131,135\}$. Thus $\text{Min-SW}_2(\alpha_X,\alpha_Z)^2 = 63/4$, and consequently

$$\text{Min-SW}_2(\alpha_X,\alpha_Z) = \frac{\sqrt{63}}{2} > \frac{\sqrt{51}}{2} = \text{Min-SW}_2(\alpha_X,\alpha_Y) + \text{Min-SW}_2(\alpha_Y,\alpha_Z).$$

□

For comparison, fixing θ before comparing the measures restores a metric on equal-weight n-point clouds whose θ-projections are injective: it is $n^{-1/2}$ times the Euclidean distance between the two tuples ordered along θ. The constrained fixed-direction extension is likewise a metric on the class of measures with atomless θ-projections [516]. It is the subsequent pair-dependent minimization over θ that destroys the triangle inequality. In dimension one no such directional choice remains, and Min-SW$_2$ = $\mathcal{W}_2$.

The right-hand side of (11.48) and the diameter estimate are absolute upper bounds, not multiplicative comparisons with $\mathcal{W}_2$. Beyond the exactness criterion above, the cited theory does not provide a universal converse of the form Min-SW$_2 \leqslant C\,\mathcal{W}_2$.

Figure 11.6 illustrates the resulting gap: the direction selected by Min-SW produces a valid lifted planar coupling, but that coupling need not be the quadratic optimal plan.

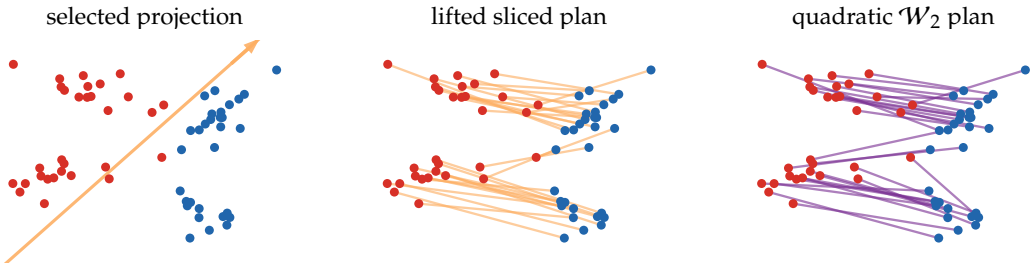


Figure 11.6: Min-SW lifted plan. A deterministic angular sweep selects a projection, after which the red and blue atoms are sorted and matched in one dimension. The middle panel lifts this matching back to the plane. Its cost upper-bounds $\mathcal{W}_2^2$, but the resulting feasible plan need not equal the quadratic optimal plan shown on the right.

11.3 Quotient Wasserstein and Wasserstein–Procrustes

This section explains how OT can factor out nuisance symmetries before comparing measures.

Quotient distances. Many comparison problems contain nuisance transformations: two shapes, images or point clouds should be considered close after translating, rotating or otherwise reparametrizing one of them. Quotient Wasserstein distances encode this idea by computing transport after optimizing over a group action. This construction is the metric analogue of passing from objects to shapes modulo symmetries, and connects OT with shape spaces, metamorphosis models and global-invariance variants of transport [522, 559, 10].

Definition 11.34 (Quotient Wasserstein distance)**.** Let a group $\mathcal{G}$ act on a metric space (X, d) by isometries, and assume that the action preserves finite pth moments. Write $g_\sharp\alpha$ for the push-forward of a measure α by $g \in \mathcal{G}$, and $[\alpha] := \{g_\sharp\alpha \,:\, g \in \mathcal{G}\}$ for its orbit. The quotient p-Wasserstein distance between the orbits of $\alpha, \beta \in \mathcal{P}_p(X)$ is

$$\mathcal{W}_{p,\mathcal{G}}([\alpha],[\beta]) := \inf_{g,h\in\mathcal{G}} \mathcal{W}_p(g_\sharp\alpha, h_\sharp\beta) = \inf_{g\in\mathcal{G}} \mathcal{W}_p(\alpha, g_\sharp\beta).$$

Proposition 11.35 (Metric property on the quotient)**.** *If $\mathcal{G}$ acts by isometries and preserves $\mathcal{P}_p(X)$, then $\mathcal{W}_{p,\mathcal{G}}$ is a pseudo-distance on $\mathcal{P}_p(X)$. If the infimum is attained between every pair of orbits, then it is a distance on the orbit space $\mathcal{P}_p(X)/\mathcal{G}$. This applies in particular to continuous actions of compact groups.*

Proof. Isometry of the action gives $\mathcal{W}_p(g_\sharp\alpha, g_\sharp\beta) = \mathcal{W}_p(\alpha,\beta)$, which proves the second formula in Definition 11.34. Non-negativity and symmetry are inherited from $\mathcal{W}_p$. For the triangle inequality, choose $g_1, g_2 \in \mathcal{G}$. Then

$$\mathcal{W}_p(\alpha,(g_1)_\sharp\beta) + \mathcal{W}_p(\beta,(g_2)_\sharp\gamma) = \mathcal{W}_p(\alpha,(g_1)_\sharp\beta) + \mathcal{W}_p((g_1)_\sharp\beta,(g_1g_2)_\sharp\gamma).$$

Applying the triangle inequality for $\mathcal{W}_p$ and then taking infima over g_1, g_2 gives the result. If the infimum is attained and the quotient distance is zero, there exist $g, h \in \mathcal{G}$ such that $\mathcal{W}_p(g_\sharp\alpha, h_\sharp\beta) = 0$, hence $g_\sharp\alpha = h_\sharp\beta$ and therefore $[\alpha] = [\beta]$. For a continuous action of a compact group, the objective is continuous on a compact parameter set, so attainment follows. Without attainment, the construction is naturally a metric only after identifying distinct orbits at zero orbit distance. □

Rigid motions and Wasserstein–Procrustes. The most common example is the Euclidean group $\mathrm{E}(d) = \mathrm{O}(d) \ltimes \mathbb{R}^d$. Its quotient distance has a standard registration name.

Definition 11.36 (Wasserstein–Procrustes distance)**.** For $\alpha, \beta \in \mathcal{P}_2(\mathbb{R}^d)$, define

$$\mathrm{WP}_2(\alpha,\beta)^2 := \inf_{\substack{R\in\mathrm{O}(d),\, t\in\mathbb{R}^d \\ \pi\in\mathcal{U}(\alpha,\beta)}} \int \|Rx + t - y\|^2 \mathrm{d}\pi(x,y).$$

Restricting R to $\mathrm{SO}(d)$ defines the orientation-preserving variant.

This is the case $p = 2$ of the quotient distance; for a general exponent p, one replaces the quadratic cost by the pth power of the Euclidean distance. Although the translation group is noncompact, the quadratic problem is well behaved: for fixed R and π, the optimal translation aligns $R\bar{x}$ with $\bar{y}$. The remaining rigid step is therefore an orthogonal Procrustes problem over the compact group $\mathrm{O}(d)$. For empirical measures, this couples a transport problem with a rigid registration problem. Classical iterative closest point methods alternate nearest-neighbor assignment and rigid least squares [53]. Wasserstein–Procrustes replaces these hard many-to-one nearest-neighbor correspondences by a mass-preserving OT plan. This makes the registration less tied to sampling density and better suited to ambiguous correspondences.

Two complementary machine-learning formulations clarify the scope of this construction. Grave, Joulin and Berthet [252] formulate unsupervised alignment of high-dimensional embeddings as a Wasserstein–Procrustes problem. In their equal-weight setting, one jointly estimates an orthogonal matrix and a permutation; they use a convex-relaxation initialization and a stochastic large-scale solver, with bilingual lexicon induction as the main application. Alvarez-Melis, Jegelka and Jaakkola [10] place the same idea in a broader framework: the coupling and a latent global transformation are optimized jointly over a flexible invariance class because cross-space costs are otherwise ill-defined. The rigid quadratic model and block updates below are the isometric Procrustes specialization of this global-invariance viewpoint.

This is an extrinsic counterpart of the Gromov–Wasserstein viewpoint developed in Section 13.3. Procrustes alignment searches only over ambient rigid motions, whereas GW compares the metric-measure structures themselves. Proposition 13.18 places the Wasserstein–Procrustes distance between GW and ordinary Wasserstein in (13.14), and Remark 13.22 explains the resulting quotient topologies. Thus a good Procrustes registration forces small GW distortion; the Mémoli profile bound in Proposition 13.25 gives a complementary inexpensive intrinsic certificate.

The empirical problem naturally suggests an alternating minimization. Given a current rigid motion $(R^{(\ell)}, t^{(\ell)})$, first compute the OT coupling between the registered source cloud and the target cloud,

$$\mathrm{P}^{(\ell)} \in \operatorname{argmin}_{\mathrm{P}\in\mathrm{U}(\mathrm{a},\mathrm{b})} \sum_{i,j} \mathrm{P}_{i,j} \|R^{(\ell)} x_i + t^{(\ell)} - y_j\|^2. \tag{11.49}$$

Then freeze this coupling and update the rigid motion by

$$(R^{(\ell+1)}, t^{(\ell+1)}) \in \operatorname{argmin}_{R\in\mathrm{O}(d),\, t\in\mathbb{R}^d} \sum_{i,j} \mathrm{P}^{(\ell)}_{i,j} \|Rx_i + t - y_j\|^2, \tag{11.50}$$

or by the same formula with $R \in \mathrm{SO}(d)$ if orientation should be preserved. The first step is an ordinary discrete OT problem with the current registered cost matrix; the next proposition shows that the second step is an orthogonal Procrustes problem with an explicit singular-value formula.

Proposition 11.37 (Rigid update for a fixed coupling)**.** *Let $\alpha = \sum_i a_i \delta_{x_i}$ and $\beta = \sum_j b_j \delta_{y_j}$, and fix $\mathrm{P} \in \mathrm{U}(\mathrm{a},\mathrm{b})$. Define*

$$\bar{x} = \sum_i a_i x_i, \qquad \bar{y} = \sum_j b_j y_j, \qquad M_{\mathrm{P}} = \sum_{i,j} \mathrm{P}_{i,j} (y_j - \bar{y})(x_i - \bar{x})^{\top}.$$

If $M_{\mathrm{P}} = U\Sigma V^{\top}$ is a singular value decomposition, then the minimizers over the full orthogonal group of

$$\min_{R\in\mathrm{O}(d),\, t\in\mathbb{R}^d} \sum_{i,j} \mathrm{P}_{i,j} \|Rx_i + t - y_j\|^2$$

are given by $R^\star = UV^\top$ *and* $t^\star = \bar{y} - R^\star \bar{x}$*, up to the usual non-uniqueness when* M_{P} *is rank deficient. If one imposes* $R \in \mathrm{SO}(d)$*, set*

$$D = \mathrm{diag}(1, \dots, 1, \det(UV^\top)), \qquad R^\star = UDV^\top, \qquad t^\star = \bar{y} - R^\star \bar{x}.$$

Proof. For fixed R, differentiating with respect to t gives $t = \bar{y} - R\bar{x}$. Substituting this value centers the variables and leaves

$$\sum_{i,j} \mathrm{P}_{i,j} \|R(x_i - \bar{x}) - (y_j - \bar{y})\|^2.$$

The two quadratic terms independent of R are fixed, so the problem is equivalent to maximizing

$$\sum_{i,j} \mathrm{P}_{i,j} \langle R(x_i - \bar{x}),\, y_j - \bar{y} \rangle = \mathrm{tr}(R^\top M_{\mathrm{P}}).$$

Von Neumann's trace inequality gives $\mathrm{tr}(R^\top U \Sigma V^\top) \leqslant \mathrm{tr}(\Sigma)$ for $R \in \mathrm{O}(d)$, with equality for $R = UV^\top$. Under the constraint $R \in \mathrm{SO}(d)$, the same argument applies with the additional determinant constraint on $U^\top R V$, which gives the displayed diagonal correction. □

Equations (11.49)–(11.50) give a block-coordinate method. The objective is not jointly convex, so the scheme should be read as a registration heuristic for the quotient problem rather than as a global solver. The exact block update moves the rigid motion fully; for visualization or continuation, one may damp the displayed motion between two successive poses.

Algorithm 11.2: Alternating Wasserstein–Procrustes alignment

Input: Weighted point clouds $(x_i, \mathrm{a}_i)_{i=1}^n$, $(y_j, \mathrm{b}_j)_{j=1}^m$, initial rigid motion $(R^{(0)}, t^{(0)})$, tolerance tol, iteration budget $L \geqslant 1$, choice $\mathrm{O}(d)$ or $\mathrm{SO}(d)$.
Output: Last coupling P and rigid motion (R, t).
Set $\bar{x} = \sum_i \mathrm{a}_i x_i$, $\bar{y} = \sum_j \mathrm{b}_j y_j$, $\ell = 0$ and $\eta^{(0)} = +\infty$.
While $\eta^{(\ell)} >$ tol and $\ell < L$ **do**:
- **Set** $\mathrm{C}^{(\ell)}_{i,j} = \|R^{(\ell)} x_i + t^{(\ell)} - y_j\|^2$.
- **Solve** $\mathrm{P}^{(\ell)} \in \mathrm{argmin}_{\mathrm{P} \in \mathrm{U}(\mathrm{a},\mathrm{b})} \sum_{i,j} \mathrm{P}_{i,j} \mathrm{C}^{(\ell)}_{i,j}$.
- **Compute** $M_{\mathrm{P}^{(\ell)}} = \sum_{i,j} \mathrm{P}^{(\ell)}_{i,j} (y_j - \bar{y})(x_i - \bar{x})^\top$.
- **Factorize** $M_{\mathrm{P}^{(\ell)}} = U \Sigma V^\top$.
- **Set** $D = \mathrm{Id}$ for $\mathrm{O}(d)$; set $D = \mathrm{diag}(1, \dots, 1, \det(UV^\top))$ for $\mathrm{SO}(d)$.
- **Set** $R^{(\ell+1)} = UDV^\top$ and $t^{(\ell+1)} = \bar{y} - R^{(\ell+1)} \bar{x}$.
- **Set** $\eta^{(\ell+1)} = \|R^{(\ell+1)} - R^{(\ell)}\|_{\mathrm{F}} + \|t^{(\ell+1)} - t^{(\ell)}\|$.
- **Set** $\ell \leftarrow \ell + 1$.

Solve $\mathrm{P}^{(\ell)} \in \mathrm{argmin}_{\mathrm{P} \in \mathrm{U}(\mathrm{a},\mathrm{b})} \sum_{i,j} \mathrm{P}_{i,j} \|R^{(\ell)} x_i + t^{(\ell)} - y_j\|^2$.
Return $\mathrm{P}^{(\ell)}, R^{(\ell)}, t^{(\ell)}$.

Figure 11.7 follows this block-coordinate scheme through a deliberately large translation, showing how the transport correspondences and the rigid registration stabilize together.

Figure 11.7: Wasserstein–Procrustes alignment of two bunny silhouettes under a strong translation and a moderate rotation. The target silhouette is shown in black. The moving source silhouette is sampled by farthest-point sampling and colored from red to blue at iterations $1, 2, 3, 5, 10$. Each step solves an equal-weight OT assignment, then updates the rigid motion by the closed-form Procrustes formula of Proposition 11.37; faint segments show selected correspondences of the current OT assignment. The displayed motion is damped only to make the registration path visible, while the underlying update is the block-coordinate method of Algorithm 11.2.

11.4 Vector Quantiles and Linear Optimal Transport

Linear OT starts from the multivariate analogue of quantile coordinates. The one-dimensional quantile function represents a probability measure by the monotone map sending a fixed reference law to it; in dimension $d > 1$, Brenier's theorem gives the corresponding construction after choosing an absolutely continuous reference probability α_{ref}, typically the uniform law on a convex body or a standard Gaussian.

Vector quantiles. Assume that α_{ref} is absolutely continuous. For a target law α with finite second moment, its vector quantile relative to α_{ref} is the Brenier map

$$T_\alpha = \nabla\varphi_\alpha, \qquad (T_\alpha)_\sharp \alpha_{\mathrm{ref}} = \alpha,$$

or equivalently the solution of

$$\min_{T_\sharp \alpha_{\mathrm{ref}} = \alpha} \int \|x - T(x)\|^2 \mathrm{d}\alpha_{\mathrm{ref}}(x).$$

This construction is canonical only after fixing α_{ref}: changing the reference law changes the coordinates used to represent α. The same transport-based quantile map has been used in several complementary statistical directions. Conditional vector quantile regression replaces scalar conditional quantiles by conditional Brenier maps [105, 106]; Monge–Kantorovich ranks and depth use transports to a spherical reference [134]; center-outward distribution and quantile functions build multivariate ranks and signs from the forward and inverse maps [266]; and scalable nonlinear vector quantile regression learns such conditional maps with flexible models [453].

Linearized Wasserstein coordinates. Linear OT replaces a nonlinear transport distance by a Hilbert norm between reference maps. It is useful when one reference measure is fixed and many nearby distributions must be compared cheaply. Let T_α be the Brenier map pushing α_{ref} to α, understood as an element of $L^2(\alpha_{\mathrm{ref}}; \mathbb{R}^d)$ and hence defined only α_{ref}-almost everywhere. The linear OT embedding is

Definition 11.38 (Linear optimal-transport embedding)**.** Fix an absolutely continuous reference measure $\alpha_{\mathrm{ref}} \in \mathcal{P}_2(\mathbb{R}^d)$, and let T_α denote the Brenier map from α_{ref} to $\alpha \in \mathcal{P}_2(\mathbb{R}^d)$. Then

$$\alpha \mapsto T_\alpha - \mathrm{Id} \in L^2(\alpha_{\mathrm{ref}}; \mathbb{R}^d), \qquad \mathrm{LOT}_{\alpha_{\mathrm{ref}}}(\alpha, \beta) = \|T_\alpha - T_\beta\|_{L^2(\alpha_{\mathrm{ref}})}. \tag{11.51}$$

If one of the two targets equals the reference, the linearized distance is exact. For instance,

$$\mathrm{LOT}_{\alpha_{\mathrm{ref}}}(\alpha_{\mathrm{ref}}, \alpha) = \|T_\alpha - \mathrm{Id}\|_{L^2(\alpha_{\mathrm{ref}})} = \mathcal{W}_2(\alpha_{\mathrm{ref}}, \alpha).$$

For two arbitrary targets, the coupling $(T_\alpha, T_\beta)_\sharp \alpha_{\mathrm{ref}}$ is admissible but not generally optimal, so $\mathrm{LOT}_{\alpha_{\mathrm{ref}}}$ is a tangent-space approximation of the Wasserstein geometry. Introduced for the analysis of image populations [546], LOT has subsequently been used for continuous image-pattern analysis [313], provable classification of transformed distributions [389], collider-event analysis [100], and scalable Wasserstein dimensionality reduction [148]. Uniqueness of the Brenier maps shows that $\mathrm{LOT}_{\alpha_{\mathrm{ref}}}$ is itself a distance on the target measures for which these maps are defined. For a family $(\alpha_s)_s$ with weights $(\lambda_s)_s$, the linearized barycenter is obtained by averaging maps,

$$\bar{T} = \sum_s \lambda_s T_{\alpha_s}, \qquad \bar{\alpha}_{\mathrm{LOT}} = \bar{T}_\sharp \alpha_{\mathrm{ref}}.$$

This is exact in one dimension, where quantile functions linearize $\mathcal{W}_2$, and it is especially useful when many barycenters with changing weights must be evaluated quickly.

Figure 11.8 makes the LOT embedding explicit: map averaging is exact in one-dimensional quantile coordinates, whereas in two dimensions its linearized barycenter can differ from the genuine McCann midpoint.

The usefulness of these coordinates depends on controlling how much they distort the underlying Wasserstein geometry. The following global estimate is therefore important: LOT always dominates $\mathcal{W}_2$, while on compact supports with a uniform reference it remains Hölder-continuous with respect to perturbations measured by $\mathcal{W}_1$.

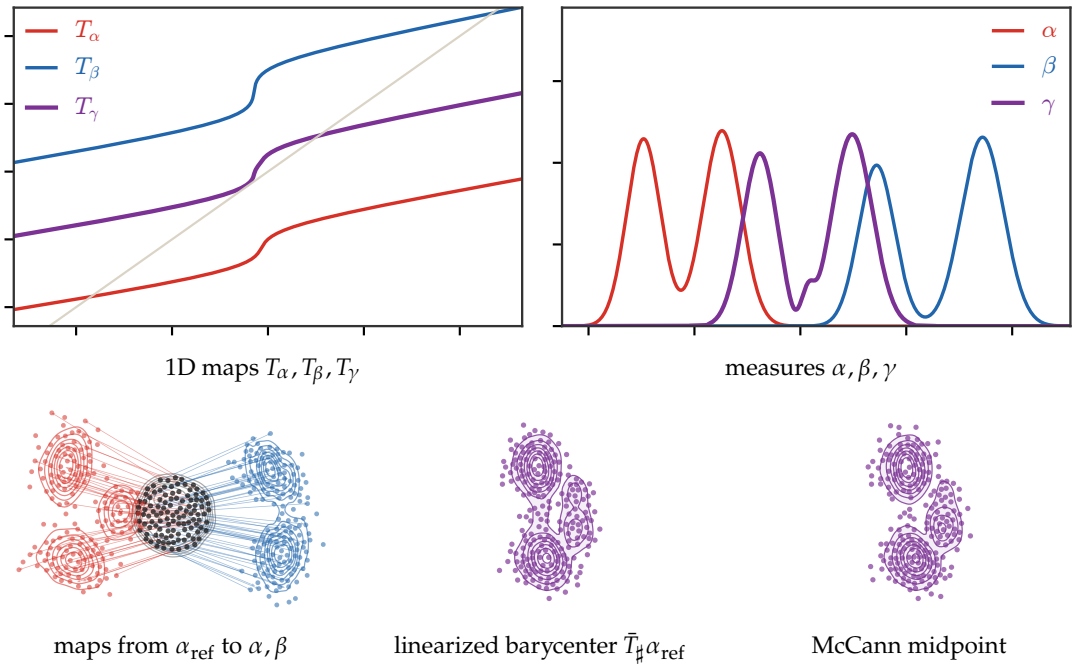


Figure 11.8: Linear OT coordinates. Fixing a reference measure α_{ref} turns each target into a map T_α from α_{ref} to α, or equivalently into the displacement field $T_\alpha - \mathrm{Id}$. In one dimension this is exactly the quantile parametrization of $\mathcal{W}_2$, so averaging the maps toward a two-component α and a two-component β gives the true Wasserstein barycenter. In two dimensions, the first panel shows the reference-to-target maps, computed on dense clouds and displayed on farthest-point subsets, with β represented by two Gaussian components. The middle purple panel shows the linearized barycenter obtained by averaging the two maps from α_{ref}. The right purple panel shows the genuine McCann midpoint between α and β, obtained by solving a direct OT problem between the two target clouds and interpolating at $t = 1/2$.

Proposition 11.39 (Quantitative stability of linear OT)**.** *Let $X, Y \subset \mathbb{R}^d$ be compact convex sets, assume that X has positive Lebesgue measure, and let α_{ref} be normalized Lebesgue measure on X. Then there exists a constant $C = C(d, X, Y)$ such that, for all $\alpha, \beta \in \mathcal{P}(Y)$,*

$$\mathcal{W}_2(\alpha, \beta) \leqslant \mathrm{LOT}_{\alpha_{\text{ref}}}(\alpha, \beta) \quad \textit{and} \quad \mathrm{LOT}_{\alpha_{\text{ref}}}(\alpha, \beta) \leqslant C\, \mathcal{W}_1(\alpha, \beta)^{2/15}.$$

Proof. The first inequality is immediate: $(T_\alpha, T_\beta)_\sharp \alpha_{\text{ref}}$ is a feasible coupling between α and β. The second inequality is the quantitative stability theorem of Mérigot, Delalande and Chazal [382]. Its proof first controls the difference of the two Kantorovich potentials in terms of $\mathcal{W}_1(\alpha, \beta)$, then combines convexity and interpolation estimates to control the gradients in $L^2(\alpha_{\text{ref}})$; the resulting dimension-independent Hölder exponent is $2/15$. In one dimension, quantile functions make the embedding isometric and the exponent improves to one. In several dimensions, the theorem is global but only Hölder; it should not be read as a global Lipschitz estimate in $\mathcal{W}_2$. □

> **Remark 11.40 (Three Hilbertian embeddings of measures).** Several constructions in this text embed measures into Hilbert spaces, but they encode different geometries. Kernel mean embeddings send α to $\int k(x, \cdot) \mathrm{d}\alpha(x)$ in an RKHS and lead to MMD distances; see Section 7.2. Quadratic sliced Wasserstein sends a measure to the collection of one-dimensional quantile functions of its projections, viewed in $L^2(\mathbb{S}^{d-1} \times [0, 1])$; see Section 11.2. Linear OT sends α to the displacement field $T_\alpha - \mathrm{Id}$ from a fixed reference α_{ref} in $L^2(\alpha_{\text{ref}}; \mathbb{R}^d)$. The first construction is linear in the measure and depends on the kernel, the second is nonlinear but reduces OT to projected one-dimensional quantiles, and the third is a tangent approximation to the full Wasserstein geometry around a chosen reference.

Principal components in linear OT coordinates. The preceding embedding turns probability measures into displacement fields in a fixed Hilbert chart, so ordinary principal component analysis can be applied to deformations rather than to densities. Given training measures $(\alpha_i)_{i=1}^N$, set

$$z_i := T_{\alpha_i} - \mathrm{Id}, \qquad \bar{z} := \frac{1}{N} \sum_{i=1}^N z_i.$$

The map $\mathrm{Id} + \bar{z}$ defines the linear OT mean $(\mathrm{Id} + \bar{z})_\sharp \alpha_{\mathrm{ref}}$. The empirical covariance operator on $L^2(\alpha_{\mathrm{ref}}; \mathbb{R}^d)$ is the finite-rank operator

$$C_{\mathrm{LOT}} h := \frac{1}{N} \sum_{i=1}^{N} \langle z_i - \bar{z},\, h \rangle_{L^2(\alpha_{\mathrm{ref}})} (z_i - \bar{z}),$$

and its leading orthonormal eigenvectors e_k define the principal linear OT modes. Equivalently, one diagonalizes the $N \times N$ Gram matrix $G_{i,j} = \frac{1}{N} \langle z_i - \bar{z},\, z_j - \bar{z} \rangle_{L^2(\alpha_{\mathrm{ref}})}$. If $G v_k = \lambda_k v_k$ with $\lambda_k > 0$ and $\|v_k\| = 1$, then

$$e_k = \frac{1}{\sqrt{N \lambda_k}} \sum_{i=1}^{N} (v_k)_i (z_i - \bar{z})$$

is the corresponding unit eigenvector of C_{LOT}. The score of α_i along mode e_k is

$$a_{i,k} := \langle z_i - \bar{z},\, e_k \rangle_{L^2(\alpha_{\mathrm{ref}})}.$$

A low-rank reconstruction, or a synthetic excursion with prescribed coefficients $a = (a_1, \ldots, a_m)$, is then obtained by pushing the reference through

$$T_a(x) := x + \bar{z}(x) + \sum_{k=1}^{m} a_k e_k(x), \qquad \alpha_a := (T_a)_\sharp \alpha_{\mathrm{ref}}.$$

For small excursions around the mean displacement, this gives a practical tangent-space PCA for probability measures: it captures dominant modes of deformation while avoiding repeated pairwise OT computations. For large coefficients, T_a may fail to be the Brenier map from α_{ref} to α_a, and may even leave the regular chart where T_a is a gradient of a convex function. Thus $a \mapsto \alpha_a$ should be read as a chart-dependent linearized visualization rather than as an intrinsic Wasserstein geodesic. This LOT-PCA viewpoint was introduced for image variability analysis in [546] and developed in transport-based signal analysis [518, 312]; it complements intrinsic principal-geodesic or geodesic-PCA approaches, which optimize directly in the curved Wasserstein space [483, 57].

Figure 11.9 first shows the exact one-dimensional case. A large training set of well-separated two-Gaussian mixtures is generated with moderate component-wise location and scale variations, although only representative densities are displayed. Each density is embedded by its quantile function on a fine grid, ordinary PCA is performed in $L^2(0,1)$, and the displayed modes are obtained by converting quantile excursions back into densities.

Figure 11.10 then illustrates a regularized numerical approximation of the same construction on MNIST digit-zero images. Each image is normalized as a probability measure on the pixel grid. A Sinkhorn barycenter is used as the reference α_{ref}, entropic couplings from α_{ref} to the training images are converted into barycentric-projection maps, and PCA is applied to the resulting approximate displacement fields.

11.5 Spectral and Robust Wasserstein Distances

Spectral OT changes the scalar quadratic cost by measuring the whole displacement covariance through a matrix gauge. The same object admits a robust projected formulation: instead of fixing one projection, one maximizes over the polar set of the gauge. Subspace robust OT is the important non-convex rank-constrained version of this idea [419]; spectral gauges provide its convex minimax counterpart and connect to recent spectral-gradient viewpoints such as Muon dynamics [427].

Definition 11.41 (Monotone spectral gauge)**.** A monotone spectral gauge on positive semidefinite matrices is a convex, positively 1-homogeneous map $\gamma : \mathbb{S}^d_+ \to \mathbb{R}_+$ such that $\gamma(M) = 0$ only for $M = 0$, $\gamma(QMQ^\top) = \gamma(M)$ for every orthogonal matrix Q, and

$$0 \preceq M \preceq N \quad \Longrightarrow \quad \gamma(M) \leqslant \gamma(N).$$

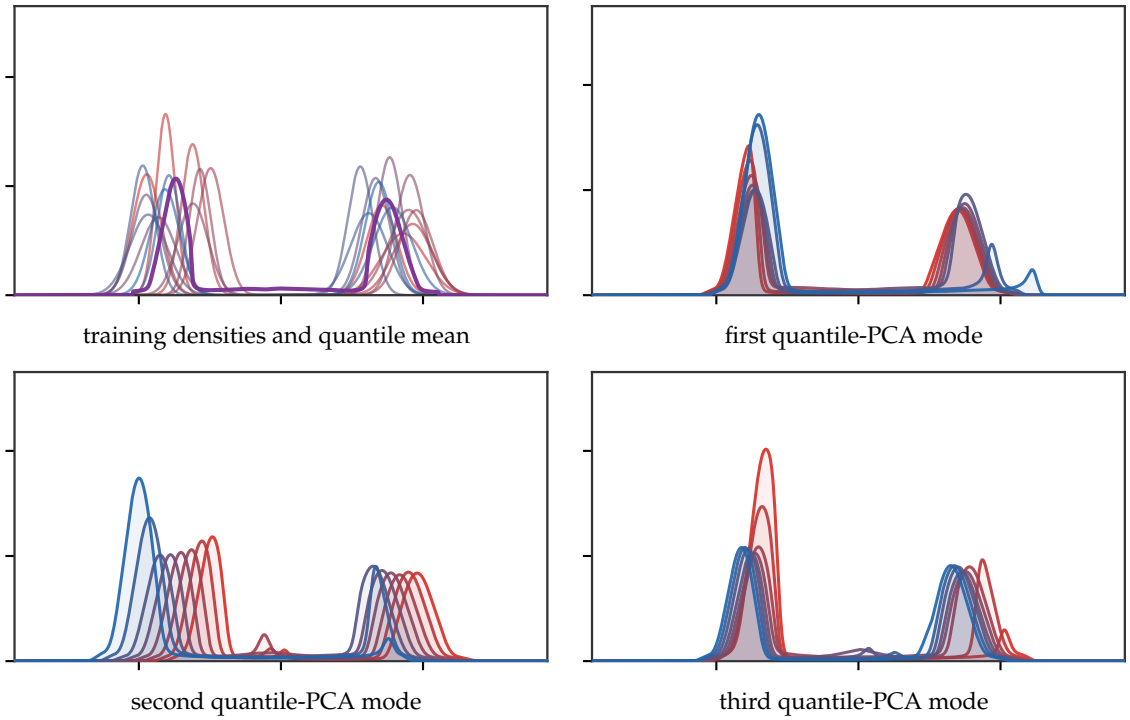


Figure 11.9: One-dimensional linear OT PCA for synthetic two-Gaussian mixtures. The PCA is fit on a large training ensemble, while the dataset panel displays only representative densities and the quantile average in violet. Each mode panel shows densities obtained from $Q_{\bar{\alpha}} + ae_k$, using slightly extrapolated coefficients a increasing from red to blue. Since the embedding is the exact quantile parametrization $Q_\alpha \in L^2(0,1)$, this PCA is performed in exact Wasserstein coordinates rather than in an approximate tangent chart.

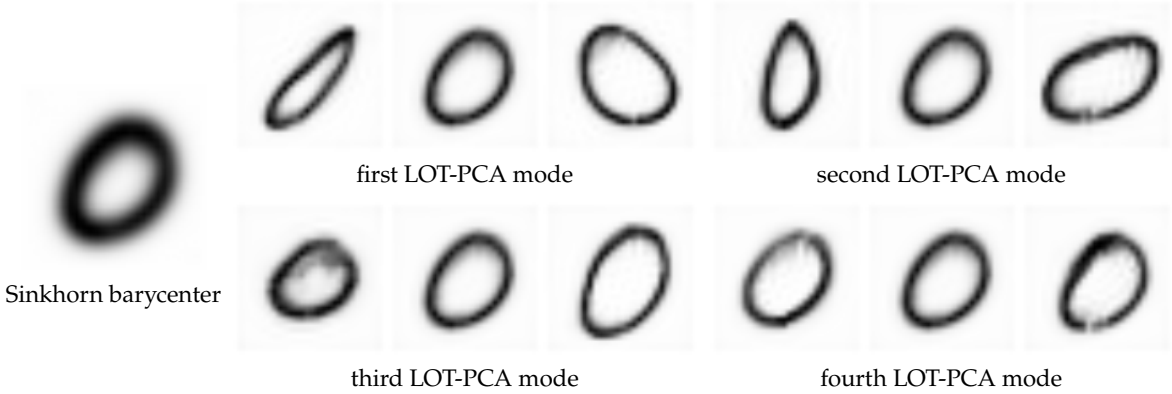


Figure 11.10: Principal components in linear OT coordinates for MNIST digit-zero histograms. The left panel shows the Sinkhorn barycenter reference α_{ref}. Each mode panel is a triptych obtained by pushing α_{ref} through $x \mapsto x + \bar{z}(x) + ae_k(x)$ with a negative, zero, and positive coefficient a. The displayed densities use white for zero displayed mass and black for high displayed mass; the underlying transport computations still use the original densities, where high pixel values carry the mass. The modes capture rotations, aspect-ratio changes, and stroke-thickness deformations in the tangent chart around the barycenter.

For $1 \leqslant q \leqslant +\infty$, the Schatten gauge

$$\gamma_q(M) := \|M\|_{S_q} = \begin{cases} \left(\sum_{i=1}^d \lambda_i(M)^q\right)^{1/q}, & 1 \leqslant q < +\infty, \\ \lambda_{\max}(M), & q = +\infty, \end{cases}$$

is a monotone spectral gauge. The cases $q = 1$, $q = 2$ and $q = +\infty$ are respectively the trace, Frobenius and spectral gauges.

The monotonicity condition means that increasing the displacement covariance in Loewner order cannot decrease the transport penalty.

Definition 11.42 (Spectral Wasserstein distance)**.** Let γ be a monotone spectral gauge. For a coupling $\pi \in \mathcal{U}(\alpha, \beta)$, define its displacement covariance

$$M_\pi := \int (x - y)(x - y)^\top \mathrm{d}\pi(x, y).$$

The spectral Wasserstein distance associated with γ is

$$\mathcal{W}_\gamma(\alpha, \beta)^2 := \inf_{\pi \in \mathcal{U}(\alpha,\beta)} \gamma(M_\pi). \tag{11.52}$$

The special case $\gamma(M) = \mathrm{tr}(M)$ gives the usual quadratic Wasserstein distance $\mathcal{W}_2$. The spectral gauge $\gamma(M) = \lambda_{\max}(M)$ instead measures the worst transported variance direction. For $A \succeq 0$, define the quadratic projected transport cost

$$\mathcal{W}_{2,A}(\alpha, \beta)^2 := \inf_{\pi \in \mathcal{U}(\alpha,\beta)} \int (x - y)^\top A(x - y) \mathrm{d}\pi(x, y) = \mathcal{W}_2((A^{1/2})_\sharp \alpha, (A^{1/2})_\sharp \beta)^2. \tag{11.53}$$

The last equality remains valid when A is singular. One inequality follows by projecting any coupling of (α, β); for the converse, disintegrate α and β over their $A^{1/2}$-images and lift a coupling of the projected measures by conditionally coupling the fibers. The polar set of the gauge is

$$\mathcal{B}_\gamma := \{A \succeq 0 \,;\, \mathrm{tr}(AM) \leqslant \gamma(M) \quad \text{for all } M \succeq 0\}, \tag{11.54}$$

so that, for a closed gauge, $\gamma(M) = \sup_{A \in \mathcal{B}_\gamma} \mathrm{tr}(AM)$. For the Schatten gauge γ_q, Schatten Hölder duality gives

$$\mathcal{B}_{\gamma_q} = \left\{A \succeq 0 \,;\, \|A\|_{S_{q^*}} \leqslant 1\right\}, \qquad \frac{1}{q} + \frac{1}{q^*} = 1,$$

with the usual endpoint conventions. Thus the trace gauge has polar set $\{A : 0 \preceq A \preceq \mathrm{Id}\}$, the Frobenius gauge is self-polar on $\mathbb{S}^d_+$, and the spectral gauge has polar set $\{A \succeq 0 : \mathrm{tr}(A) \leqslant 1\}$.

Proposition 11.43 (Robust representation and metric equivalence)**.** *Assume, for simplicity, that the measures are compactly supported and that γ is closed and finite on the positive semidefinite cone. Then*

$$\mathcal{W}_\gamma(\alpha, \beta)^2 = \sup_{A \in \mathcal{B}_\gamma} \mathcal{W}_{2,A}(\alpha, \beta)^2.$$

If there exist constants $0 < a \leqslant b < +\infty$ such that $a\mathrm{Id} \in \mathcal{B}_\gamma$ and $\mathcal{B}_\gamma \subset \{A \,;\, 0 \preceq A \preceq b\mathrm{Id}\}$, equivalently

$$a\,\mathrm{tr}(M) \leqslant \gamma(M) \leqslant b\,\mathrm{tr}(M) \qquad (M \succeq 0),$$

then

$$\sqrt{a}\,\mathcal{W}_2(\alpha, \beta) \leqslant \mathcal{W}_\gamma(\alpha, \beta) \leqslant \sqrt{b}\,\mathcal{W}_2(\alpha, \beta).$$

The robust representation proves that $\mathcal{W}_\gamma$ is a distance: the supremum of pseudodistances gives symmetry and the triangle inequality, while the lower comparison with $\mathcal{W}_2$ gives definiteness. When γ is the restriction of a norm to the positive semidefinite cone, these bounds hold automatically in finite dimension, so $\mathcal{W}_\gamma$ is equivalent to $\mathcal{W}_2$ on measures with finite second moments.

Proof. Using the polar representation of γ,

$$\mathcal{W}_\gamma(\alpha,\beta)^2 = \inf_{\pi\in\mathcal{U}(\alpha,\beta)} \sup_{A\in\mathcal{B}_\gamma} \operatorname{tr}(AM_\pi).$$

The coupling set is convex and compact for weak convergence under compact support. Since γ is a finite gauge on a finite-dimensional cone and vanishes only at the origin, it is equivalent to the trace norm on the slice $\operatorname{tr}(M) = 1$, so $\mathcal{B}_\gamma$ is convex and compact. The map $(\pi, A) \mapsto \operatorname{tr}(AM_\pi)$ is affine in each variable and continuous. Sion's minimax theorem gives

$$\inf_\pi \sup_{A\in\mathcal{B}_\gamma} \operatorname{tr}(AM_\pi) = \sup_{A\in\mathcal{B}_\gamma} \inf_\pi \operatorname{tr}(AM_\pi) = \sup_{A\in\mathcal{B}_\gamma} \mathcal{W}_{2,A}(\alpha,\beta)^2.$$

For fixed $A \succeq 0$, $\mathcal{W}_{2,A}$ is the Wasserstein pseudodistance associated with the seminorm $x \mapsto \|A^{1/2}x\|$. Since all terms are nonnegative, the robust identity also gives

$$\mathcal{W}_\gamma(\alpha,\beta) = \sup_{A\in\mathcal{B}_\gamma} \mathcal{W}_{2,A}(\alpha,\beta).$$

A supremum of pseudodistances is symmetric and satisfies the triangle inequality.

If $a\mathrm{Id} \in \mathcal{B}_\gamma$ and $A \preceq b\mathrm{Id}$ for all $A \in \mathcal{B}_\gamma$, then

$$a\,\mathcal{W}_2(\alpha,\beta)^2 = \mathcal{W}_{2,a\mathrm{Id}}(\alpha,\beta)^2 \leqslant \mathcal{W}_\gamma(\alpha,\beta)^2 \leqslant b\,\mathcal{W}_2(\alpha,\beta)^2,$$

which proves definiteness and equivalence with $\mathcal{W}_2$. The equivalence between these operator bounds and $a\operatorname{tr}(M) \leqslant \gamma(M) \leqslant b\operatorname{tr}(M)$ follows directly from the polar formula. In finite dimension, any norm restricted to the positive semidefinite cone is equivalent to the trace norm on that cone. The finite-second-moment case follows by truncation when these norm-equivalence bounds hold. □

Definition 11.44 (Subspace robust Wasserstein)**.** For $1 \leqslant k \leqslant d$, the Paty–Cuturi subspace robust Wasserstein distance is

$$\mathrm{SRW}_{2,k}(\alpha,\beta) := \sup_{U^\top U=\mathrm{Id}_k} \mathcal{W}_2((U^\top)_\sharp\alpha,(U^\top)_\sharp\beta) = \sup_{P^2=P=P^\top,\ \operatorname{tr}(P)=k} \mathcal{W}_{2,P}(\alpha,\beta).$$

For the Ky Fan gauge

$$\gamma_k(M) = \sum_{\ell=1}^{k} \lambda_\ell(M),$$

where the eigenvalues are sorted in decreasing order, the polar set is

$$\mathcal{B}_{\gamma_k} = \{A\ ;\ 0 \preceq A \preceq \mathrm{Id},\ \operatorname{tr}(A) \leqslant k\}\,.$$

Thus $k = d$ gives $\gamma_d(M) = \operatorname{tr}(M)$ and recovers $\mathcal{W}_2$. The convex hull of rank-k projectors is

$$\{A\ ;\ 0 \preceq A \preceq \mathrm{Id},\ \operatorname{tr}(A) = k\}\,,$$

and, since $M \succeq 0$, the associated support function is the same Ky Fan gauge. Thus $\mathcal{W}_{\gamma_k}$ is the convexified spectral counterpart of $\mathrm{SRW}_{2,k}$, while $\mathrm{SRW}_{2,k}$ keeps the original non-convex rank constraint. More precisely,

$$\mathrm{SRW}_{2,k}(\alpha,\beta) \leqslant \mathcal{W}_{\gamma_k}(\alpha,\beta) \leqslant \mathcal{W}_2(\alpha,\beta), \qquad \sqrt{\frac{k}{d}}\,\mathcal{W}_2(\alpha,\beta) \leqslant \mathcal{W}_{\gamma_k}(\alpha,\beta),$$

because

$$\left\{\frac{k}{d}\mathrm{Id}\right\} \cup \left\{P\ ;\ P^2 = P = P^\top,\ \operatorname{tr}(P) = k\right\} \subset \mathcal{B}_{\gamma_k} \subset \{A\ ;\ 0 \preceq A \preceq \mathrm{Id}\}\,.$$

For $k = 1$, $\gamma_1(M) = \lambda_{\max}(M)$ and $\mathcal{B}_{\gamma_1} = \{A \succeq 0\ ;\ \operatorname{tr}(A) \leqslant 1\}$. This top-eigenvalue spectral Wasserstein geometry is the case connected to Muon-type spectral dynamics in [427].

Figure 11.11 compares the trace and top-eigenvalue geometries at both levels: the selected transport plans and the displacement interpolations they induce.

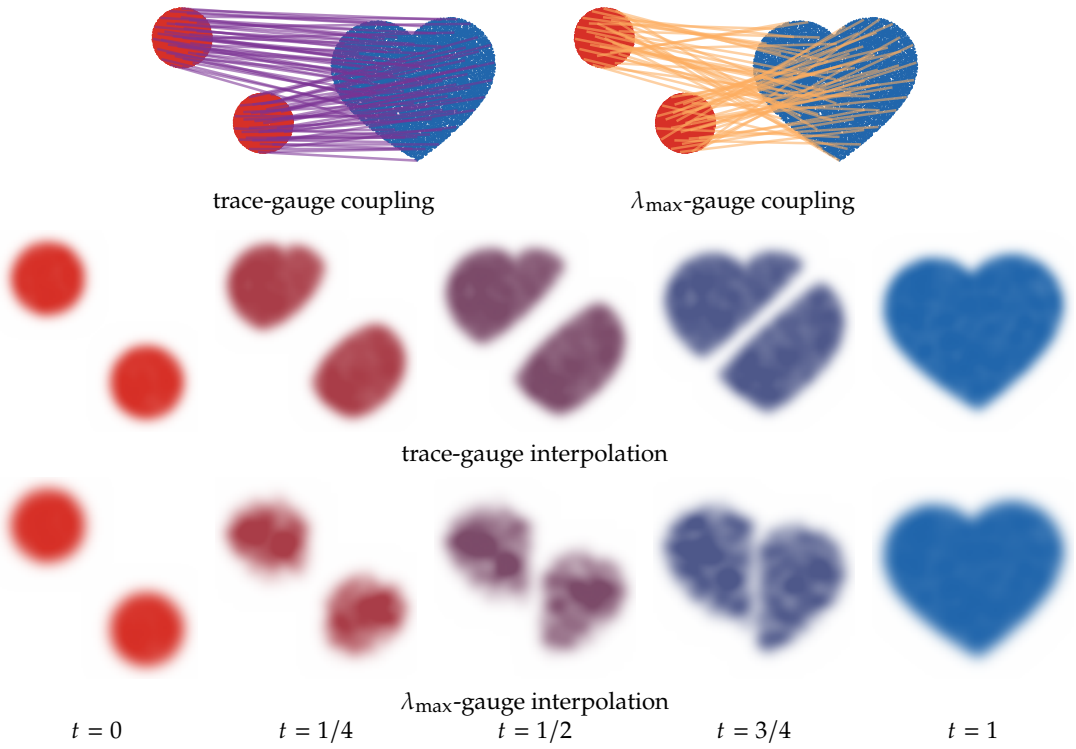


Figure 11.11: Trace and spectral gauges for displacement covariances. The trace gauge minimizes the average squared displacement and gives the usual quadratic transport plan from a shifted two-disk source silhouette to a shifted heart-shaped target. The $\lambda_{\max}$ gauge penalizes the worst projected displacement variance; the displayed plan is obtained by approximating the robust formulation with finitely many directions. The last two rows render the corresponding displacement interpolations from 70,000-point silhouette samples and kernel-smoothed lifted plans, with the same convention as Figure 2.6: white means zero density, and high density saturates in the red-to-blue interpolation color of the corresponding time. The spatial separation makes the different displacement geometries easier to read.

11.6 Conditional Wasserstein Distances

Many applications compare probability laws while keeping an external condition fixed: a class label, a time variable, a spatial location, or, later in Section 15.8, the depth of a residual network. The resulting geometry is a fiberwise, or conditional, version of optimal transport. It is based on disintegration of measures and is closely related to conditional and constrained variants of transport used in weak transport and conditional simulation [539, 469, 26, 408, 29].

The recent literature uses this same fiberwise constraint in several complementary directions. Peszek and Poyato study heterogeneous gradient flows in the topology of fibered optimal transport, emphasizing fixed-fiber transport and PDEs with heterogeneities [421]. Hosseini, Hsu and Taghvaei develop conditional optimal transport on function spaces through triangular maps and Kantorovich relaxations, motivated by amortized Bayesian inference [279]. Chemseddine, Hagemann, Steidl and Wald introduce conditional Wasserstein distances for Bayesian inverse problems and OT flow matching, with restricted couplings that compare posterior laws condition by condition [125]. Kerrigan, Migliorini and Smyth give a dynamic conditional OT formulation and use it to build simulation-free conditional flows [300]. The definition below isolates the common geometric core: transport is ordinary within each fiber and forbidden across distinct conditions.

Definition 11.45 (Conditional couplings and conditional OT)**.** Let (S, λ) be a standard Borel probability space of conditions and let (Ω, d) be a Polish metric space. For $p \geqslant 1$, denote by $\mathcal{P}_{p,\lambda}(S \times \Omega)$ the set of probability measures on $S \times \Omega$ whose first marginal is λ and whose disintegration

$$\alpha(\mathrm{d}s, \mathrm{d}x) = \alpha_s(\mathrm{d}x)\lambda(\mathrm{d}s)$$

satisfies $\int_S \int_\Omega d(x, x_0)^p \, \mathrm{d}\alpha_s(x)\mathrm{d}\lambda(s) < +\infty$ for some, hence every, $x_0 \in \Omega$.

For $\alpha, \beta \in \mathcal{P}_{p,\lambda}(S \times \Omega)$, a conditional coupling is a measure

$$\Pi(\mathrm{d}s, \mathrm{d}x, \mathrm{d}y) = \pi_s(\mathrm{d}x, \mathrm{d}y)\lambda(\mathrm{d}s)$$

such that $\pi_s \in \mathcal{U}(\alpha_s, \beta_s)$ for λ-a.e. s. The set of such couplings is denoted $\mathcal{U}_\lambda(\alpha, \beta)$. Let $(s, x, y) \mapsto c_s(x, y)$ be jointly Borel, nonnegative, and lower semicontinuous in (x, y) for every s. The conditional OT value is

$$\begin{aligned} \mathcal{L}_c^\lambda(\alpha, \beta) &:= \inf_{\Pi \in \mathcal{U}_\lambda(\alpha,\beta)} \int_S \int_{\Omega\times\Omega} c_s(x, y)\mathrm{d}\pi_s(x, y)\mathrm{d}\lambda(s) \\ &= \int_S \inf_{\pi_s \in \mathcal{U}(\alpha_s,\beta_s)} \int_{\Omega\times\Omega} c_s(x, y)\mathrm{d}\pi_s(x, y)\mathrm{d}\lambda(s). \end{aligned} \tag{11.55}$$

The equality in (11.55) is not merely a formal interchange of an infimum and an integral. The joint measurability and lower-semicontinuity assumptions make the fiberwise value measurable and permit a measurable selection of optimal plans; measurable near-minimizers give the same equality when an optimizer is unavailable. Equivalently, conditional transport is ordinary transport on $S \times \Omega$ with an infinite cost for moving mass between different values of s.

Definition 11.46 (Conditional Wasserstein distance)**.** For the fiber-independent cost $c_s(x, y) = c_p(x, y) := d(x, y)^p$, define

$$\mathcal{W}_{p,\lambda}(\alpha, \beta) := \left(\mathcal{L}_{c_p}^\lambda(\alpha, \beta)\right)^{1/p} = \left(\int_S \mathcal{W}_p^p(\alpha_s, \beta_s)\mathrm{d}\lambda(s)\right)^{1/p}. \tag{11.56}$$

Thus $\mathcal{L}_c^\lambda$ is the general conditional Kantorovich value, whereas $\mathcal{W}_{p,\lambda}$ is its metric specialization to the constant family $c_s = c_p = d^p$, after taking the pth root. Equivalently,

$$\mathcal{W}_{p,\lambda}(\alpha, \beta)^p = \mathcal{L}_{c_p}^\lambda(\alpha, \beta);$$

the two definitions use exactly the same conditional coupling problem.

Proposition 11.47 (Metric property)**.** *For $p \geqslant 1$, $\mathcal{W}_{p,\lambda}$ is a distance on $\mathcal{P}_{p,\lambda}(S \times \Omega)$. Moreover, this metric space is Polish.*

Proof. Non-negativity and symmetry follow from the corresponding properties of $\mathcal{W}_p$ on each fiber. If $\mathcal{W}_{p,\lambda}(\alpha, \beta) = 0$, then $\mathcal{W}_p(\alpha_s, \beta_s) = 0$ for λ-a.e. s, hence $\alpha_s = \beta_s$ for λ-a.e. s and the disintegrated measures α and β coincide. For the triangle inequality, let $\gamma \in \mathcal{P}_{p,\lambda}(S \times \Omega)$. Since $\mathcal{W}_p(\alpha_s, \gamma_s) \leqslant \mathcal{W}_p(\alpha_s, \beta_s) + \mathcal{W}_p(\beta_s, \gamma_s)$ for a.e. s, Minkowski's inequality in $L^p(S, \lambda)$ gives

$$\mathcal{W}_{p,\lambda}(\alpha, \gamma) \leqslant \mathcal{W}_{p,\lambda}(\alpha, \beta) + \mathcal{W}_{p,\lambda}(\beta, \gamma).$$

To see completeness and separability, identify a measure with the λ-a.e. equivalence class of its measurable disintegration map $s \mapsto \alpha_s \in \mathcal{P}_p(\Omega)$. Formula (11.56) is the corresponding $L^p(S, \lambda; \mathcal{P}_p(\Omega))$ metric. Since $(\mathcal{P}_p(\Omega), \mathcal{W}_p)$ is Polish, the standard L^p space of measurable metric-valued maps is Polish as well. □

Proposition 11.48 (Conditional geodesics)**.** *Assume that (Ω, d) is a geodesic Polish space and let $\alpha_0, \alpha_1 \in \mathcal{P}_{p,\lambda}(S \times \Omega)$. For λ-a.e. s, choose a constant-speed $\mathcal{W}_p$ geodesic $t \mapsto \alpha_{t,s}$ between $\alpha_{0,s}$ and $\alpha_{1,s}$, measurably in s, and set*

$$\alpha_t(\mathrm{d}s, \mathrm{d}x) = \alpha_{t,s}(\mathrm{d}x)\lambda(\mathrm{d}s).$$

Then $t \mapsto \alpha_t$ is a constant-speed geodesic for $\mathcal{W}_{p,\lambda}$.

Proof. For $0 \leqslant r \leqslant t \leqslant 1$, the fiberwise geodesic property gives

$$\mathcal{W}_p(\alpha_{r,s}, \alpha_{t,s}) = (t - r)\,\mathcal{W}_p(\alpha_{0,s}, \alpha_{1,s}) \quad \text{for } \lambda\text{-a.e. } s.$$

Integrating the pth power over S gives

$$\mathcal{W}_{p,\lambda}(\alpha_r, \alpha_t) = (t - r)\,\mathcal{W}_{p,\lambda}(\alpha_0, \alpha_1).$$

Hence the conditional curve has constant speed and realizes the distance between its endpoints. In Euclidean space one may take, for each s, an optimal plan $\pi_s \in \mathcal{U}(\alpha_{0,s}, \alpha_{1,s})$ and set $\alpha_{t,s} = ((1-t)\mathrm{p}_1 + t\mathrm{p}_2)_\sharp \pi_s$, where $\mathrm{p}_1, \mathrm{p}_2$ are the two coordinate projections. On a general geodesic space, the same construction uses a measurable family of optimal dynamical plans. Such a family can be selected measurably in the Polish setting; when geodesics are non-unique, different selections can produce different conditional geodesics. □

CHAPTER 12

Generalized OT Problems

This chapter changes the optimization problem rather than only the ground distance. Barycenters average several measures, multi-marginal OT couples many measures at once, low-rank and capacity constraints restrict the admissible plans, inverse OT learns the cost from observed transport, and weak or martingale OT acts on conditional laws. These models remain close to Kantorovich optimization, but the unknown can now be a family of couplings, a factored plan, a learned cost, or a coupling subject to nonlinear conditional constraints.

12.1 OT Barycenters

Barycenters ask how to average probability measures rather than points. This section explains the variational definition, the special closed forms in one dimension and for Gaussians, and the entropic algorithms used in practice.

Fréchet means. The natural formulation is a Fréchet-mean problem on the space of probability measures: the unknown is the barycenter measure itself, and its support is not prescribed. It uses the continuous Kantorovich value $\mathcal{L}_c$ defined in (3.7).

Definition 12.1 (Optimal-transport barycenter)**.** Given input measures $(\beta_s)_{s=1}^S$ on a space $\mathcal{X}$ and weights $\lambda \in \Sigma_S$, discard any zero-weight inputs. An *optimal-transport barycenter* of this weighted family is any solution of

$$\min_{\alpha \in \mathcal{M}_+^1(\mathcal{X})} \sum_{s=1}^{S} \lambda_s \mathcal{L}_c(\alpha, \beta_s). \tag{12.1}$$

Unlike a coupling, the barycenter is a new probability measure on $\mathcal{X}$. Since the weights λ_s are nonnegative, problem (12.1) is convex in α: Proposition 3.32 shows that the continuous Kantorovich value is jointly convex in its two marginals.

This barycenter problem was introduced by Agueh and Carlier [1], following earlier ideas of Carlier and Ekeland [110]. For the quadratic cost on $\mathcal{X} = \mathbb{R}^d$, a barycenter exists under the usual finite-second-moment assumption. It is unique if at least one positive-weight input is absolutely continuous; more general uniqueness criteria replace absolute continuity by conditions ensuring an essentially unique multi-marginal barycentric map. Discrete existence, consistency and fixed-point constructions are further studied in [14, 9, 329].

Fixed-support restriction. For computation, one often turns the preceding infinite-dimensional problem into a finite one by prescribing possible barycenter locations. Assume the inputs are discrete,

$$\beta_s = \sum_{j=1}^{n_s} b_{s,j} \delta_{x_{s,j}}, \qquad \mathrm{b}_s = (b_{s,j})_{j=1}^{n_s} \in \Sigma_{n_s}.$$

Choose candidate barycenter sites $(y_i)_{i=1}^n$ and restrict the unknown to $\alpha = \sum_i a_i \delta_{y_i}$. For each input s, the cost $\mathcal{L}_c(\alpha, \beta_s)$ then becomes a finite Kantorovich problem with cost matrix

$$(\mathrm{C}_s)_{i,j} = c(y_i, x_{s,j}) \in \mathbb{R}^{n \times n_s}.$$

Thus its value is $\mathrm{L}_{\mathrm{C}_s}(\mathrm{a}, \mathrm{b}_s)$, in the notation of the discrete Kantorovich problem (3.2).

Definition 12.2 (Fixed-support discrete OT barycenter)**.** With the candidate sites and cost matrices above, a *fixed-support discrete OT barycenter* is a measure $\alpha^\star = \sum_i a_i^\star \delta_{y_i}$ whose weight vector $\mathrm{a}^\star = (a_i^\star)_{i=1}^n$ solves

$$\min_{\mathrm{a}\in\Sigma_n} \sum_{s=1}^{S} \lambda_s \mathrm{L}_{\mathrm{C}_s}(\mathrm{a}, \mathrm{b}_s), \tag{12.2}$$

This construction is a finite-dimensional restriction, not an exact discrete reduction of the general barycenter problem. In the ordinary two-marginal Kantorovich problem, once both marginals are discrete, the two supports are known and the whole problem is exactly the matrix optimization (3.2) on their product support. For barycenters, the input supports do not determine the support of the unknown barycenter: a minimizer may place mass outside the chosen sites $(y_i)_i$ and outside the union of the input supports. Once the candidate support is fixed, however, the nonnegative weights λ_s and Proposition 3.6 show that problem (12.2) is convex in a: the discrete Kantorovich value is jointly convex in its two histograms.

Instead of fixing the barycenter support in advance, one may optimize simultaneously over its weights and locations. For quadratic costs on $\mathbb{R}^d$, this gives

$$\min_{\substack{\mathrm{a}\in\Sigma_n \\ Y=(y_i)_{i=1}^n\in(\mathbb{R}^d)^n}} \sum_{s=1}^{S} \lambda_s \mathrm{L}_{\mathrm{C}_s(Y)}(\mathrm{a}, \mathrm{b}_s), \qquad [\mathrm{C}_s(Y)]_{i,j} = \|y_i - x_{s,j}\|^2.$$

For fixed Y, the objective is the convex problem (12.2) in a; its dependence on Y is generally nonconvex, so the joint problem is nonconvex. Cuturi and Doucet [169] proposed alternating weight updates, obtained from dual OT solutions, and location updates, obtained from primal OT plans. Thus choosing the support is itself a significant computational part of the barycenter problem, rather than a harmless preprocessing step.

For quadratic costs, the multi-marginal formulation of Section 12.2 shows that, for discrete inputs, one may choose a barycenter supported on weighted averages $\sum_s \lambda_s x_{s,i_s}$ of one support point from each input. This exact candidate set can contain $\prod_s n_s$ points, but Corollary 12.21 shows that there exists a barycenter for which at most $\sum_s n_s - S + 1$ of them carry positive mass. Prescribing the support $(y_i)_i$ before solving (12.2) is nevertheless a numerical approximation, because the active weighted averages are not known in advance.

Example 12.3 (Application to fair score repair). Let (X, S) be a random pair taking values in $\mathcal{X} \times \{1, \dots, K\}$, where S is the protected attribute and $p_s = \mathbb{P}(S = s) > 0$. Let $f : \mathcal{X} \times \{1, \dots, K\} \to \mathbb{R}$ be a measurable score such that $\mathbb{E}\,|f(X, S)|^2 < +\infty$. For each group, denote the conditional score law by

$$\beta_s := \mathrm{Law}\big(f(X, S) \mid S = s\big).$$

Demographic parity asks these laws to be independent of s. Let $\alpha^\star$ be the quadratic Wasserstein barycenter defined by (12.1), with inputs $(\beta_s)_{s=1}^K$ and weights $(p_s)_{s=1}^K$. Assume that every β_s has a density, and denote by T_s the Brenier map satisfying

$$(T_s)_\sharp \beta_s = \alpha^\star.$$

Since the scores are one-dimensional, T_s is the monotone rearrangement from β_s to $\alpha^\star$. Define the repaired scoring rule by

$$\widetilde{f}(x, s) := T_s\big(f(x, s)\big).$$

Then, for every group s,

$$\mathrm{Law}\big(\widetilde{f}(X, S) \mid S = s\big) = (T_s)_\sharp \beta_s = \alpha^\star.$$

Thus $\widetilde{f}(X, S)$ is independent of S and satisfies demographic parity. Moreover, among repairs obtained by transporting every group to a common score law, the barycenter and the maps $(T_s)_s$ minimize the expected squared score modification [248, 146, 92, 280].

With only two input measures, the barycenter problem introduces no new averaging operation: it selects an intermediate point of the ambient Wasserstein geometry.

Proposition 12.4 (Two-measure barycenters are Wasserstein geodesics)**.** *Let $\beta_0, \beta_1 \in \mathcal{P}_2(\mathbb{R}^d)$ and $t \in [0,1]$. For $S = 2$, $c(x,y) = \|x-y\|^2$ and weights $(1-t, t)$, the minimizers of the barycenter problem (12.1) are exactly the time-t points of constant-speed $\mathcal{W}_2$ geodesics from β_0 to β_1. The minimum value is*

$$t(1-t)\,\mathcal{W}_2^2(\beta_0, \beta_1).$$

In particular, if π is any optimal coupling between β_0 and β_1, then

$$\alpha_t := \big((x,y) \mapsto (1-t)x + ty\big)_\sharp \pi$$

is a barycenter. If β_0 has a density, the barycenter is unique and

$$\alpha_t = ((1-t)\mathrm{Id} + tT)_\sharp \beta_0,$$

where T is the Brenier map from β_0 to β_1.

Proof. The endpoint cases $t \in \{0,1\}$ are immediate, so assume $0 < t < 1$. For any candidate $\alpha \in \mathcal{P}_2(\mathbb{R}^d)$, set

$$A = \mathcal{W}_2(\beta_0, \alpha), \qquad B = \mathcal{W}_2(\alpha, \beta_1), \qquad D = \mathcal{W}_2(\beta_0, \beta_1).$$

The identity

$$(1-t)A^2 + tB^2 = t(1-t)(A+B)^2 + \big((1-t)A - tB\big)^2$$

and the triangle inequality $A + B \geqslant D$ give

$$(1-t)\,\mathcal{W}_2^2(\alpha, \beta_0) + t\,\mathcal{W}_2^2(\alpha, \beta_1) \geqslant t(1-t)D^2.$$

For the interpolation α_t induced by an optimal coupling, Proposition 4.13 gives $A = tD$ and $B = (1-t)D$, so the lower bound is attained.

Conversely, every minimizer must attain equality in both inequalities above. Hence $A + B = D$ and $(1-t)A = tB$, which imply $A = tD$ and $B = (1-t)D$. Parameterize a constant-speed geodesic from β_0 to α on $[0,t]$ and one from α to β_1 on $[t,1]$; both pieces then have speed D. Their concatenation is globally geodesic: any shorter path between points on opposite pieces, combined with the remaining subsegments, would contradict $D = \mathcal{W}_2(\beta_0, \beta_1)$. Thus the concatenated geodesic passes through α at time t. Finally, if β_0 has a density, Brenier's theorem (Theorem 2.23) gives the unique optimal coupling $(\mathrm{Id}, T)_\sharp \beta_0$, and the displayed McCann interpolation follows from Definition 2.29. □

Example 12.5 (Dirac inputs recover Fréchet means). The general barycenter formulation above extends the computation of barycenters of points $(x_s)_{s=1}^S \in \mathcal{X}^S$ to arbitrary measures. Indeed, if $\beta_s = \delta_{x_s}$ is a single Dirac mass, then a solution is $\delta_{x^\star}$, where $x^\star$ minimizes $x \mapsto \sum_s \lambda_s c(x, x_s)$.

Figure 12.1 moves beyond this degenerate case and compares barycenter grids obtained from one-dimensional quantile averaging and two-dimensional entropic transport under the same bilinear corner weights.

Quadratic barycenters inherit both the average translation and a sharp localization property from their inputs.

Proposition 12.6 (Mean and support of quadratic Wasserstein barycenters)**.** *Let $\beta_1, \dots, \beta_S \in \mathcal{P}_2(\mathbb{R}^d)$, let $\lambda \in \Sigma_S$, and let $\alpha^\star$ be a barycenter for the squared Euclidean cost. Define*

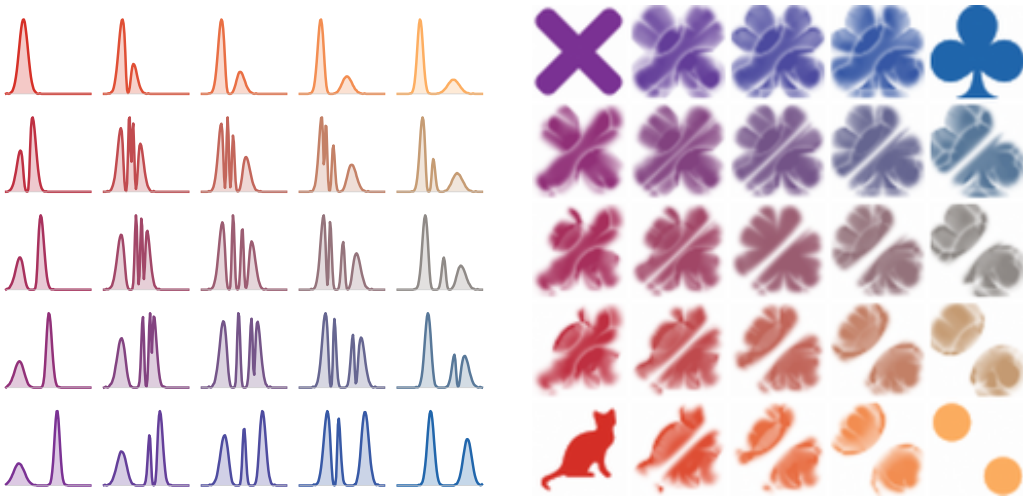
$$K := \overline{\mathrm{conv}}\left(\bigcup_{\{s:\lambda_s > 0\}} \mathrm{supp}(\beta_s) \right).$$

Then

$$\mathrm{supp}(\alpha^\star) \subset K, \qquad \int_{\mathbb{R}^d} x \,\mathrm{d}\alpha^\star(x) = \sum_{s=1}^S \lambda_s \int_{\mathbb{R}^d} x \,\mathrm{d}\beta_s(x).$$

Proof. Write $m_\alpha = \int x \,\mathrm{d}\alpha(x)$ and let α^0 be its centered translate, as in Proposition 4.9. That proposition gives, for every candidate α,

$$\sum_{s=1}^S \lambda_s\, \mathcal{W}_2^2(\alpha, \beta_s) = \sum_{s=1}^S \lambda_s \|m_\alpha - m_{\beta_s}\|^2 + \sum_{s=1}^S \lambda_s\, \mathcal{W}_2^2(\alpha^0, \beta_s^0).$$



one-dimensional quantile barycenters two-dimensional entropic barycenters

Figure 12.1: Wasserstein barycenter grids for four corner measures. The left panel uses the one-dimensional formula $Q_{u,v} = \sum_{i,j} \lambda_{i,j}(u,v) Q_{i,j}$ for one Gaussian law and three asymmetric two-Gaussian mixtures, and displays densities reconstructed from the averaged quantiles. The right panel computes entropic Wasserstein barycenters on a common pixel grid for the cat, two-disk, cross and clover silhouettes, using the normalized squared ground cost, $\varepsilon = 4 \cdot 10^{-4}$ and a Sinkhorn tolerance of $5 \cdot 10^{-8}$. The barycenters are rendered as density images with values clamped at their 95% quantile rather than by threshold contours. Colors interpolate between the four corners and encode the same bilinear weights in both panels.

For a fixed centered law α^0, the second sum is independent of m_α, whereas the first is uniquely minimized at $m_\alpha = \sum_s \lambda_s m_{\beta_s}$. Translating $\alpha^\star$ without changing its centered law therefore proves the mean identity.

For the support statement, let P_K be the Euclidean metric projection onto the closed convex set K. For every $y \in K$, the projection inequality gives

$$\|x - y\|^2 \geqslant \|P_K(x) - y\|^2 + \|x - P_K(x)\|^2.$$

For each s with $\lambda_s > 0$, let π_s be an optimal coupling between $\alpha^\star$ and β_s. Since $\beta_s(K) = 1$, pushing π_s forward by (P_K, Id) and integrating the preceding inequality yields

$$\mathcal{W}_2^2\big((P_K)_\sharp \alpha^\star, \beta_s\big) \leqslant \mathcal{W}_2^2(\alpha^\star, \beta_s) - \int_{\mathbb{R}^d} \mathrm{dist}(x, K)^2 \, \mathrm{d}\alpha^\star(x).$$

After multiplication by λ_s and summation, barycenter optimality forces the last integral to vanish. Hence $\alpha^\star(K) = 1$, and the closedness of K gives $\mathrm{supp}(\alpha^\star) \subset K$. □

If all input supports are compact, their finite union is compact and so is its convex hull; in that case the closure in the definition of K is unnecessary.

One-dimensional case. On the line, barycenters become linear after the quantile change of variables. This gives the rare case where the barycenter is explicit rather than the solution of a high-dimensional optimization problem.

Proposition 12.7 (Quantile barycenters on the line)**.** *For $\mathcal{X} = \mathbb{R}$ and $c(x,y) = |x-y|^2$, the quantile function of a Wasserstein barycenter is the weighted average of the input quantile functions:*

$$C_{\alpha^\star}^{-1}(r) = \sum_{s=1}^{S} \lambda_s C_{\beta_s}^{-1}(r), \qquad r \in (0,1).$$

Proof. The one-dimensional formula (2.11) gives

$$\sum_s \lambda_s \, \mathcal{W}_2^2(\alpha, \beta_s) = \int_0^1 \sum_s \lambda_s \left| C_\alpha^{-1}(r) - C_{\beta_s}^{-1}(r) \right|^2 \mathrm{d}r.$$

The minimization decouples pointwise in r. For each fixed r, the minimizer of $z \mapsto \sum_s \lambda_s |z - C_{\beta_s}^{-1}(r)|^2$ is the weighted average $\sum_s \lambda_s C_{\beta_s}^{-1}(r)$. This function is nondecreasing because it is a positive weighted sum of nondecreasing quantile functions, hence it is a valid quantile function. □

Gaussian case. Gaussian barycenters show that the same separation as in the Gaussian Wasserstein formula (2.19) persists: means average linearly, while covariances average according to the Bures–Wasserstein geometry.

Proposition 12.8 (Nondegenerate Gaussian inputs remain Gaussian)**.** *Let the positive-weight inputs be $\beta_s = \mathcal{N}(m_s, \Sigma_s)$, with $\Sigma_s \succeq 0$, and assume that at least one Σ_s is positive definite. The quadratic Wasserstein barycenter is unique and has the form $\alpha^\star = \mathcal{N}(m, \Sigma)$, where*

$$m = \sum_s \lambda_s m_s.$$

Its positive-definite covariance Σ is the unique minimizer of the Bures objective

$$\mathcal{E}(X) := \sum_s \lambda_s \mathcal{B}(X, \Sigma_s)^2. \tag{12.3}$$

For $X \in \mathbb{S}^d_{++}$, define

$$R_s(X) := \left(X^{1/2}\Sigma_s X^{1/2}\right)^{1/2}, \qquad T_s(X) := X^{-1/2}R_s(X)X^{-1/2}, \tag{12.4}$$

$$\overline{T}(X) := \sum_s \lambda_s T_s(X), \qquad \Psi_\lambda(X) := \sum_s \lambda_s R_s(X). \tag{12.5}$$

Here $T_s(X)$ is the positive-semidefinite linear optimal map from $\mathcal{N}(0, X)$ to $\mathcal{N}(0, \Sigma_s)$. The covariance Σ is equivalently characterized by either of the stationarity equations

$$\overline{T}(\Sigma) = \mathrm{Id} \qquad \Longleftrightarrow \qquad \Psi_\lambda(\Sigma) = \Sigma. \tag{12.6}$$

In dimension one, if $\sigma_s = \sqrt{\Sigma_s}$ denotes the input standard deviation, then the barycenter standard deviation is $\sigma = \sum_s \lambda_s \sigma_s$.

Proof. Let $\mathrm{Gauss}(\alpha)$ denote the Gaussian measure with the same mean and covariance as α. For every competitor $\alpha \in \mathcal{P}_2(\mathbb{R}^d)$ and every Gaussian input β_s, the Gelbrich contraction of Theorem 16.30 gives

$$\mathcal{W}_2^2\big(\mathrm{Gauss}(\alpha), \beta_s\big) = \mathcal{W}_2^2\big(\mathrm{Gauss}(\alpha), \mathrm{Gauss}(\beta_s)\big) \leqslant \mathcal{W}_2^2(\alpha, \beta_s).$$

Summing with weights λ_s shows that moment-matched Gaussian projection cannot increase the barycenter objective. Since a barycenter exists, projecting any minimizer therefore produces a Gaussian barycenter. The input with positive-definite covariance is absolutely continuous; the uniqueness criterion stated after Definition 12.1 then implies that the barycenter itself is this Gaussian measure.

For a Gaussian candidate, formula (2.19) separates the objective as

$$\sum_s \lambda_s \mathcal{W}_2^2\big(\mathcal{N}(m, \Sigma), \mathcal{N}(m_s, \Sigma_s)\big) = \sum_s \lambda_s \|m - m_s\|^2 + \sum_s \lambda_s \mathcal{B}(\Sigma, \Sigma_s)^2.$$

The first term is uniquely minimized at $m = \sum_s \lambda_s m_s$. The second is the Bures barycenter problem. Uniqueness of the Wasserstein barycenter makes its covariance minimizer unique, and the presence of a positive-definite input makes this minimizer positive definite [9, 54].

It remains to derive the fixed-point equation. First suppose that $\Sigma_s \succ 0$. Cyclicity of the trace gives

$$\operatorname{tr} R_s(X) = \operatorname{tr}\left(\Sigma_s^{1/2} X \Sigma_s^{1/2}\right)^{1/2}.$$

For a symmetric perturbation Z, differentiation of the trace of the matrix square root yields

$$D[\operatorname{tr} R_s(X)]\,[Z] = \frac{1}{2}\operatorname{tr}\big(T_s(X)Z\big).$$

Consequently,
$$D\mathcal{E}(X)[Z] = \operatorname{tr}\big((\mathrm{Id} - \overline{T}(X))Z\big). \tag{12.7}$$

The same identity for singular Σ_s follows by replacing Σ_s with $\Sigma_s + \delta\mathrm{Id}$ and letting $\delta \downarrow 0$. Since the minimizer lies in $\mathbb{S}^d_{++}$, first-order optimality is $\overline{T}(\Sigma) = \mathrm{Id}$. Finally, definitions (12.4)–(12.5) give

$$\Psi_\lambda(X) = X^{1/2}\overline{T}(X)X^{1/2},$$

so $\overline{T}(\Sigma) = \mathrm{Id}$ is equivalent to $\Psi_\lambda(\Sigma) = \Sigma$. Conversely, a positive-definite solution of this equation is stationary for the convex objective (12.3), hence is its unique minimizer. In dimension one, $\mathcal{B}(\sigma^2, \sigma_s^2)^2 = (\sigma - \sigma_s)^2$, whose minimizer is $\sigma = \sum_s \lambda_s \sigma_s$. □

If all input covariances are singular, the same contraction argument still gives a Gaussian barycenter, but the uniqueness step can fail and non-Gaussian barycenters may coexist. Thus the nondegeneracy assumption is essential when asserting that every barycenter is Gaussian.

Remark 12.9 (Forward KL barycenter of Gaussian laws). The contrast with Wasserstein averaging is particularly sharp for the forward KL barycenter. Assume now that every $\Sigma_s \succ 0$, and define

$$\alpha_{\mathrm{KL}} := \operatorname{argmin}_{\alpha \in \mathcal{P}(\mathbb{R}^d)} \sum_s \lambda_s \, \mathrm{KL}(\alpha \mid \beta_s).$$

Then $\alpha_{\mathrm{KL}} = \mathcal{N}(m_{\mathrm{KL}}, \Sigma_{\mathrm{KL}})$, with

$$\Sigma_{\mathrm{KL}} = \left(\sum_s \lambda_s \Sigma_s^{-1} \right)^{-1}, \qquad m_{\mathrm{KL}} = \Sigma_{\mathrm{KL}} \sum_s \lambda_s \Sigma_s^{-1} m_s.$$

Thus Σ_{KL} is the weighted harmonic mean of the covariance matrices. Unlike the Wasserstein barycenter in Proposition 12.8, its mean is precision-weighted and therefore depends on the input covariances. Indeed, normalizing the geometric mean $\prod_s (\mathrm{d}\beta_s/\mathrm{d}x)^{\lambda_s}$ gives precisely this Gaussian, and the KL objective differs from $\mathrm{KL}(\alpha \mid \alpha_{\mathrm{KL}})$ only by a constant. This is directional: minimizing $\sum_s \lambda_s \, \mathrm{KL}(\beta_s \mid \alpha)$ over arbitrary α instead gives the mixture $\sum_s \lambda_s \beta_s$, which is generally not Gaussian; see Section 7.3.

The equation $\Psi_\lambda(X) = X$ is a characterization, but it does not by itself prescribe the most natural iteration. Its transport meaning leads to a different update. Once the mean has been set to $m = \sum_s \lambda_s m_s$, the optimal affine map from $\mathcal{N}(m, X)$ to $\mathcal{N}(m_s, \Sigma_s)$ is

$$x \longmapsto m_s + T_s(X)(x - m).$$

Their weighted average is $x \mapsto m + \overline{T}(X)(x - m)$. Pushing the current Gaussian forward by this averaged map gives the *transport fixed-point update* of Álvarez-Esteban et al. [9], whose covariance is

$$\mathcal{K}_1(X) := \overline{T}(X) X \overline{T}(X) = X^{-1/2} \Psi_\lambda(X)^2 X^{-1/2}. \tag{12.8}$$

More generally, damping the displacement by $\eta \in (0, 1]$ replaces the averaged map by

$$B_\eta(X) := (1 - \eta)\mathrm{Id} + \eta \overline{T}(X)$$

and gives

$$\mathcal{K}_\eta(X) := B_\eta(X) X B_\eta(X). \tag{12.9}$$

Thus $\eta = 1$ is the original Álvarez-Esteban et al. iteration [9], while $0 < \eta < 1$ moves only a fraction η along the Bures geodesic from the current Gaussian to its transport update. Since $B_\eta(X)$ is the unique positive-definite optimal map from $\mathcal{N}(0, X)$ to $\mathcal{N}(0, \mathcal{K}_\eta(X))$, one has $\mathcal{K}_\eta(X) = X$ if and only if $B_\eta(X) = \mathrm{Id}$, hence if and only if (12.6) holds. Every $\mathcal{K}_\eta$ therefore has the same positive-definite fixed point.

Proposition 12.10 (Convergence of the Gaussian transport fixed-point iteration)**.** *Assume that $\lambda_s > 0$, $\Sigma_s \succeq 0$, and that at least one Σ_s is positive definite. Fix $\eta \in (0, 1]$ and $X^{(0)} \succ 0$, and iterate*

$$X^{(\ell+1)} = \mathcal{K}_\eta(X^{(\ell)}).$$

Then every $X^{(\ell)}$ is positive definite and $X^{(\ell)} \to \Sigma$, where Σ is the covariance of the Gaussian barycenter in Proposition 12.8. More precisely, with

$$\mathcal{R}(X) := \mathrm{tr}\big((\mathrm{Id} - \overline{T}(X)) X (\mathrm{Id} - \overline{T}(X))\big),$$

one has

$$\mathcal{E}(X^{(\ell+1)}) \leqslant \mathcal{E}(X^{(\ell)}) - \eta(2 - \eta)\mathcal{R}(X^{(\ell)}), \tag{12.10}$$

and therefore $\sum_{\ell \geqslant 0} \mathcal{R}(X^{(\ell)}) < +\infty$. Equivalently,

$$\mathcal{B}\big(X^{(\ell)}, X^{(\ell+1)}\big)^2 = \eta^2 \mathcal{R}(X^{(\ell)}).$$

Proof. Fix $X \succ 0$, abbreviate $T_s = T_s(X)$, $\overline{T} = \overline{T}(X)$ and $B = B_\eta(X)$, and let $Z \sim \mathcal{N}(0, X)$. Since $T_s Z \sim \mathcal{N}(0, \Sigma_s)$, the pair $(BZ, T_s Z)$ is an admissible coupling between $\mathcal{N}(0, \mathcal{K}_\eta(X))$ and $\mathcal{N}(0, \Sigma_s)$. For every $z \in \mathbb{R}^d$, the Euclidean variance decomposition around $\overline{T}z = \sum_s \lambda_s T_s z$ gives

$$\sum_s \lambda_s \|Bz - T_s z\|^2 = \sum_s \lambda_s \|\overline{T}z - T_s z\|^2 + (1-\eta)^2 \|z - \overline{T}z\|^2,$$
$$\sum_s \lambda_s \|z - T_s z\|^2 = \sum_s \lambda_s \|\overline{T}z - T_s z\|^2 + \|z - \overline{T}z\|^2.$$

Integrating and bounding each Wasserstein cost by the admissible coupling $(BZ, T_s Z)$ proves (12.10). Moreover, $B \succ 0$: this is immediate when $\eta < 1$, and for $\eta = 1$ it follows because the weighted sum $\overline{T}$ contains at least one positive-definite term. Hence $z \mapsto Bz$ is the optimal map from $\mathcal{N}(0, X)$ to $\mathcal{N}(0, \mathcal{K}_\eta(X))$, which proves the displayed identity for the Bures step length.

It remains to rule out escape to the boundary of the positive cone. Set

$$\delta(X) := \det(X)^{1/(2d)}, \qquad c_0 := \sum_s \lambda_s \det(\Sigma_s)^{1/(2d)} > 0.$$

The Minkowski determinant inequality, applied to the positive-semidefinite matrices Id and $T_s(X)$ with weights $1-\eta$ and $\eta\lambda_s$, gives

$$\begin{aligned} \delta\big(\mathcal{K}_\eta(X)\big) &= \delta(X) \det(B_\eta(X))^{1/d} \\ &\geqslant (1-\eta)\delta(X) + \eta \sum_s \lambda_s \delta(X) \det(T_s(X))^{1/d} \\ &= (1-\eta)\delta(X) + \eta c_0, \end{aligned}$$

where the last identity uses $T_s(X) X T_s(X) = \Sigma_s$. It follows that $\delta(X^{(\ell)}) \geqslant \min\{\delta(X^{(0)}), c_0\} > 0$ for every ℓ.

The decreasing energy also bounds the traces. Indeed, the Schatten Cauchy–Schwarz inequality in the Bures formula gives

$$\mathcal{B}(X, \Sigma_s)^2 \geqslant \left(\sqrt{\operatorname{tr} X} - \sqrt{\operatorname{tr} \Sigma_s}\right)^2,$$

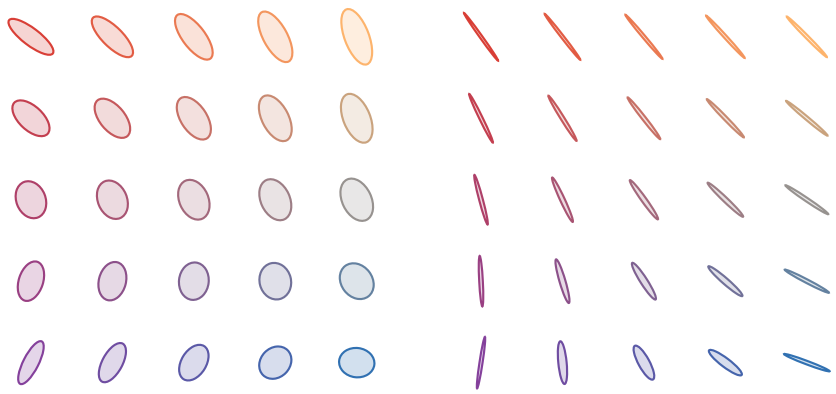
so (12.10) bounds $\operatorname{tr} X^{(\ell)}$ uniformly. A uniform trace bound together with the determinant lower bound places all iterates in a compact subset of $\mathbb{S}^d_{++}$.

Summing (12.10) gives $\mathcal{R}(X^{(\ell)}) \to 0$. Every subsequence therefore has a further subsequence converging to some $X_\infty \succ 0$. Continuity of the matrix square root and inverse on the preceding compact set yields $\mathcal{R}(X_\infty) = 0$, hence $\overline{T}(X_\infty) = \mathrm{Id}$. Proposition 12.8 identifies X_∞ with the unique covariance Σ. Since every cluster point is Σ, the full sequence converges to Σ. The undamped case $\eta = 1$ is the Gaussian specialization of Theorem 4.2 of Álvarez-Esteban et al. [9]; the argument above also proves the damped extension. □

> **Remark 12.11 (Raw covariance Picard iteration).** The transport update (12.8) should not be confused with the raw Picard iteration $X^{(\ell+1)} = \Psi_\lambda(X^{(\ell)})$ suggested by the covariance equation. In dimension one, writing $c = \sum_s \lambda_s \sigma_s$ gives
>
> $$\Psi_\lambda(r) = c\sqrt{r}, \qquad \mathcal{K}_\eta(r) = \big((1-\eta)\sqrt{r} + \eta c\big)^2.$$
>
> Thus the undamped transport update reaches the barycenter variance c^2 in one step, whereas raw Picard generally does not. Proposition 12.10 concerns the transport updates $\mathcal{K}_\eta$, not the raw Picard map; see also [462, 9, 54].

Figure 12.2 illustrates the nonlinear covariance interpolation characterized above; increasing anisotropy makes the simultaneous rotation and rescaling of the Bures–Wasserstein barycenter especially visible.

Sinkhorn for barycenters. A key difference with the regularized two-marginal OT problem is that there is no canonical reference measure $\alpha \otimes \beta$, because the barycenter α is unknown. To reduce complexity, one usually fixes a candidate support for the barycenter and solves the discrete problem (12.2); this introduces a discretization error but keeps the number of unknowns manageable.



moderate anisotropy strong anisotropy

Figure 12.2: Bures–Wasserstein barycenters of centered Gaussian covariance matrices. Each panel shows a 5×5 grid of barycenter ellipses for four corner covariances, without separate input panels: the corner ellipses are the four input covariances themselves. The right grid uses more anisotropic inputs, making the nonlinear rotation and scaling of covariance barycenters more visible.

One can then use the entropy-only convention of (8.1) and approximate (12.2) by

$$\min_{a \in \Sigma_n} \sum_{s=1}^{S} \lambda_s L^{\varepsilon}_{C_s}(a, b_s) \tag{12.11}$$

for some $\varepsilon > 0$. This is a smooth convex minimization problem, which can be tackled using gradient descent [169]. An alternative is to use a descent method, typically quasi-Newton, on the semi-dual [170]; this is useful when adding extra regularization on the barycenter, for instance to impose smoothness.

A simple but effective approach, developed in [37], observes that (12.11) has the same minimizers as the weighted KL projection problem

$$\min_{(P_s)_s} \left\{ \varepsilon \sum_s \lambda_s \mathrm{KL}(P_s | K_s) \; ; \; \forall\, s,\; {P_s}^{\top} \mathbb{1}_n = b_s, \quad P_1 \mathbb{1}_{n_1} = \ldots = P_S \mathbb{1}_{n_S} \right\}, \tag{12.12}$$

where $K_s \coloneqq e^{-C_s/\varepsilon}$. The barycenter a is implicitly encoded in the common row marginal

$$a = P_1 \mathbb{1}_{n_1} = \ldots = P_S \mathbb{1}_{n_S}.$$

The two objectives differ only by constants depending on $(C_s, \varepsilon)_s$, not on the couplings or on the barycenter. Assume below that every C_s is finite and every b_s is positive; zero-weight target atoms can be deleted before running the iteration. The optimal couplings solving (12.12) then have scaling form

$$P_s = \mathrm{diag}(u_s) K_s \, \mathrm{diag}(v_s), \tag{12.13}$$

and the generalized Sinkhorn iterations are

$$\forall\, s \in [\![1, S]\!], \qquad v_s^{(\ell+1)} \coloneqq \frac{b_s}{{K_s}^{\top} u_s^{(\ell)}}, \tag{12.14}$$

$$\forall\, s \in [\![1, S]\!], \qquad u_s^{(\ell+1)} \coloneqq \frac{a^{(\ell+1)}}{K_s v_s^{(\ell+1)}}, \tag{12.15}$$

$$\text{where} \quad a^{(\ell+1)} \coloneqq \prod_s (K_s v_s^{(\ell+1)})^{\lambda_s}. \tag{12.16}$$

The geometric mean in (12.16) enforces the fact that all couplings share the same barycenter marginal.

The scaling cycle has an exact dual-optimization interpretation; it is not merely a sequence of marginal normalizations. Write $u_s = e^{f_s/\varepsilon}$ and $v_s = e^{g_s/\varepsilon}$, with all operations understood componentwise.

Algorithm 12.1: Entropic barycenter Sinkhorn

Input: Finite costs C_s, positive target histograms b_s, barycenter weights $\lambda \in \operatorname{int}(\Sigma_S)$, regularization $\varepsilon > 0$, tolerance tol, iteration budget $L \geqslant 1$.
Output: Barycenter weights a and couplings P_s.
Initialize: Set $K_s = e^{-C_s/\varepsilon}$, $u_s^{(0)} = \mathbb{1}_n$ for all s, $r^{(0)} = +\infty$, and $\ell = 0$.
While $r^{(\ell)} \geqslant$ tol and $\ell < L$ **do**:
  Set $\ell \leftarrow \ell + 1$.
  For each marginal s **do**:
    $v_s^{(\ell)} = \frac{b_s}{K_s{}^\top u_s^{(\ell-1)}}$.
  Compute barycenter marginal: $a^{(\ell)} = \prod_s \left(K_s v_s^{(\ell)}\right)^{\lambda_s}$.
  For each marginal s **do**:
    $u_s^{(\ell)} = \frac{a^{(\ell)}}{K_s v_s^{(\ell)}}$.
  Set $P_s^{(\ell)} = \operatorname{diag}(u_s^{(\ell)}) K_s \operatorname{diag}(v_s^{(\ell)})$ for all s.
  Set $r^{(\ell)} = \max_s \max\{\|P_s^{(\ell)}\mathbb{1} - a^{(\ell)}\|_1, \|(P_s^{(\ell)})^\top \mathbb{1} - b_s\|_1\}$.
Return $a^{(\ell)}$ and $P_s^{(\ell)}$.

Proposition 12.12 below shows that the potentials maximize the concave dual (12.17). With $(f_s)_s$ fixed, this objective separates over s, and its exact maximizer in each g_s is

$$g_s^+ = \varepsilon \log\left(\frac{b_s}{K_s{}^\top e^{f_s/\varepsilon}}\right),$$

which is the v_s-update (12.14). Conversely, with $(g_s^+)_s$ fixed, exact maximization over the coupled block $(f_s)_s$ under $\sum_s \lambda_s f_s = 0$ gives

$$q_s := K_s e^{g_s^+/\varepsilon}, \qquad a^+ = \prod_s q_s^{\lambda_s}, \qquad f_s^+ = \varepsilon \log\left(\frac{a^+}{q_s}\right).$$

Thus a complete generalized Sinkhorn cycle is exact two-block coordinate ascent on the dual, or equivalently alternating minimization of its negative. Indeed, the constraint follows from $\log a^+ = \sum_s \lambda_s \log q_s$, while the first-order conditions require $e^{f_s^+/\varepsilon} q_s = a^+$ for every s. These are precisely (12.15)–(12.16). In the primal formulation (12.12), the same cycle alternates weighted KL projections onto the target-column constraints and the common-row-marginal constraint [37].

Proposition 12.12 (Dual of entropic barycenters)**.** *Under the preceding positivity assumptions, the optimal scalings in* (12.13) *can be written as* $(u_s, v_s) = (e^{f_s/\varepsilon}, e^{g_s/\varepsilon})$*, where* $(f_s, g_s)_s$ *solve the dual problem*

$$\max_{(f_s,g_s)_s} \left\{ \sum_s \lambda_s \left(\langle g_s, b_s\rangle - \varepsilon \langle K_s e^{g_s/\varepsilon}, e^{f_s/\varepsilon}\rangle + \varepsilon \sum_{i,j} (K_s)_{i,j} \right) ; \sum_s \lambda_s f_s = 0 \right\}. \tag{12.17}$$

Proof. Introduce Lagrange multipliers in (12.12):

$$\min_{(P_s)_s, a} \max_{(f_s,g_s)_s} \sum_s \lambda_s \Big(\varepsilon \mathrm{KL}(P_s|K_s) + \langle a - P_s \mathbb{1}_{n_s}, f_s\rangle + \langle b_s - P_s{}^\top \mathbb{1}_n, g_s\rangle \Big).$$

The explicit constraint $a \in \Sigma_n$ may be dropped here: nonnegativity of the couplings, together with $P_s \mathbb{1}_{n_s} = a$ and $P_s{}^\top \mathbb{1}_n = b_s$, already forces $a \in \Sigma_n$. We may therefore minimize the Lagrangian over $a \in \mathbb{R}^n$. Strong duality holds, so one can exchange the minimum and maximum. Finiteness of the minimum with respect to a gives the vector constraint $\sum_s \lambda_s f_s = 0$, while minimization with respect to P_s gives the Legendre transform of $\mathrm{KL}(\cdot|K_s)$:

$$\max_{(f_s,g_s)_s} \sum_s \lambda_s \left[\langle g_s, b_s\rangle - \varepsilon \mathrm{KL}^*\left(\frac{f_s \oplus g_s}{\varepsilon} \middle| K_s \right) \right], \qquad \sum_s \lambda_s f_s = 0.$$

The separable conjugate is

$$\mathrm{KL}^*(U|K) = \sum_{i,j} K_{i,j} \left(e^{U_{i,j}} - 1\right), \tag{12.18}$$

because for $k > 0$,

$$\sup_{r \geqslant 0} ur - \big(r \log(r/k) - r + k\big) = k(e^u - 1),$$

and the case $k = 0$ follows by lower semicontinuity. Substituting this conjugate gives (12.17), including its additive constant. The coordinate maximization in g_s gives (12.14); the block maximization in all $(f_s)_s$ under $\sum_s \lambda_s f_s = 0$ gives the weighted geometric mean (12.16) and then (12.15). □

Classical applications include two-dimensional image interpolation, three-dimensional shape interpolation, and barycenters on surfaces where the ground cost is the square of the geodesic distance; see [495] for applications to computer graphics and imaging.

Wasserstein-over-Wasserstein and barycenters. The barycenter formula does not require a finite list of inputs. The Wasserstein-over-Wasserstein viewpoint of Section 13.2 makes it possible to replace the discrete family $(\beta_s, \lambda_s)_s$ by a law $\mathfrak{A} \in \mathcal{P}(\mathcal{P}(\Omega))$ over probability measures. Such *population Wasserstein barycenters* were studied for random probability measures by Bigot and Klein, in general geodesic settings by Le Gouic and Loubes, and on Riemannian manifolds by Kim and Pass [58, 329, 306]. Assume, for instance, that c is lower semicontinuous and that there exists at least one $\beta_0 \in \mathcal{P}(\Omega)$ with

$$\int_{\mathcal{P}(\Omega)} \mathcal{L}_c(\beta_0, \alpha) \mathrm{d}\mathfrak{A}(\alpha) < +\infty,$$

together with the usual compactness or coercivity hypotheses ensuring existence of minimizers. For instance, these assumptions are automatic when Ω is compact and c is continuous. The barycenter correspondence is then

$$\mathcal{B}_c(\mathfrak{A}) := \operatorname*{Argmin}_{\beta \in \mathcal{P}(\Omega)} \int_{\mathcal{P}(\Omega)} \mathcal{L}_c(\beta, \alpha) \mathrm{d}\mathfrak{A}(\alpha).$$

When this set is a singleton, we denote its element by

$$\widetilde{\alpha}_{\mathfrak{A}} := \operatorname*{argmin}_{\beta \in \mathcal{P}(\Omega)} \int_{\mathcal{P}(\Omega)} \mathcal{L}_c(\beta, \alpha) \mathrm{d}\mathfrak{A}(\alpha),$$

which defines a nonlinear flattening map $\mathfrak{A} \mapsto \widetilde{\alpha}_{\mathfrak{A}}$ from laws over measures back to measures on Ω. When $\mathfrak{A} = \sum_s \lambda_s \delta_{\beta_s}$, this is exactly the finite barycenter problem above. This map should be contrasted with the linear collapsed, or barycentric, mixture $\bar{\alpha}_{\mathfrak{A}}$ of Definition 13.9, which simply averages the input measures themselves. The two operations agree in degenerate linear situations, but in general $\widetilde{\alpha}_{\mathfrak{A}}$ is a geometric average in transport space, whereas $\bar{\alpha}_{\mathfrak{A}}$ is an ordinary mixture in the ambient linear space of measures.

The next result records the corresponding law of large numbers. It is useful when a dataset is itself made of probability measures, for instance populations of histograms, posterior distributions or shapes. We state it in the compact setting, where no moment or tightness side conditions are needed; non-compact extensions require the usual integrability assumptions. Consistency of Wasserstein barycenters and related statistical constructions is developed in [71, 329, 559]; streaming and large-scale uses of many input measures appear for instance in [504, 502, 503].

Proposition 12.13 (Law of large numbers for barycenters over measures)**.** *Let Ω be a compact metric space, let $c : \Omega \times \Omega \to \mathbb{R}$ be continuous, and let $\mathfrak{A} \in \mathcal{P}(\mathcal{P}(\Omega))$. Let $\alpha_1, \alpha_2, \dots$ be independent random probability measures with common law $\mathfrak{A}$, and set*

$$\widehat{\mathfrak{A}}_p := \frac{1}{p} \sum_{i=1}^{p} \delta_{\alpha_i} \in \mathcal{P}(\mathcal{P}(\Omega)).$$

Then, almost surely,

$$\widehat{\mathfrak{A}}_p \rightharpoonup \mathfrak{A} \ \text{ in } \mathcal{P}(\mathcal{P}(\Omega)), \qquad \bar{\alpha}_{\widehat{\mathfrak{A}}_p} \rightharpoonup \bar{\alpha}_{\mathfrak{A}} \ \text{ in } \mathcal{P}(\Omega). \tag{12.19}$$

Moreover, if $\mathcal{B}_c(\mathfrak{A}) = \{\widetilde{\alpha}_{\mathfrak{A}}\}$ is a singleton and if $\widetilde{\alpha}_{\widehat{\mathfrak{A}}_p} \in \mathcal{B}_c(\widehat{\mathfrak{A}}_p)$ is any empirical barycenter, then, almost surely,

$$\widetilde{\alpha}_{\widehat{\mathfrak{A}}_p} \rightharpoonup \widetilde{\alpha}_{\mathfrak{A}} \quad \text{in } \mathcal{P}(\Omega). \tag{12.20}$$

Proof. Set $K = \mathcal{P}(\Omega)$. Since Ω is compact metric, K is compact metric for weak convergence, and so is $\mathcal{P}(K)$. The space $C(K)$ is separable for the uniform norm. Applying the scalar strong law to a countable dense family of test functions and then using uniform approximation gives, almost surely, for every $\Phi \in C(K)$,

$$\int \Phi(\alpha)\mathrm{d}\widehat{\mathfrak{A}}_p(\alpha) = \frac{1}{p}\sum_{i=1}^{p} \Phi(\alpha_i) \longrightarrow \int \Phi(\alpha)\mathrm{d}\mathfrak{A}(\alpha)$$

This is exactly (12.19).

For the collapsed mixtures, take $f \in C(\Omega)$ and define $\Phi_f(\alpha) = \int_\Omega f\,\mathrm{d}\alpha$. This function is continuous on $\mathcal{P}(\Omega)$. Hence

$$\int_\Omega f\,\mathrm{d}\bar{\alpha}_{\widehat{\mathfrak{A}}_p} = \int_{\mathcal{P}(\Omega)} \Phi_f(\alpha)\mathrm{d}\widehat{\mathfrak{A}}_p(\alpha) \longrightarrow \int_{\mathcal{P}(\Omega)} \Phi_f(\alpha)\mathrm{d}\mathfrak{A}(\alpha) = \int_\Omega f\,\mathrm{d}\bar{\alpha}_{\mathfrak{A}},$$

which proves the second convergence in (12.19).

It remains to prove the nonlinear barycenter consistency. Since c is continuous on the compact product $\Omega \times \Omega$, the value $(\beta, \alpha) \mapsto \mathcal{L}_c(\beta, \alpha)$ is continuous on K^2. Therefore the map $\beta \mapsto h_\beta$, where $h_\beta(\alpha) = \mathcal{L}_c(\beta, \alpha)$, is continuous from the compact space K to $C(K)$. Its image $\mathcal{H} := \{h_\beta : \beta \in K\}$ is compact in $C(K)$. The convergence of $\widehat{\mathfrak{A}}_p$ to $\mathfrak{A}$ is uniform over $\mathcal{H}$: given $\eta > 0$, cover $\mathcal{H}$ by finitely many η-balls in $\|\cdot\|_\infty$, use weak convergence for the centers, and bound the error on the balls by the total variation of $\widehat{\mathfrak{A}}_p - \mathfrak{A}$. Hence the empirical objectives

$$F_p(\beta) := \int \mathcal{L}_c(\beta, \alpha)\mathrm{d}\widehat{\mathfrak{A}}_p(\alpha)$$

converge uniformly on K to

$$F(\beta) := \int \mathcal{L}_c(\beta, \alpha)\mathrm{d}\mathfrak{A}(\alpha).$$

Let $\beta_p \in \mathcal{B}_c(\widehat{\mathfrak{A}}_p)$. Compactness gives a subsequence $\beta_{p_k} \rightharpoonup \beta$. Uniform convergence, continuity of F, and optimality of β_{p_k} give, for any $\gamma \in K$,

$$F(\beta) = \lim_k F_{p_k}(\beta_{p_k}) \leqslant \lim_k F_{p_k}(\gamma) = F(\gamma),$$

so $\beta \in \mathcal{B}_c(\mathfrak{A})$. If this set is the singleton $\{\widetilde{\alpha}_{\mathfrak{A}}\}$, every converging subsequence has the same limit, and therefore the whole sequence $\widetilde{\alpha}_{\widehat{\mathfrak{A}}_p}$ converges to $\widetilde{\alpha}_{\mathfrak{A}}$, proving (12.20). □

The first convergence in (12.19) is the classical law of large numbers on Wasserstein space, and the second is its linear image under the collapse map. By contrast, (12.20) is nonlinear: it recomputes a Wasserstein barycenter from the empirical law over measures. The number p of input measures should not be confused with the number n of samples used to approximate each input measure, studied in Section 10.2. In applications one often observes p empirical measures, each made of roughly n atoms, hence about np points in total. Balancing the error due to finitely many input laws against the error due to finitely sampled input laws is a separate statistical and computational tradeoff.

Toward central limit theorems on Wasserstein space. The same hierarchy suggests a central-limit refinement of the preceding law of large numbers, but the nonlinear geometry makes this substantially more delicate. For the linear collapse $\bar{\alpha}_{\widehat{\mathfrak{A}}_p}$, testing against a fixed $f \in C(\Omega)$ reduces the question to the classical scalar central limit theorem for the random variable $\alpha \mapsto \int f\,\mathrm{d}\alpha$. For the nonlinear barycenter $\widetilde{\alpha}_{\widehat{\mathfrak{A}}_p}$, however, there is no canonical vector difference $\widetilde{\alpha}_{\widehat{\mathfrak{A}}_p} - \widetilde{\alpha}_{\mathfrak{A}}$ inside $\mathcal{P}(\Omega)$. One has to choose a local linearization. When the population barycenter is sufficiently regular so that the optimal map T_p from $\widetilde{\alpha}_{\mathfrak{A}}$ to $\widetilde{\alpha}_{\widehat{\mathfrak{A}}_p}$ exists, this amounts to asking whether $\sqrt{p}\,(T_p - \mathrm{Id})$ converges in a Hilbert space such as $L^2(\widetilde{\alpha}_{\mathfrak{A}})$. In nonsmooth settings one must instead work with optimal-plan or logarithmic-map coordinates. Even after such a linearization, an infinite-dimensional CLT requires tightness of the rescaled tangent variables and a genuine Radon Gaussian random element; in a Hilbert space, the associated covariance must be trace class. A cylindrical Gaussian limit alone is therefore not a probability law on the tangent space. This obstruction explains why Wasserstein-space CLTs are more rigid than the weak laws above.

There are nevertheless important settings where such results can be proved. In one dimension, the quantile representation linearizes $\mathcal{W}_2$, so barycenter fluctuations can be studied through empirical averages of quantile functions. Another finite-dimensional case is the family of non-degenerate Gaussian measures in fixed dimension, where $\mathcal{W}_2$ reduces to the Bures geometry of means and covariance matrices. Agueh and Carlier [2] formulate this Wasserstein-barycenter CLT precisely in tangent coordinates and prove it in a few special cases, including the one-dimensional non-atomic setting and finite laws supported on non-degenerate Gaussian measures. Entropic barycenters give a smoother variant for which central-limit theorems for empirical barycenters are also available [109]. These results should be read as nonlinear analogues of the statistical limits discussed in Chapter 10, not as a generic Hilbert-space CLT valid on all of $\mathcal{P}(\Omega)$.

Sliced and Radon barycenters. Slicing gives a scalable surrogate for high-dimensional barycenters by applying the one-dimensional quantile formula of Proposition 12.7 in every projection direction. For measures on $\mathbb{R}^d$, define the sliced barycenter by replacing $\mathcal{W}_2$ with SW_2, introduced in Definition 11.22 and interpreted through the Radon transform in Section 11.2:

$$\min_{\alpha\in\mathcal{M}_+^1(\mathbb{R}^d)} \sum_{s=1}^S \lambda_s \,\mathrm{SW}_2(\alpha,\beta_s)^2 = \min_{\alpha\in\mathcal{M}_+^1(\mathbb{R}^d)} \int_{\mathbb{S}^{d-1}} \sum_{s=1}^S \lambda_s \,\mathcal{W}_2\left((P_\theta)_\sharp\alpha,(P_\theta)_\sharp\beta_s\right)^2 \mathrm{d}\sigma(\theta).$$

The constraint that all projected measures come from the same α is the nontrivial part of this problem. A cheaper Radon-domain approximation drops this consistency constraint and minimizes directly over collections $(\gamma_\theta)_{\theta\in\mathbb{S}^{d-1}}$ of one-dimensional probability measures,

$$\min_{(\gamma_\theta)_\theta} \int_{\mathbb{S}^{d-1}} \sum_{s=1}^S \lambda_s \,\mathcal{W}_2(\gamma_\theta,\beta_{s,\theta})^2 \mathrm{d}\sigma(\theta), \qquad \beta_{s,\theta} := (P_\theta)_\sharp\beta_s.$$

This problem decouples over directions. For each θ, the minimizer is the one-dimensional barycenter γ_θ with quantile

$$C_{\gamma_\theta}^{-1}(r) = \sum_{s=1}^S \lambda_s C_{\beta_{s,\theta}}^{-1}(r), \qquad r\in[0,1].$$

For two inputs β_0,β_1 and interpolation weight $t\in[0,1]$, it is useful to make this directionwise construction explicit. Define

$$Q_i(\theta,r) := C_{\beta_{i,\theta}}^{-1}(r), \qquad Q_t(\theta,r) := (1-t)Q_0(\theta,r)+tQ_1(\theta,r), \qquad \gamma_{t,\theta} := \big(Q_t(\theta,\cdot)\big)_\sharp\mathrm{Leb}_{[0,1]}. \tag{12.21}$$

Thus Q_t is the quantile field of the independently interpolated projected laws $(\gamma_{t,\theta})_\theta$. When these laws have densities, we denote them by $h_t(\theta,\cdot)$. This relaxed value is a lower bound on the sliced barycenter value. If the minimizing family is Radon-consistent, meaning that there exists a probability measure $\bar\alpha$ on $\mathbb{R}^d$ such that $\gamma_\theta = (P_\theta)_\sharp\bar\alpha$ for σ-almost every θ, then this lower bound is attained and $\bar\alpha$ is an exact sliced barycenter. In general, however, independently computed one-dimensional barycenters do not satisfy the range conditions of the Radon transform. One therefore reconstructs a density in a least-squares sense, usually through a regularized Radon pseudoinverse, as explained next.

Let $h(\theta,t)$ denote a density of γ_θ. We use the one-dimensional Fourier transform in the variable t,

$$\widehat{h}(\theta,\omega) := \int_{\mathbb{R}} e^{-\mathrm{i}\omega t}h(\theta,t)\mathrm{d}t, \qquad h(\theta,t) = \frac{1}{2\pi}\int_{\mathbb{R}} e^{\mathrm{i}\omega t}\widehat{h}(\theta,\omega)\mathrm{d}\omega, \tag{12.22}$$

whenever the inversion formula is valid.

Proposition 12.14 (Radon least-squares pseudoinverse)**.** *Let $d\geqslant 2$, let σ be the uniform probability measure on $\mathbb{S}^{d-1}$, and let R be the density Radon transform from Remark 11.24. Write*

$$\mathcal{D}(R) := \left\{\rho\in L^2(\mathbb{R}^d)\,;\; R\rho\in L^2(\mathbb{S}^{d-1}\times\mathbb{R},\mathrm{d}\sigma\,\mathrm{d}t)\right\}.$$

Let $h\in L^2(\mathbb{S}^{d-1}\times\mathbb{R},\mathrm{d}\sigma\,\mathrm{d}t)$ and assume that the Fourier expressions below define an element of $\mathcal{D}(R)$. Then the unique solution of

$$\min_{\rho\in\mathcal{D}(R)} \int_{\mathbb{S}^{d-1}}\int_{\mathbb{R}} |R\rho(\theta,t)-h(\theta,t)|^2\,\mathrm{d}t\,\mathrm{d}\sigma(\theta) \tag{12.23}$$

is $\rho^\dagger = R^\dagger h$, where

$$R^\dagger h(x) = \frac{\left|\mathbb{S}^{d-1}\right|}{2(2\pi)^d} \int_{\mathbb{S}^{d-1}} \int_{\mathbb{R}} e^{\mathrm{i}\omega\langle\theta, x\rangle} |\omega|^{d-1} \widehat{h}(\theta, \omega) \mathrm{d}\omega \, \mathrm{d}\sigma(\theta). \tag{12.24}$$

Thus $R^\dagger h$ is the density of the pseudoinverse reconstruction, which is a priori a signed measure. If $h = R\rho$ is Radon-consistent and ρ is sufficiently regular, then $R^\dagger R\rho = \rho$.

Proof. For the d-dimensional Fourier transform, use the compatible convention

$$\widehat{\rho}(\xi) = \int_{\mathbb{R}^d} e^{-\mathrm{i}\langle\xi, x\rangle} \rho(x) \mathrm{d}x, \qquad \rho(x) = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} e^{\mathrm{i}\langle\xi, x\rangle} \widehat{\rho}(\xi) \mathrm{d}\xi.$$

The Fourier-slice theorem gives $$\widehat{R\rho}(\theta, \omega) = \widehat{\rho}(\omega\theta). \tag{12.25}$$

Applying the one-dimensional Plancherel identity in t rewrites the objective in (12.23), up to the positive factor $1/(2\pi)$, as

$$\int_{\mathbb{S}^{d-1}} \int_{\mathbb{R}} \left|\widehat{\rho}(\omega\theta) - \widehat{h}(\theta, \omega)\right|^2 \mathrm{d}\omega \, \mathrm{d}\sigma(\theta).$$

Every $\xi \neq 0$ has the two signed-polar representations $(\theta, \omega) = (\xi/\|\xi\|, \|\xi\|)$ and $(-\xi/\|\xi\|, -\|\xi\|)$. The pointwise least-squares minimizer is therefore

$$\widehat{\rho^\dagger}(\xi) = \frac{1}{2}\left(\widehat{h}\left(\frac{\xi}{\|\xi\|}, \|\xi\|\right) + \widehat{h}\left(-\frac{\xi}{\|\xi\|}, -\|\xi\|\right)\right). \tag{12.26}$$

This also proves uniqueness under the stated existence assumptions. Applying the inverse d-dimensional Fourier transform to (12.26) and using signed polar coordinates yields (12.24); the factor $\left|\mathbb{S}^{d-1}\right|/2$ accounts for the normalization of σ and the two signed representations of each nonzero frequency. Finally, if $h = R\rho$, then (12.25) makes both terms in (12.26) equal to $\widehat{\rho}(\xi)$, proving $R^\dagger R\rho = \rho$. □

Formula (12.24) is the filtered back-projection representation of the Radon pseudoinverse used in tomography [274]; $|\omega|^{d-1}$ is its ramp multiplier. Because this multiplier amplifies high frequencies, one typically chooses a bandwidth $\Omega > 0$ and replaces it by

$$m_\Omega(\omega) := |\omega|^{d-1} \chi(\omega/\Omega), \qquad R_\Omega^\dagger h(x) := \frac{\left|\mathbb{S}^{d-1}\right|}{2(2\pi)^d} \int_{\mathbb{S}^{d-1}} \int_{\mathbb{R}} e^{\mathrm{i}\omega\langle\theta, x\rangle} m_\Omega(\omega) \widehat{h}(\theta, \omega) \mathrm{d}\omega \, \mathrm{d}\sigma(\theta), \tag{12.27}$$

where χ is an even low-pass window with $\chi(0) = 1$. For example, the numerical reconstruction in Figure 12.3 uses the super-Gaussian window $\chi(s) = e^{-|s|^4}$. For the two-input path (12.21), choose $\eta_t \geqslant 0$ so that the positive part below has nonzero mass, and define the nonnegative, unit-mass reconstruction

$$A_t(x) := \frac{\left((R_\Omega^\dagger h_t)(x) - \eta_t\right)_+}{\displaystyle\int_{\mathbb{R}^d} \left((R_\Omega^\dagger h_t)(z) - \eta_t\right)_+ \mathrm{d}z}. \tag{12.28}$$

For the endpoints, set $A_i = \rho_i$ when $\beta_i = \rho_i \mathrm{d}x$, $i \in \{0, 1\}$. In Figure 12.3, the small threshold η_t only suppresses finite-angle inversion ghosts. The resulting regularized density is generally only a least-squares approximation to the independently averaged slices. This construction is fast and was introduced for sliced and Radon Wasserstein barycenters in [76], but it should not be confused with the exact variational sliced barycenter. In the continuous all-directions setting, R is injective by the Cramér–Wold theorem [162]; inconsistency comes instead from failure of Radon range conditions such as antipodal symmetry and moment consistency. With finitely many directions, the sampled Radon operator is also non-injective, so the pseudoinverse reconstruction and the constrained sliced barycenter may differ.

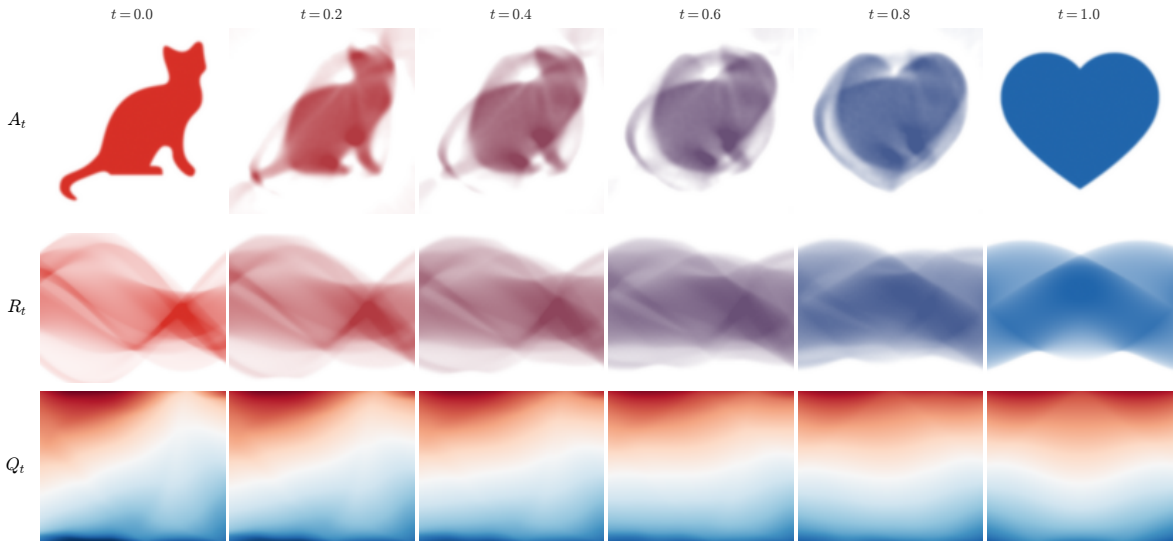


Figure 12.3: Radon-domain sliced barycentric interpolation between the cat and heart densities. The columns correspond to $t = 0, 0.2, \dots, 1$. The first row shows the endpoint densities and the intermediate reconstructions A_t defined in (12.28) from the windowed pseudoinverse (12.27). The second row shows the projected-density fields h_t (labeled R_t in the figure), obtained by converting the directionwise quantile barycenters back into one-dimensional densities. The third row shows the quantile fields Q_t defined in (12.21). This illustrates both the simplicity of one-dimensional quantile interpolation and the extra pseudoinverse reconstruction step needed to return from Radon data to an image density.

12.2 Multimarginal OT

Multi-marginal OT couples more than two measures at once. It is the natural language for barycenters, matching with teams and several-body costs, but its tensor dimension is the main computational obstacle.

Definition and basic structure. The multi-marginal formulation replaces a coupling between two measures by a joint distribution with several prescribed marginals.

Definition 12.15 (Multimarginal optimal transport)**.** Let $\alpha_s \in \mathcal{P}(\mathcal{X}_s)$ for $s = 1, \dots, S$, and denote by $\mathcal{U}(\alpha_1, \dots, \alpha_S)$ the probability measures whose s-th marginal is α_s. For a measurable cost c bounded from below, define

$$\mathrm{MMOT}_c(\alpha_1, \dots, \alpha_S) := \inf_{\pi \in \mathcal{U}(\alpha_1, \dots, \alpha_S)} \int_{\mathcal{X}_1 \times \cdots \times \mathcal{X}_S} c(x_1, \dots, x_S) \mathrm{d}\pi(x_1, \dots, x_S). \tag{12.29}$$

This is still a linear program in the discrete setting, but its ambient tensor has size $\prod_s n_s$.

Monge structure and splitting-set twist. As in the two-marginal case, one would like to know when the optimal joint law is induced by deterministic maps from one marginal. The relevant non-degeneracy assumption is stronger than pairwise twist, because the other $S - 1$ variables have to be recovered simultaneously. The condition below is the standard multi-marginal analogue used in the Monge-structure theory of Gangbo–Swiech and Pass [232, 416, 417, 418].

Definition 12.16 (Twist on splitting sets)**.** Fix $x_1 \in \mathcal{X}_1$. A set $M \subset \mathcal{X}_2 \times \cdots \times \mathcal{X}_S$ is a c-splitting set at x_1 if there exist functions $u_s : \mathcal{X}_s \to \mathbb{R} \cup \{-\infty\}$, for $s = 2, \dots, S$, such that

$$\sum_{s=2}^{S} u_s(x_s) \leqslant c(x_1, x_2, \dots, x_S)$$

for all $(x_2, \dots, x_S)$, with equality on M. Assume c is differentiable in x_1. The cost is twisted on splitting sets if, for every x_1 and every c-splitting set M at x_1, the map

$$(x_2, \dots, x_S) \longmapsto \nabla_{x_1} c(x_1, x_2, \dots, x_S)$$

is injective on M.

Proposition 12.17 (Multi-marginal Monge structure)**.** *Assume that $\mathcal{X}_s \subset \mathbb{R}^d$, that c is continuous and differentiable with respect to x_1, and that c is twisted on splitting sets. Suppose that Kantorovich dual maximizers $(\varphi_s)_{s=1}^S$ exist, that α_1 is absolutely continuous, and that φ_1 is differentiable α_1-a.e. Then every optimal plan $\pi^\star \in \mathcal{U}(\alpha_1, \dots, \alpha_S)$ is concentrated on the graph of maps*

$$\pi^\star = (\mathrm{Id}, T_2, \dots, T_S)_\sharp \alpha_1, \qquad (T_s)_\sharp \alpha_1 = \alpha_s.$$

In particular, under these hypotheses the optimizer is unique.

Proof. Let $(\varphi_s)_{s=1}^S$ be optimal dual potentials. Complementary slackness gives a Borel contact set Γ of full $\pi^\star$-measure on which $\sum_s \varphi_s(x_s) = c(x_1, \dots, x_S)$. After disintegrating with respect to the first marginal, fix a point x_1 where φ_1 is differentiable and where the conditional plan is concentrated on the fiber

$$M(x_1) = \{(x_2, \dots, x_S) : (x_1, x_2, \dots, x_S) \in \Gamma\}.$$

For this fixed x_1, the fiber is a splitting set. Indeed, set $u_2 = \varphi_2 + \varphi_1(x_1)$ and $u_s = \varphi_s$ for $s \geqslant 3$. Dual feasibility gives $\sum_{s=2}^S u_s(x_s) \leqslant c(x_1, x_2, \dots, x_S)$, and equality holds on $M(x_1)$. If $(x_2, \dots, x_S) \in M(x_1)$, the function

$$z \longmapsto c(z, x_2, \dots, x_S) - \sum_{s=2}^S \varphi_s(x_s)$$

touches φ_1 from above at $z = x_1$. Differentiating at this contact point gives

$$\nabla \varphi_1(x_1) = \nabla_{x_1} c(x_1, x_2, \dots, x_S).$$

All points in the fiber therefore have the same value of $\nabla_{x_1} c$. Twist on splitting sets makes the fiber a singleton for α_1-a.e. x_1. Disintegrating $\pi^\star$ with respect to its first marginal gives Dirac conditional measures, hence measurable maps $(T_2, \dots, T_S)$. If π^1 and π^2 are two optimal plans, their average is also optimal. The conditional measure of this average over x_1 is the average of the two Dirac conditionals, and it must again be a Dirac mass by the preceding argument. Hence the two Dirac masses coincide for α_1-a.e. x_1, proving uniqueness. □

Remark 12.18 (Recovery of the two-marginal theory). When $S = 2$, twist on splitting sets is exactly the usual twist condition of Definition 2.35. Indeed, for fixed x, the whole target space is a splitting set by taking $u_2(y) = c(x, y)$; hence the condition requires $y \mapsto \nabla_x c(x, y)$ to be injective. At a dual contact point,

$$\nabla \varphi_1(x) = \nabla_x c(x, y),$$

so this injectivity selects a unique $y = T(x)$. Provided the Kantorovich problem admits an optimizer, the proposition therefore makes that optimizer equal to $(\mathrm{Id}, T)_\sharp \alpha_1$: the relaxation is tight and T is an optimal Monge map. This is precisely the two-marginal mechanism isolated in Proposition 2.36. For the quadratic cost $c(x, y) = \|x - y\|^2$, one has $\nabla_x c(x, y) = 2(x - y)$ and therefore

$$T(x) = x - \frac{1}{2} \nabla \varphi_1(x) = \nabla \left(\frac{1}{2} \|x\|^2 - \frac{1}{2} \varphi_1(x) \right).$$

Choosing the usual c-concave representative of the quadratic dual potential makes the potential in parentheses convex; it is differentiable α_1-almost everywhere because α_1 is absolutely continuous. Thus the case $S = 2$ recovers the convex-gradient map, uniqueness, and tightness conclusions of Brenier's theorem 2.23 and Corollary 3.36. For $S > 2$, the splitting-set twist is the stronger requirement that the same first-order identity recover the entire tuple $(x_2, \dots, x_S)$ at once.

Coulomb cost and density-functional theory. A second canonical example, besides barycenters, comes from electronic structure. For N electrons in $\mathbb{R}^3$, the repulsive Coulomb interaction is the multi-body cost

$$c_{\mathrm{Coul}}(x_1, \dots, x_N) := \sum_{1 \leqslant i < j \leqslant N} \frac{1}{\|x_i - x_j\|},$$

with the value $+\infty$ on the collision set. Proposition 12.17 therefore does not apply verbatim: the Coulomb cost is neither finite nor differentiable on the whole product space. Any finite-energy plan gives zero

mass to exact collisions, so the cost is smooth at almost every point charged by the plan, but this removes only the singularity; one must still establish differentiability of the dual potential and twist on the relevant splitting sets. Away from collisions,

$$\nabla_{x_1} c_{\mathrm{Coul}}(x_1, \dots, x_N) = -\sum_{j=2}^{N} \frac{x_1 - x_j}{\|x_1 - x_j\|^3}.$$

For $N = 2$, the map from x_2 to this vector is injective, so the ordinary two-marginal twist argument can be recovered under the required existence, duality and differentiability hypotheses. For $N \geqslant 3$, however, the displayed total force does not by itself determine the entire tuple $(x_2, \dots, x_N)$; twist on splitting sets, and hence a Monge representation, is not automatic. The previous proposition thus supplies a mechanism to verify in special Coulomb models, not a general existence theorem for co-motion maps.

If ρ is an electron density with $\int_{\mathbb{R}^3} \rho(x)\mathrm{d}x = N$ and $\alpha = \rho/N$ is the associated probability density, the strictly-correlated-electrons relaxation of density-functional theory is the equal-marginal problem

$$V_{\mathrm{ee}}^{\mathrm{SCE}}[\rho] := \inf_{\pi \in \mathcal{U}(\alpha,\dots,\alpha)} \int_{(\mathbb{R}^3)^N} c_{\mathrm{Coul}}(x_1, \dots, x_N)\mathrm{d}\pi(x_1, \dots, x_N).$$

Since the cost and constraints are permutation invariant, symmetrizing any admissible plan does not change its value, so one may equivalently minimize over symmetric plans. This functional gives the smallest possible electron–electron repulsion compatible with the prescribed one-particle density; it appears as the strong-interaction limit in density-functional theory and was connected to optimal transport in [249, 91, 158, 187]. The deterministic ansatz writes a plan through co-motion maps

$$\pi = (\mathrm{Id}, T_2, \dots, T_N)_\sharp \alpha, \qquad (T_i)_\sharp \alpha = \alpha,$$

so that the position of one electron determines the positions of the others. The following cyclic version is the most common structural form of this ansatz in the strictly-correlated-electrons literature.

Proposition 12.19 (Cyclic co-motion plans)**.** *Let $\alpha \in \mathcal{P}(\mathbb{R}^3)$ and let $T : \mathbb{R}^3 \to \mathbb{R}^3$ be measurable with $T_\sharp \alpha = \alpha$ and $T^N = \mathrm{Id}$ α-a.e. Set $T^0 = \mathrm{Id}$ and*

$$\pi_T = (T^0, T^1, \dots, T^{N-1})_\sharp \alpha.$$

Then $\pi_T \in \mathcal{U}(\alpha, \dots, \alpha)$. If $R_\sigma(x_1, \dots, x_N) = (x_{\sigma(1)}, \dots, x_{\sigma(N)})$ for $\sigma \in \mathrm{Perm}(N)$, then

$$\bar{\pi}_T := \frac{1}{N!} \sum_{\sigma \in \mathrm{Perm}(N)} (R_\sigma)_\sharp \pi_T$$

is a symmetric admissible plan and

$$\int c_{\mathrm{Coul}} \mathrm{d}\bar{\pi}_T = \int c_{\mathrm{Coul}} \mathrm{d}\pi_T = \int_{\mathbb{R}^3} \sum_{0 \leqslant i < j \leqslant N-1} \frac{1}{\|T^i(x) - T^j(x)\|} \mathrm{d}\alpha(x).$$

Proof. Since $T_\sharp \alpha = \alpha$, all iterates T^i preserve α, so every marginal of π_T is α. The plan $\bar{\pi}_T$ is an average of coordinate permutations of π_T, hence has the same marginals and is invariant under coordinate permutations. The Coulomb cost is symmetric in its arguments, so its integral is unchanged by each R_σ. The last identity follows by evaluating c_{Coul} on the graph $(x, T(x), \dots, T^{N-1}(x))$. □

Proposition 12.17 explains the general mechanism that can force graph solutions, while Proposition 12.19 records the additional equal-marginal symmetry used by co-motion maps. For the Coulomb cost, however, the singular repulsion and permutation symmetry make the structure delicate: co-motion maps are optimal in special geometries, but they are not universally optimal, and counterexamples are known [151, 59]. Thus the DFT problem is both a central application and a warning that multi-marginal OT is richer than a naive deterministic matching problem.

Figure 12.4 shows the same phenomenon in a deliberately small one-dimensional model. Three identical two-bump marginals, with unequal component masses and widths, interact through a softened pairwise Coulomb cost. The displayed object is the two-particle marginal of the entropic three-marginal optimizer, with its one-dimensional marginals glued to the top and right sides of the plan. As ε

decreases, mass is pushed away from the collision diagonal $x_1 = x_2$, revealing the repulsive geometry that underlies the strictly-correlated-electrons model.

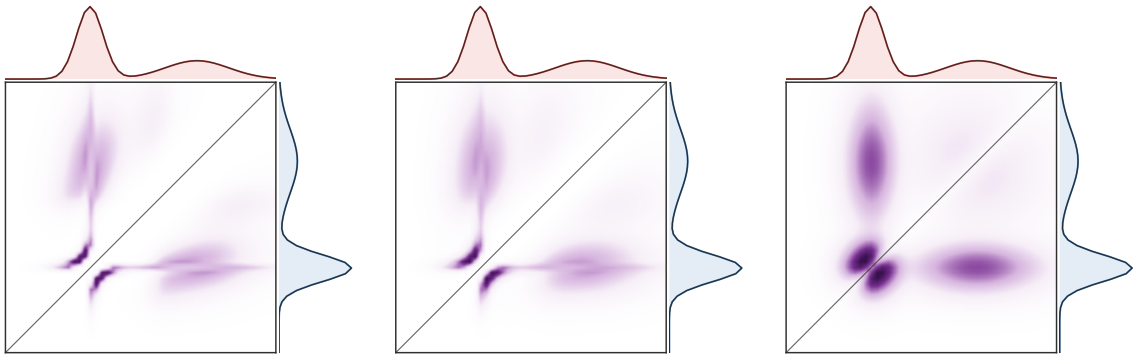


Figure 12.4: Entropic three-marginal Coulomb transport in one dimension. The three marginals are the same asymmetric two-Gaussian mixture, and the pairwise cost is $((x_r - x_s)^2 + \delta^2)^{-1/2}$. Each panel shows the (X_1, X_2) marginal of the tensor Sinkhorn solution; the red top strip and blue right strip are the prescribed marginals, while the thin dark curves are the marginals obtained by summing the displayed pairwise plan. Small regularization concentrates the plan away from the collision diagonal, while larger regularization blurs the repulsive arrangement toward the independent reference.

Multi-marginal formulation of barycenters. Wasserstein barycenters are the central example. For the squared Euclidean cost, one can introduce a latent barycenter point and eliminate it explicitly, leading to the multi-marginal cost

$$c_{\mathrm{bar}}(x_1, \dots, x_S) = \min_{x \in \mathbb{R}^d} \sum_{s=1}^{S} \lambda_s \|x - x_s\|^2.$$

Proposition 12.20 (Multi-marginal formula for quadratic barycenters)**.** *Let $\beta_1, \dots, \beta_S \in \mathcal{P}_2(\mathbb{R}^d)$ and $\lambda \in \Sigma_S$. Define*

$$B(x_1, \dots, x_S) = \sum_{s=1}^{S} \lambda_s x_s, \qquad c_{\mathrm{bar}}(x_1, \dots, x_S) = \min_x \sum_s \lambda_s \|x - x_s\|^2.$$

If $\pi^\star$ solves the multi-marginal OT problem with marginals $(\beta_s)_s$ and cost c_{bar}, then $\alpha^\star = B_\sharp \pi^\star$ is a Wasserstein barycenter. Conversely, every barycenter is obtained this way from an optimal multi-marginal plan.

Proof. For any candidate barycenter α and couplings $\pi_s \in \mathcal{U}(\alpha, \beta_s)$, glue the couplings along their common α marginal to obtain a joint law of $(X, Y_1, \dots, Y_S)$. Conditioning on $(Y_s)_s$ and minimizing over X gives

$$\sum_s \lambda_s \mathbb{E}\|X - Y_s\|^2 \geqslant \mathbb{E} \min_x \sum_s \lambda_s \|x - Y_s\|^2 = \mathbb{E} c_{\mathrm{bar}}(Y_1, \dots, Y_S).$$

Taking the infimum over the couplings gives that the barycenter value is at least the multi-marginal value. Conversely, from an optimal multi-marginal plan $\pi^\star$, set $X = B(Y_1, \dots, Y_S)$. The couplings between X and each Y_s are feasible for the barycenter problem and attain exactly the multi-marginal cost, proving equality and the formula. If $\alpha^\star$ is any barycenter, choose optimal couplings between $\alpha^\star$ and each β_s and glue them along the common $\alpha^\star$ marginal. Since the barycenter and multi-marginal values are equal, the conditional minimization inequality above must be an equality. Thus $X = B(Y_1, \dots, Y_S)$ almost surely for the induced optimal multi-marginal plan, and $\alpha^\star = B_\sharp \pi^\star$. □

Corollary 12.21 (Sparse discrete barycenters)**.** *Let*

$$\beta_s = \sum_{i_s=1}^{n_s} b_{s,i_s} \delta_{x_{s,i_s}}, \qquad s = 1, \dots, S,$$

be discrete input measures. There exists a quadratic Wasserstein barycenter $\alpha^\star$ supported on weighted averages $\sum_s \lambda_s x_{s,i_s}$ of one support point from each input and satisfying

$$\#\operatorname{supp}(\alpha^\star) \leqslant \sum_{s=1}^{S} n_s - S + 1.$$

Proof. Write a discrete multi-marginal plan as a nonnegative tensor $\mathrm{P} = (\mathrm{P}_{i_1,\dots,i_S})$ and let $\mathcal{A}_{\mathrm{marg}}$ collect its S marginals:

$$\big(\mathcal{A}_{\mathrm{marg}}\mathrm{P}\big)_{s,i_s} = \sum_{(i_r)_{r\neq s}} \mathrm{P}_{i_1,\dots,i_S}.$$

After vectorizing P, Proposition 3.7 applies to the multi-marginal linear program. Its constraint operator has rank

$$\operatorname{rank}(\mathcal{A}_{\mathrm{marg}}) = \sum_{s=1}^{S} n_s - S + 1.$$

Indeed, a family of vectors $u_s \in \mathbb{R}^{n_s}$ belongs to the annihilator of its image precisely when

$$\sum_{s=1}^{S} u_{s,i_s} = 0 \qquad \text{for every } (i_1,\dots,i_S).$$

Varying one index at a time shows that every u_s is constant, say $u_s = c_s \mathbb{1}_{n_s}$, and the remaining condition is $\sum_s c_s = 0$. The annihilator therefore has dimension $S-1$, which proves the rank formula.

The rank-controlled sparsity result therefore supplies an optimal multi-marginal tensor $\mathrm{P}^\star$ with at most $\sum_s n_s - S + 1$ positive entries. Proposition 12.20 gives the barycenter $\alpha^\star = B_\sharp \mathrm{P}^\star$. Each positive tensor entry produces at most one atom under B, and collisions can only reduce the support size. This proves the claim. □

This linear support bound, rather than the cardinality of the full product grid, is the standard sparsity estimate for discrete Wasserstein barycenters [14].

Entropic regularization of multi-marginal OT. As in the two-marginal case, adding an entropic penalty with respect to the product measure $\alpha_1 \otimes \cdots \otimes \alpha_S$ leads to scaling algorithms:

$$\inf_{\pi \in \mathcal{U}(\alpha_1,\dots,\alpha_S)} \int c\,\mathrm{d}\pi + \varepsilon\,\mathrm{KL}(\pi | \alpha_1 \otimes \cdots \otimes \alpha_S).$$

The optimizer has the generalized Gibbs form

$$\mathrm{d}\pi^\star(x_1,\dots,x_S) = \exp\!\left(\frac{\sum_s f_s(x_s) - c(x_1,\dots,x_S)}{\varepsilon}\right) \prod_s \mathrm{d}\alpha_s(x_s),$$

and generalized Sinkhorn iterations alternately update one potential f_s so that the s-th marginal is correct. This formula is direct but, without additional structure, it is mostly a conceptual baseline. In the discrete case, even storing the Gibbs tensor or the coupling requires $\prod_s n_s$ entries, and the unstructured multi-marginal Sinkhorn complexity inherits this exponential dependence on the number of marginals [341].

Treewidth and graphical structure. The important exception is when the cost factors over a sparse interaction graph. Let $G = (V, E)$ be a finite undirected graph with $V = \{1,\dots,S\}$ and suppose, for simplicity, that

$$c(x_1,\dots,x_S) = \sum_{(r,s)\in E} c_{r,s}(x_r, x_s). \tag{12.30}$$

The relevant complexity parameter is not merely the number of edges, but the largest intermediate interaction created when variables are summed out.

Definition 12.22 (Tree decomposition and treewidth)**.** A *tree decomposition* of a finite graph $G = (V, E)$ is a tree $\mathcal{T} = (Q, F)$ together with bags $B_q \subseteq V$, $q \in Q$, such that

1. $\bigcup_{q \in Q} B_q = V$;
2. for every edge $(r, s) \in E$, some bag contains both r and s;
3. for every vertex $s \in V$, the nodes $\{q \in Q \,;\, s \in B_q\}$ form a connected subtree of $\mathcal{T}$.

The width of this decomposition is $\max_{q \in Q} |B_q| - 1$, and the *treewidth* $\mathrm{tw}(G)$ is the minimum width over all tree decompositions.

The last condition is the *running-intersection property*. A tree decomposition equipped with factors assigned to its bags is also called a *junction tree*. Equivalently, choose an order in which to eliminate the vertices of G. Just before eliminating a vertex, connect all its remaining neighbors, thereby adding *fill-in edges*; the induced width of the order is the largest number of remaining neighbors encountered. The treewidth is the minimum induced width over all elimination orders, and the corresponding bags consist of each eliminated vertex together with those neighbors. This equivalent viewpoint explains why treewidth controls exact summation.

The same construction applies when the cost contains higher-order factors $c_A(x_A)$ indexed by subsets $A \subseteq V$: one first forms the primal interaction graph by connecting every pair of variables that occurs in a common factor. Its treewidth controls the resulting contractions.

Example 12.23 (Treewidth of familiar interaction graphs). A tree, including a chain or a star, has treewidth one. For the chain $1 - 2 - \cdots - S$, the edge bags $B_q = \{q, q+1\}$ form a width-one decomposition. A cycle C_S, $S \geqslant 3$, has treewidth two: fixing vertex 1, the bags $\{1, q, q+1\}$ for $q = 2, \dots, S-1$ form a width-two decomposition. The complete graph K_S has treewidth $S - 1$, so it offers no asymptotic time reduction over enumerating all S indices. The star appearing when a barycenter variable is retained as a hub also has treewidth one; in that graphical formulation, the leaf marginals are prescribed while the hub marginal is induced by the optimizer [211].

Junction-tree contractions inside Sinkhorn. The treewidth reduction is realized by replacing every full tensor contraction in Sinkhorn by exact sum-product messages. In the discrete setting, define the edge Gibbs matrices and the current unary factors

$$\mathrm{K}^{r,s}_{i_r,i_s} = \exp\!\left(-\frac{\mathrm{C}^{r,s}_{i_r,i_s}}{\varepsilon}\right), \qquad h_s(i_s) = (\mathrm{a}_s)_{i_s} (u_s)_{i_s}.$$

Under (12.30), the current scaled coupling is represented implicitly as

$$\mathrm{P}_{i_1,\dots,i_S} = \prod_{s \in V} h_s(i_s) \prod_{(r,s) \in E} \mathrm{K}^{r,s}_{i_r,i_s}. \tag{12.31}$$

When G is a tree, let $N_G(r)$ denote the neighbors of r. The directed sum-product messages satisfy

$$m_{r \to s}(i_s) = \sum_{i_r = 1}^{n_r} \mathrm{K}^{r,s}_{i_r,i_s} h_r(i_r) \prod_{q \in N_G(r) \setminus \{s\}} m_{q \to r}(i_r). \tag{12.32}$$

Cutting the edge (r, s) separates the tree into two components: $m_{r \to s}(i_s)$ is precisely the total contribution of the component containing r, conditional on the boundary state i_s. Conditioning first on i_r gives (12.32). A leaf-to-root pass followed by a root-to-leaf pass computes every directed message and hence every current marginal,

$$(\widehat{\mathrm{a}}_s)_{i_s} = h_s(i_s) \prod_{r \in N_G(s)} m_{r \to s}(i_s). \tag{12.33}$$

For the block s currently selected by the cyclic scaling algorithm, the exact coordinate update is then

$$u_s \leftarrow u_s \odot \frac{\mathrm{a}_s}{\widehat{\mathrm{a}}_s},$$

with componentwise products and quotients. This changes only the unary factor h_s. Updating all blocks at once would instead define a Jacobi scheme. Thus the expensive denominator of one generalized Sinkhorn block update is evaluated by messages rather than by enumerating all multi-indices.

For a general tree decomposition, choose one host bag for each unary factor, assign each edge factor in (12.31) to a bag containing both endpoints, and denote the product assigned to bag B_q by $\psi_q(i_{B_q})$. For adjacent bags $q, q' \in Q$, set $\mathcal{S}_{q,q'} = B_q \cap B_{q'}$. Junction-tree messages take the form

$$M_{q\to q'}(i_{\mathcal{S}_{q,q'}}) = \sum_{i_{B_q\setminus\mathcal{S}_{q,q'}}} \psi_q(i_{B_q}) \prod_{\ell\in N_{\mathcal{T}}(q)\setminus\{q'\}} M_{\ell\to q}(i_{\mathcal{S}_{\ell,q}}). \tag{12.34}$$

After a collect-distribute pass, the calibrated bag belief is

$$\mathfrak{b}_q(i_{B_q}) = \psi_q(i_{B_q}) \prod_{\ell\in N_{\mathcal{T}}(q)} M_{\ell\to q}(i_{\mathcal{S}_{\ell,q}}). \tag{12.35}$$

For any $s \in B_q$, summing $\mathfrak{b}_q$ over $B_q \setminus \{s\}$ gives $\widehat{\mathrm{a}}_s$; the running-intersection property ensures that every bag containing s gives the same result.

Redundant adjacent bags may be contracted, so one can assume that the decomposition is reduced. If its width is w, every bag contains at most $w+1$ variables and every separator contains at most w. Consequently, if $n_s \leqslant n$, one collect-distribute inference pass for fixed scalings costs

$$O\Big(|Q|\, n^{w+1}\Big) \quad \text{time}, \qquad O(|Q|\, n^{w}) \quad \text{message memory}.$$

More precisely, the time is bounded by

$$O\left(\sum_{q\in Q}\big(1+\deg_{\mathcal{T}}(q)\big)\prod_{s\in B_q} n_s\right),$$

and the stored messages occupy

$$O\left(\sum_{(q,q')\in F}\prod_{s\in\mathcal{S}_{q,q'}} n_s\right)$$

beyond the original factors and scaling vectors. The message-memory bound assumes that bag contractions are streamed; materializing every dense belief (12.35) instead requires $O(|Q|n^{w+1})$ working memory.

These are fixed-scaling inference bounds, not the cost of an entire Sinkhorn solve. They also lead to a sharper block implementation. Assign each constrained marginal s to a host bag $q(s)$. After updating u_s, only the messages along the unique path from $q(s)$ to the host bag of the next updated marginal must be refreshed. A path of length ℓ costs $O(\ell n^{w+1})$, hence at most $O(\mathrm{diam}(\mathcal{T})n^{w+1})$ per block update; this is the junction-tree iterative-scaling mechanism analyzed in [263, 211]. For a tree interaction graph, direct edge messages specialize to $O(\sum_{(r,s)\in E} n_r n_s)$ for a full pass, or $O(Sn^2)$ with equal support sizes. A cycle admits $O(Sn^3)$ full junction-tree passes, while the complete graph still costs $O(n^S)$ in time.

The structured implementation therefore never forms K or P as full tensors. It stores the original local factors and separator messages, evaluates bag products on the fly, replaces the explicit contraction in Algorithm 12.2 by (12.33) or (12.34), and returns the coupling through its factors and scalings. Materializing every entry of P would itself require $\prod_s n_s$ operations, so the saving is meaningful when only marginals, costs, samples, or other low-order statistics are needed. For small ε, the same recursions are evaluated in the log domain with log-sum-exp operations. This viewpoint connects entropic multi-marginal OT with exact inference in probabilistic graphical models and with Schrödinger bridge computations [262, 263, 211, 8].

A representative fluid-mechanics example is the time discretization of Brenier's generalized incompressible Euler problem: the kinetic action couples neighboring time slices and the periodic incompressibility constraint closes the chain into a cycle, hence a graph of treewidth two. Entropic Bregman/Sinkhorn schemes exploit this circular structure through low-order contractions [37, 39].

Practical barycenter solvers therefore exploit separability of the cost, low-rank structure, convolutional kernels, or a fixed barycenter support.

In finite dimension, the direct generalized scaling scheme is the tensor version of Sinkhorn. The explicit algorithm below assumes positive marginals and a finite cost tensor; forbidden configurations with infinite cost can instead be removed from the tensor support, provided the remaining marginal constraints are feasible.

Algorithm 12.2: Multi-marginal Sinkhorn

Input: Positive marginals $a_s \in \Sigma_{n_s}$, finite tensor cost C, regularization $\varepsilon > 0$, tolerance tol, iteration budget $L \geqslant 1$.
Output: Multi-marginal entropic coupling tensor P.
Build $K_{i_1,\dots,i_S} = \exp\left(-\frac{C_{i_1,\dots,i_S}}{\varepsilon}\right) \prod_{s=1}^{S} (a_s)_{i_s}$.
Initialize: Set $u_s = \mathbb{1}_{n_s}$ for all s, residual $\eta = +\infty$, and $\ell = 0$.
While $\eta >$ tol and $\ell < L$ **do**:
  Set $\ell \leftarrow \ell + 1$.
  For $s = 1, \dots, S$ **do**:
    Use the newest available values of the other scaling vectors.
    $(u_s)_i \leftarrow \frac{(a_s)_i}{\sum_{i_1,\dots,i_{s-1},i_{s+1},\dots,i_S} K_{i_1,\dots,i_{s-1},i,i_{s+1},\dots,i_S} \prod_{r \neq s} (u_r)_{i_r}}$.
  Set $P_{i_1,\dots,i_S} = K_{i_1,\dots,i_S} \prod_s (u_s)_{i_s}$.
  Set $\eta = \max_s \|(\mathrm{proj}_s)_\sharp P - a_s\|_1$.
Return P.

12.3 Low-Rank Optimal Transport

Low-rank OT reduces the size of a transport plan by forcing the coupling itself to pass through a small latent measure. This is useful when the mass exchange is expected to be organized by a few hidden clusters or prototypes, and it is distinct from approximating the Sinkhorn kernel by a low-rank matrix. The idea was introduced statistically through factored couplings by Forrow, Hütter, Nitzan, Rigollet, Schiebinger and Weed [218], and developed algorithmically for arbitrary costs by Scetbon, Cuturi and Peyré [474].

Definition 12.24 (Low-rank factored couplings)**.** Let $a \in \Sigma_n$, $b \in \Sigma_m$ and let $r \geqslant 1$. A rank-r factored coupling is a triple (Q, R, g) such that

$$Q \in \mathbb{R}_+^{n \times r}, \qquad R \in \mathbb{R}_+^{m \times r}, \qquad g \in \Sigma_r,$$

with

$$Q\mathbb{1}_r = a, \qquad R\mathbb{1}_r = b, \qquad Q^\top \mathbb{1}_n = R^\top \mathbb{1}_m = g.$$

It induces the coupling

$$P(Q, R, g) = Q\,\mathrm{diag}(g)^{-1} R^\top, \qquad P_{i,j} = \sum_{k=1}^{r} \frac{Q_{i,k} R_{j,k}}{g_k}, \tag{12.36}$$

where columns with $g_k = 0$ are discarded before applying the formula.

The latent interpretation is immediate. The vector g is the law of an intermediate variable $Z \in [\![1, r]\!]$; Q is a coupling of the source index X with Z; R is a coupling of the target index Y with the same Z. Formula (12.36) is the law of (X, Y) obtained by sampling $Z \sim g$, then sampling X and Y conditionally independently given Z. Equivalently, OT is replaced by a succession of two transports through an abstract intermediate measure $\eta = \sum_{k=1}^{r} g_k \delta_{z_k}$ on an r-point space. The locations z_k only label the latent atoms and do not enter the original cost.

Proposition 12.25 (Factored couplings and nonnegative rank)**.** *For every feasible triple (Q, R, g) in Definition 12.24, the matrix $P(Q, R, g)$ belongs to $U(a, b)$ and has nonnegative rank at most r. Conversely, every coupling $P \in U(a, b)$ with nonnegative rank at most r admits a representation of the form (12.36), after possibly deleting zero latent components.*

Proof. The marginal constraints follow by direct summation:

$$\mathrm{P}\mathbb{1}_m = \mathrm{Q}\,\mathrm{diag}(g)^{-1}\mathrm{R}^\top\mathbb{1}_m = \mathrm{Q}\,\mathrm{diag}(g)^{-1}g = \mathrm{Q}\mathbb{1}_r = \mathrm{a},$$

and similarly $\mathrm{P}^\top\mathbb{1}_n = \mathrm{b}$. Since P is a sum of r nonnegative rank-one matrices $\mathrm{Q}_{:,k}\mathrm{R}_{:,k}^\top/g_k$, its nonnegative rank is at most r.

Conversely, suppose $\mathrm{P} = UV^\top$ with $U \in \mathbb{R}_+^{n\times r}$ and $V \in \mathbb{R}_+^{m\times r}$. Set $q_k = \sum_i U_{i,k}$ and $s_k = \sum_j V_{j,k}$. Components with $q_k s_k = 0$ do not contribute and can be removed. Define $g_k = q_k s_k$, $\mathrm{Q}_{i,k} = U_{i,k}s_k$ and $\mathrm{R}_{j,k} = V_{j,k}q_k$. Since P has total mass one, $g \in \Sigma_r$. Moreover $\mathrm{Q}\mathbb{1}_r = \mathrm{P}\mathbb{1}_m = \mathrm{a}$, $\mathrm{R}\mathbb{1}_r = \mathrm{P}^\top\mathbb{1}_n = \mathrm{b}$, and both column marginals are equal to g. Finally,

$$\sum_k \frac{\mathrm{Q}_{i,k}\mathrm{R}_{j,k}}{g_k} = \sum_k \frac{U_{i,k}s_k V_{j,k}q_k}{q_k s_k} = \mathrm{P}_{i,j}.$$

□

For a cost matrix $\mathrm{C} \in \mathbb{R}^{n\times m}$, the low-rank constrained OT value is therefore

$$\min_{(\mathrm{Q},\mathrm{R},g)} \sum_{k=1}^{r} \frac{\mathrm{Q}_{:,k}{}^\top \mathrm{C}\mathrm{R}_{:,k}}{g_k} \quad \text{subject to Definition 12.24.}$$

This problem is non-convex. Scetbon, Cuturi and Peyré regularize the joint variables $(\mathrm{Q}, \mathrm{R}, g)$ by the sum of their entropies and optimize them by constrained mirror descent [474]. To isolate the simpler block mechanism used in Figure 12.5, fix here a positive latent law $g \in \Sigma_r$ and optimize only the two sub-couplings. The resulting pedagogical model is

$$\min_{\substack{\mathrm{Q}\mathbb{1}_r=\mathrm{a},\ \mathrm{Q}^\top\mathbb{1}_n=g\\ \mathrm{R}\mathbb{1}_r=\mathrm{b},\ \mathrm{R}^\top\mathbb{1}_m=g}} \sum_{k=1}^{r} \frac{\mathrm{Q}_{:,k}{}^\top \mathrm{C}\mathrm{R}_{:,k}}{g_k} + \varepsilon\mathrm{KL}(\mathrm{Q}|\mathrm{a}\otimes g) + \varepsilon\mathrm{KL}(\mathrm{R}|\mathrm{b}\otimes g), \tag{12.37}$$

where Q, R are nonnegative. For fixed g, this differs only by constants from adding $-\varepsilon(\mathrm{H}(\mathrm{Q}) + \mathrm{H}(\mathrm{R}))$ to the factorized transport cost. Each block subproblem is a strictly convex entropic OT problem with an effective cost, although the joint objective remains non-convex because of its bilinear Q–R term.

Algorithm 12.3: Alternating low-rank Sinkhorn with fixed latent mass

Input: Positive marginals $\mathrm{a} \in \mathrm{int}(\Sigma_n)$ and $\mathrm{b} \in \mathrm{int}(\Sigma_m)$, cost $\mathrm{C} \in \mathbb{R}^{n\times m}$, rank r, positive latent mass $g \in \mathrm{int}(\Sigma_r)$, regularization $\varepsilon > 0$, iteration budget $L \geqslant 1$, tolerance tol.
Output: Factored coupling $\mathrm{P} = \mathrm{Q}\,\mathrm{diag}(g)^{-1}\mathrm{R}^\top$.
Initialize: Set $\mathrm{Q}^{(0)} = \mathrm{a}\otimes g$, $\mathrm{R}^{(0)} = \mathrm{b}\otimes g$ and $J^{(0)} = +\infty$.
For $\ell = 0, \dots, L-1$ **do**:
  Set $A^{(\ell)} = \mathrm{C}\mathrm{R}^{(\ell)}\,\mathrm{diag}(g)^{-1}$.
  Update $\mathrm{Q}^{(\ell+1)}$ by Sinkhorn: $\mathrm{Q}^{(\ell+1)} = \mathrm{argmin}_{\mathrm{Q}\mathbb{1}_r=\mathrm{a},\,\mathrm{Q}^\top\mathbb{1}_n=g}\langle A^{(\ell)}, \mathrm{Q}\rangle + \varepsilon\mathrm{KL}(\mathrm{Q}|\mathrm{a}\otimes g)$.
  Set $B^{(\ell+1)} = \mathrm{C}^\top\mathrm{Q}^{(\ell+1)}\,\mathrm{diag}(g)^{-1}$.
  Update $\mathrm{R}^{(\ell+1)}$ by Sinkhorn: $\mathrm{R}^{(\ell+1)} = \mathrm{argmin}_{\mathrm{R}\mathbb{1}_r=\mathrm{b},\,\mathrm{R}^\top\mathbb{1}_m=g}\langle B^{(\ell+1)}, \mathrm{R}\rangle + \varepsilon\mathrm{KL}(\mathrm{R}|\mathrm{b}\otimes g)$.
  Set $\mathrm{P}^{(\ell+1)} = \mathrm{Q}^{(\ell+1)}\,\mathrm{diag}(g)^{-1}(\mathrm{R}^{(\ell+1)})^\top$.
  Set $J^{(\ell+1)}$ to the value of (12.37) at $(\mathrm{Q}^{(\ell+1)}, \mathrm{R}^{(\ell+1)})$.
  If $\ell \geqslant 1$ and $|J^{(\ell+1)} - J^{(\ell)}| \leqslant \mathrm{tol}\max\{1, |J^{(\ell)}|\}$ **then**:
    Return $\mathrm{P}^{(\ell+1)}$.
Return $\mathrm{P}^{(L)}$.

Each update in Algorithm 12.3 exactly minimizes one factor block while keeping the other fixed. The objective values therefore decrease monotonically. With the positive marginals assumed above, the entropy makes each block minimizer unique and keeps it in the relative interior of its transport polytope. Compactness and the standard convergence argument for exact cyclic block-coordinate descent then imply that every accumulation point is a coordinatewise minimizer, hence a stationary point of the constrained problem. The non-convexity prevents this argument from guaranteeing a globally optimal rank-r coupling.

Figure 12.5 visualizes both the intermediate latent measure and the improvement of the factored coupling as the prescribed rank increases.

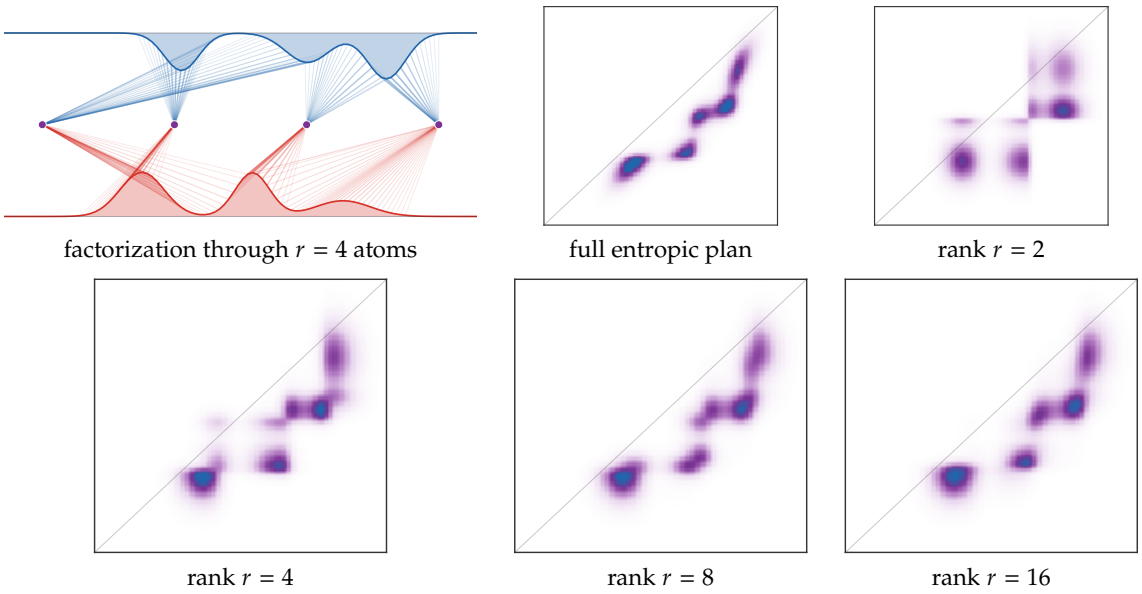


Figure 12.5: Low-rank entropic OT on a one-dimensional Gaussian-mixture example with small regularization. The first panel shows the factorization through an intermediate four-atom measure: red mass couples to violet latent atoms, which then couple to blue mass. The remaining panels compare the full entropic coupling with fixed-latent-mass low-rank couplings of increasing rank. The approximation improves with r, but this is deliberately not a favorable example for low rank: in one dimension with a small entropic parameter, quadratic OT is close to a sparse Monge graph rather than to a genuinely low-rank matrix.

12.4 Capacity-Constrained Optimal Transport

Classical Kantorovich transport only fixes the marginals: if a pair (x, y) is cheap, the optimizer may concentrate as much mass as the marginal constraints allow on this pair. Capacity-constrained OT adds a local congestion rule on the coupling itself. It is useful when edges, facilities or matchings have limited throughput, and it also gives a clean mathematical way to interpolate between a sparse OT plan and the independent product coupling. The systematic study of the continuous problem, including existence and the geometry of active saturated regions, was developed by Korman and McCann [315].

The capacity is most naturally expressed as an upper bound on the density of the plan relative to the product measure.

Definition 12.26 (Capacity-constrained optimal transport)**.** Let $\alpha \in \mathcal{M}_+^1(\mathcal{X})$, $\beta \in \mathcal{M}_+^1(\mathcal{Y})$, let $c : \mathcal{X} \times \mathcal{Y} \to [0, +\infty]$ be measurable, and let $u : \mathcal{X} \times \mathcal{Y} \to [0, +\infty]$ be a measurable capacity. Define

$$\mathcal{L}_c^u(\alpha, \beta) := \inf_{\pi \in \mathcal{U}(\alpha,\beta)} \left\{ \int_{\mathcal{X}\times\mathcal{Y}} c(x, y) \mathrm{d}\pi(x, y) \; ; \; \pi \ll \alpha \otimes \beta, \quad \frac{\mathrm{d}\pi}{\mathrm{d}(\alpha \otimes \beta)}(x, y) \leqslant u(x, y) \quad (\alpha \otimes \beta)\text{-a.e.} \right\}. \tag{12.38}$$

For discrete measures with cost matrix C and capacity matrix $\overline{\mathrm{P}}_{i,j} = u_{i,j} a_i b_j$, the corresponding problem is

$$\min_{\mathrm{P} \in \mathrm{U(a,b)}} \langle \mathrm{C}, \mathrm{P} \rangle \quad \text{subject to} \quad 0 \leqslant \mathrm{P}_{i,j} \leqslant \overline{\mathrm{P}}_{i,j} \quad \forall (i, j). \tag{12.39}$$

By convention, $u \equiv +\infty$ removes both the density bound and the absolute-continuity requirement, so that $\mathcal{L}_c^{+\infty}(\alpha, \beta) = \mathcal{L}_c(\alpha, \beta)$.

The product coupling $\alpha \otimes \beta$ is feasible whenever $u \geqslant 1$. Thus the constraint is not meant to promote the product plan; rather, it prevents the optimizer from using any pair more than the prescribed density ratio. At the opposite extreme, $u = 1$ forces the independent coupling itself, because a density bounded by one and integrating to one must equal one almost everywhere. Values close to one therefore enforce diffuse plans close to this reference.

For discrete measures $\alpha = \sum_i a_i \delta_{x_i}$ and $\beta = \sum_j b_j \delta_{y_j}$, a capacity is an upper matrix $\overline{\mathrm{P}} \in \mathbb{R}_+^{n \times m}$. The density-ratio discretization in Definition 12.26 gives a capped linear program. Feasibility is now a genuine issue: the upper matrix must contain enough mass in every row-column cut to support the prescribed marginals. In particular, the usual transport polytope is recovered when $\overline{\mathrm{P}}_{i,j} = +\infty$, while small capacities select a smaller capped transportation polytope.

The cut condition can be stated exactly. For index sets $I \subset \{1, \dots, n\}$ and $J \subset \{1, \dots, m\}$, write $a(I) = \sum_{i \in I} a_i$, $b(J) = \sum_{j \in J} b_j$, and $\overline{\mathrm{P}}(I, J) = \sum_{i \in I, j \in J} \overline{\mathrm{P}}_{i,j}$.

Proposition 12.27 (Feasibility of a capped transport polytope)**.** *The constraints in* (12.39) *are feasible if and only if*

$$a(I) + b(J) - 1 \leqslant \overline{\mathrm{P}}(I, J) \qquad \text{for every } I \subset \{1, \dots, n\},\ J \subset \{1, \dots, m\}. \tag{12.40}$$

Proof. If P is feasible, then

$$\mathrm{P}(I, J) = a(I) - \mathrm{P}(I, J^c) \geqslant a(I) - b(J^c) = a(I) + b(J) - 1.$$

Since $\mathrm{P}(I, J) \leqslant \overline{\mathrm{P}}(I, J)$, condition (12.40) is necessary.

For sufficiency, build a flow network with capacity a_i from the source to row vertex i, capacity $\overline{\mathrm{P}}_{i,j}$ from row i to column j, and capacity b_j from column j to the sink. A cut containing row set I and column set J^c has capacity $1 - a(I) + \overline{\mathrm{P}}(I, J) + 1 - b(J)$. Condition (12.40) says that every such cut has capacity at least one. The max-flow/min-cut theorem therefore gives a unit flow, and its row-to-column edge values form the required matrix P. □

Entropic smoothing gives a direct Sinkhorn-like algorithm. Assume that the cut condition (12.40) holds. With $\mathrm{K}_{i,j} = a_i b_j e^{-\mathrm{C}_{i,j}/\varepsilon}$, the regularized problem is

$$\min_{\mathrm{P} \in \mathrm{U}(\mathrm{a}, \mathrm{b}),\, 0 \leqslant \mathrm{P} \leqslant \overline{\mathrm{P}}} \langle \mathrm{C}, \mathrm{P} \rangle + \varepsilon \mathrm{KL}(\mathrm{P} | \mathrm{a} \otimes \mathrm{b}). \tag{12.41}$$

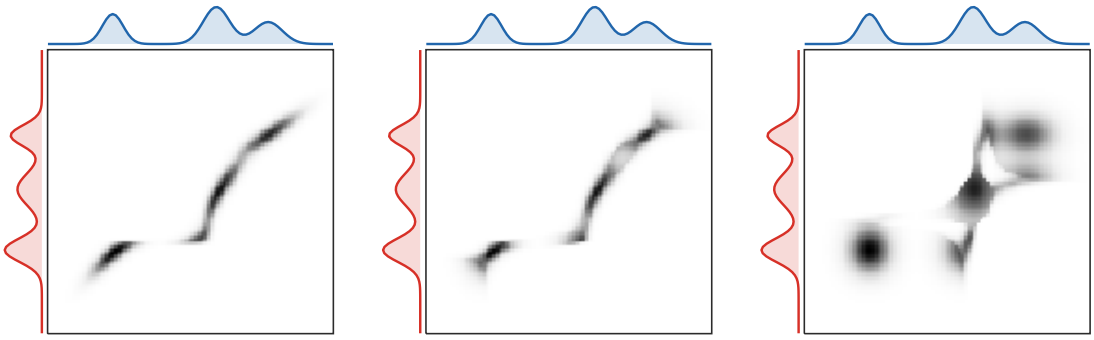
Equivalently, up to additive constants, the objective is $\varepsilon \mathrm{KL}(\mathrm{P}|\mathrm{K})$. The problem is therefore the KL projection of K onto the intersection of three convex sets: the row constraints, the column constraints and the box $\mathrm{P} \leqslant \overline{\mathrm{P}}$. Alternating KL projections with Dykstra correction factors [198, 30], in the same spirit as the Bregman projection formulation of Sinkhorn [37], gives a simple capacity-constrained scaling scheme.

Algorithm 12.4: Capacity-constrained Sinkhorn by KL-Dykstra

Input: Positive marginals $\mathrm{a} \in \Sigma_n$, $\mathrm{b} \in \Sigma_m$, finite cost C, feasible capacity $\overline{\mathrm{P}}$, regularization $\varepsilon > 0$, tolerance tol, iteration budget $L \geqslant 1$.
Output: Entropic capped coupling P.
Restrict all arrays to the active edge set $E = \{(i, j) : \overline{\mathrm{P}}_{i,j} > 0\}$; entries outside E remain zero.
Convention: All products and divisions below are entrywise on E.
Set $\mathrm{K}_{i,j} = a_i b_j e^{-\mathrm{C}_{i,j}/\varepsilon}$ for $(i, j) \in E$.
Initialize: Set $\mathrm{P} = \mathrm{K}$, correction matrices $\mathrm{R}_1 = \mathrm{R}_2 = \mathrm{R}_3 = \mathbb{1}_{n \times m}$, residual $r = +\infty$, and $\ell = 0$.
While $r >$ tol and $\ell < L$ **do**:
  Set $\ell \leftarrow \ell + 1$ and $\mathrm{P}_{\mathrm{old}} = \mathrm{P}$.
  Set $Z = \mathrm{P} \odot \mathrm{R}_1$.
  Project rows: $\mathrm{P} = \operatorname{diag}(\mathrm{a} / (Z \mathbb{1}_m))\, Z$.
  Update $\mathrm{R}_1 = Z \oslash \mathrm{P}$.
  Set $Z = \mathrm{P} \odot \mathrm{R}_2$.
  Project columns: $\mathrm{P} = Z \operatorname{diag}(\mathrm{b} / (Z^\top \mathbb{1}_n))$.
  Update $\mathrm{R}_2 = Z \oslash \mathrm{P}$.
  Set $Z = \mathrm{P} \odot \mathrm{R}_3$.
  Project capacity: $\mathrm{P} = \min\{Z, \overline{\mathrm{P}}\}$ entrywise.
  Update $\mathrm{R}_3 = Z \oslash \mathrm{P}$.
  Set r to the maximum of the three feasibility defects and the fixed-point defect: $r = \max\{\|\mathrm{P}\mathbb{1}_m - \mathrm{a}\|_\infty, \|\mathrm{P}^\top \mathbb{1}_n - \mathrm{b}\|_\infty, \|(\mathrm{P} - \overline{\mathrm{P}})_+\|_\infty, \|\mathrm{P} - \mathrm{P}_{\mathrm{old}}\|_1\}$.
Return P.

On the active support, all iterates and correction factors are positive, so every division in Algorithm 12.4 is well defined. KL-Dykstra convergence on finite-dimensional affine and box constraints then shows that, when the capped polytope is nonempty, the iterates converge to the unique minimizer of (12.41). Deleting zero-capacity edges before the iteration is essential: writing the correction step directly with a zero cap would create undefined products of zero and infinite correction factors.

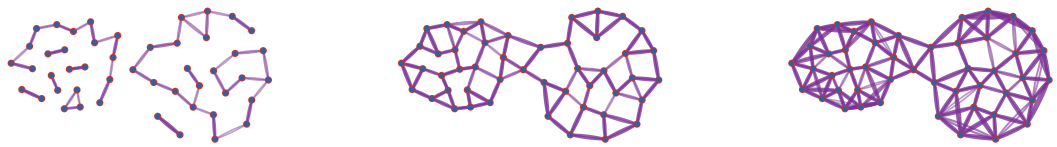
Figure 12.6 shows how lowering the entrywise cap progressively spreads a one-dimensional coupling while preserving both prescribed marginals.



unconstrained, $u = +\infty$ medium capacity, $u = 10$ small capacity, $u = 2.6$

Figure 12.6: Capacity-constrained entropic OT between two one-dimensional Gaussian-mixture histograms. Each matrix solves (12.41) with the same marginals and the same squared-distance cost. The left panel is the unconstrained entropic OT solution, corresponding to the density-ratio cap $\overline{\mathrm{P}}_{i,j} = +\infty$. The other panels impose $\overline{\mathrm{P}}_{i,j} = u a_i b_j$. The red side strip is the source marginal and the blue top strip is the target marginal. As the capacity decreases, more entries saturate and the coupling is forced to spread.

For the empirical self-coupling in Figure 12.7, the cap is chosen to prescribe a minimum number of outgoing connections per source point. With uniform weights $a_i = 1/n$, imposing $\mathrm{P}_{i,j} \leqslant 1/(qn)$ is equivalent to the conditional bound $\mathrm{P}_{i,j}/a_i \leqslant 1/q$. Since each row has total mass $1/n$, this forces each source row to use at least q target atoms, up to the small extra spreading introduced by entropic smoothing.



one connection, $q = 1$ three connections, $q = 3$ five connections, $q = 5$

Figure 12.7: Capacity-constrained local self-couplings on a two-dimensional empirical Gaussian mixture. The source and target are the same semi-regular uniform empirical measure on n atoms, but the diagonal is removed to avoid the trivial identity plan and reveal the local graph selected by the capped problem. Each panel solves (12.41) with the off-diagonal cap $\overline{\mathrm{P}}_{i,j} = 1/(qn)$, equivalently $\mathrm{P}_{i,j}/a_i \leqslant 1/q$ since $a_i = 1/n$. The three panels therefore impose at least one, three and five outgoing connections per source atom; increasing q forces the local matching to spread over more neighbors.

12.5 Metric Learning and Inverse OT

Metric learning differentiates a forward transport loss through a parameterized cost, whereas inverse OT starts from an observed plan and asks which cost makes it optimal. The first viewpoint is often bilevel; the second admits a direct convex formulation for affine cost families, provided the intrinsic cost invariances are removed.

12.5.1 Differentiating OT losses

This subsection records the sensitivity formulas used when the cost or the marginals are learned from data.

First variations of OT values. Inverse OT and metric learning repeatedly differentiate a forward OT value with respect to the input law and to the ground cost. The two resulting objects are precisely the two certificates of optimality: a Kantorovich potential for perturbations of the marginal and an optimal coupling for perturbations of the cost. The main caveat is non-uniqueness. In the unregularized case, the correct objects are one-sided directional derivatives, or equivalently subgradients in the measure variable and supergradients in the cost variable. Entropic regularization selects a unique plan and, for positive finite histograms, gives ordinary derivatives on the simplex interiors.

Proposition 12.28 (First variations of unregularized OT)**.** *Let $\alpha \in \mathcal{P}(\mathcal{X})$ and $\beta \in \mathcal{P}(\mathcal{Y})$, where $\mathcal{X}, \mathcal{Y}$ are compact metric spaces, and let $c \in \mathcal{C}(\mathcal{X} \times \mathcal{Y})$. Define*

$$\mathcal{V}_c(\alpha, \beta) \coloneqq \inf_{\pi \in \mathcal{U}(\alpha,\beta)} \int_{\mathcal{X}\times\mathcal{Y}} c(x,y)\mathrm{d}\pi(x,y).$$

Let $\mathcal{O}_c(\alpha,\beta)$ be the set of optimal couplings and let $\mathcal{D}_c(\alpha,\beta)$ be the set of optimal dual potentials (f,g) with $f(x)+g(y) \leqslant c(x,y)$. If χ is a signed measure with $\chi(\mathcal{X}) = 0$ and $\alpha_t = \alpha + t\chi$ is a probability measure for $0 \leqslant t \leqslant t_0$, then

$$\left.\frac{\mathrm{d}}{\mathrm{d}t}\right|_{t=0^+} \mathcal{V}_c(\alpha_t, \beta) = \sup_{(f,g)\in\mathcal{D}_c(\alpha,\beta)} \int_{\mathcal{X}} f(x)\mathrm{d}\chi(x).$$

If $h \in \mathcal{C}(\mathcal{X}\times\mathcal{Y})$ and $c_t = c + th$, then

$$\left.\frac{\mathrm{d}}{\mathrm{d}t}\right|_{t=0^+} \mathcal{V}_{c_t}(\alpha,\beta) = \inf_{\pi\in\mathcal{O}_c(\alpha,\beta)} \int_{\mathcal{X}\times\mathcal{Y}} h(x,y)\mathrm{d}\pi(x,y).$$

In particular, if the normalized optimal potential $f^\star$ and the optimal plan $\pi^\star$ are unique, then

$$\frac{\delta\mathcal{V}_c}{\delta\alpha} = f^\star, \qquad \frac{\delta\mathcal{V}_c}{\delta c} = \pi^\star.$$

The second identity means that the first variation with respect to the function c is represented by the optimal measure $\pi^\star$ on $\mathcal{X}\times\mathcal{Y}$.

Proof. Kantorovich duality writes

$$\mathcal{V}_c(\alpha,\beta) = \sup_{f\oplus g\leqslant c} \int f\mathrm{d}\alpha + \int g\mathrm{d}\beta.$$

This is a supremum of affine functions of α, so Danskin's theorem gives the one-sided directional derivative as the supremum over active maximizers, namely the optimal dual potentials. The condition $\chi(\mathcal{X}) = 0$ makes the formula independent of the additive gauge $(f,g) \mapsto (f+\lambda, g-\lambda)$.

For the cost variable,

$$\mathcal{V}_{c_t}(\alpha,\beta) = \inf_{\pi\in\mathcal{U}(\alpha,\beta)} \int c\mathrm{d}\pi + t\int h\mathrm{d}\pi$$

is an infimum of affine functions of t. Danskin's theorem for a minimum gives the right directional derivative as the infimum of $\int h\mathrm{d}\pi$ over the active minimizers. If the active dual potential or coupling is unique, the corresponding directional derivative is linear in the perturbation, which is the displayed first variation. □

In the discrete case, this proposition says that any optimal dual vector $f^\star$ is a subgradient with respect to the source weights a, while any optimal plan $\mathrm{P}^\star$ is a supergradient with respect to the cost matrix C, because the value is concave in C:

$$f^\star \in \partial_{\mathrm{a}} \mathrm{L}_{\mathrm{C}}(\mathrm{a}, \mathrm{b}), \qquad \mathrm{P}^\star \in \partial_{\mathrm{C}}^{\sup} \mathrm{L}_{\mathrm{C}}(\mathrm{a}, \mathrm{b}).$$

Here $\partial_{\mathrm{C}}^{\sup}$ denotes the superdifferential of the concave map $\mathrm{C} \mapsto \mathrm{L}_{\mathrm{C}}(\mathrm{a}, \mathrm{b})$. When the corresponding objects are unique, these inclusions become the gradients $\nabla_{\mathrm{a}} \mathrm{L}_{\mathrm{C}}(\mathrm{a}, \mathrm{b}) = f^\star$ on the tangent space $\{\langle \mathbb{1}, \chi \rangle = 0\}$ and $\nabla_{\mathrm{C}} \mathrm{L}_{\mathrm{C}}(\mathrm{a}, \mathrm{b}) = \mathrm{P}^\star$. Without uniqueness, the exact directional derivative with respect to C in a direction $\Delta\mathrm{C}$ is the minimum of $\langle \Delta\mathrm{C}, \mathrm{P} \rangle$ over all optimal plans.

Proposition 12.29 (First variations of entropic OT)**.** *Under the compact-space and continuous-cost assumptions of Proposition 12.28, let $\varepsilon > 0$ and define the KL-normalized entropic value*

$$\mathcal{V}_{c,\varepsilon}(\alpha, \beta) \coloneqq \inf_{\pi \in \mathcal{U}(\alpha,\beta)} \int c \mathrm{d}\pi + \varepsilon \, \mathrm{KL}(\pi | \alpha \otimes \beta).$$

Choose optimal entropic potentials $(f_\varepsilon, g_\varepsilon)$ in soft-transform form, so that the following density has marginals α and β:

$$\mathrm{d}\pi_\varepsilon(x, y) = \exp\left(\frac{f_\varepsilon(x) + g_\varepsilon(y) - c(x, y)}{\varepsilon}\right) \mathrm{d}\alpha(x) \mathrm{d}\beta(y).$$

This defines the unique optimal coupling. For the same perturbations $\alpha_t = \alpha + t\chi$ and $c_t = c + th$ as in Proposition 12.28,

$$\frac{\mathrm{d}}{\mathrm{d}t}\bigg|_{t=0^+} \mathcal{V}_{c,\varepsilon}(\alpha_t, \beta) = \int f_\varepsilon \mathrm{d}\chi, \qquad \frac{\mathrm{d}}{\mathrm{d}t}\bigg|_{t=0^+} \mathcal{V}_{c_t,\varepsilon}(\alpha, \beta) = \int h \mathrm{d}\pi_\varepsilon.$$

Equivalently,

$$\frac{\delta \mathcal{V}_{c,\varepsilon}}{\delta \alpha} = f_\varepsilon, \qquad \frac{\delta \mathcal{V}_{c,\varepsilon}}{\delta c} = \pi_\varepsilon.$$

In finite dimension, for positive histograms, these are ordinary derivatives on the relative interior of the simplices.

Proof. The cost derivative follows directly from the primal envelope theorem, because the entropic optimizer is unique. For the measure derivative, use the dual formula (8.23). At an optimal pair, the soft-transform equations imply the row normalization

$$\int_{\mathcal{Y}} \exp\left(\frac{f_\varepsilon(x) + g_\varepsilon(y) - c(x, y)}{\varepsilon}\right) \mathrm{d}\beta(y) = 1 \qquad \text{for all } x.$$

Differentiating the dual objective with respect to α at fixed optimal potentials gives

$$\int f_\varepsilon \mathrm{d}\chi - \varepsilon \int_{\mathcal{X}\times\mathcal{Y}} \left(e^{(f_\varepsilon(x)+g_\varepsilon(y)-c(x,y))/\varepsilon} - 1\right) \mathrm{d}\chi(x) \mathrm{d}\beta(y).$$

The second term vanishes by the previous normalization identity, leaving $\int f_\varepsilon \mathrm{d}\chi$. The gauge ambiguity of $(f_\varepsilon, g_\varepsilon)$ again disappears because $\chi(\mathcal{X}) = 0$. □

For a finite-dimensional parametrization c_θ or C_θ, the entropic formula gives the backpropagation rule

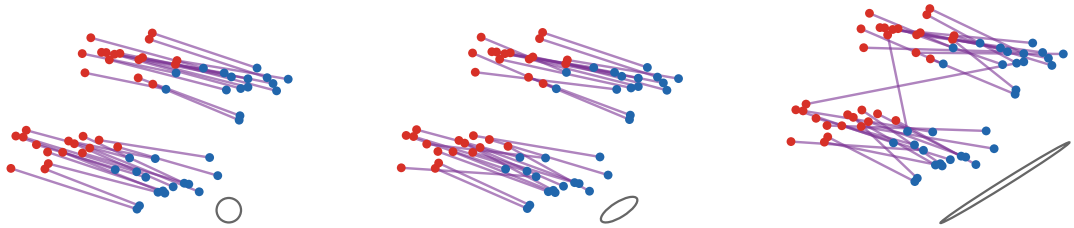
$$\partial_\theta \mathcal{V}_{c_\theta,\varepsilon} = \int \partial_\theta c_\theta(x, y) \mathrm{d}\pi_\varepsilon(x, y),$$

where π_ε is the entropic optimizer. For the unregularized value $\mathcal{V}_{c_\theta}$, uniqueness of the optimal plan $\pi^\star$ gives $\partial_\theta \mathcal{V}_{c_\theta} = \int \partial_\theta c_\theta \mathrm{d}\pi^\star$. Without uniqueness, the directional derivative in a parameter direction $\dot\theta$ is obtained by minimizing $\int \dot\theta \cdot \partial_\theta c_\theta \mathrm{d}\pi$ over the optimal face, while any selected optimal plan gives a valid supergradient with respect to the cost. This is the calculus behind ground-metric learning, which was explicitly studied in [167] and connects to the broader metric-learning literature [319, 34]. If one uses the entropy-only discrete convention (8.1) instead of the KL-normalized value, then, for positive source weights,

$$\mathrm{L}_{\mathrm{C}}^{\varepsilon}(\mathrm{a}, \mathrm{b}) = \mathcal{V}_{\mathrm{C},\varepsilon}(\mathrm{a}, \mathrm{b}) - \varepsilon \mathrm{H}(\mathrm{a}) - \varepsilon \mathrm{H}(\mathrm{b}),$$

so its derivative with respect to a is represented on the simplex tangent space by $f_\varepsilon + \varepsilon \log \mathrm{a}$, up to an irrelevant additive constant.

Figure 12.8 gives the geometric counterpart of this differentiation rule: changing an anisotropic quadratic cost changes which transport segments are selected.



Euclidean cost moderate anisotropy strong anisotropy

Figure 12.8: Changing the ground metric changes the optimal coupling. The same red and blue empirical measures are matched with $c_A(x, y) = (x - y)^\top A(x - y)$ for the Euclidean metric and two increasingly anisotropic Mahalanobis metrics. The small gray ellipse shows the unit ball of the metric: directions in which the ellipse is elongated are cheaper, and this deforms the transport segments selected by the OT plan.

12.5.2 Inverse Optimal Transport

Inverse OT asks for a ground cost that explains observed matchings or flows as optimal transport plans. In its most direct form, one observes a plan $\widehat{\pi}$ with marginals (α, β) and seeks a cost c such that $\widehat{\pi}$ is optimal for

$$\inf_{\pi \in \mathcal{U}(\alpha,\beta)} \int c(x, y) \mathrm{d}\pi(x, y).$$

This is ill-posed without structure. Adding $u(x) + v(y)$ to a cost shifts every feasible objective by the same marginal-dependent constant, and multiplying a cost by a positive scalar does not change its minimizers. Moreover, the zero cost rationalizes every feasible plan. An identifiable inverse model must therefore quotient or normalize these gauge and scale freedoms, and a sparse observed plan can still be compatible with a nontrivial cone of normalized costs.

A useful statistical methodology is to measure the suboptimality of the observed plan through a Fenchel–Young loss. Write the score as $s = -c$ and define the convex regularized prediction value

$$G_\varepsilon(s) = \sup_{\pi \in \mathcal{U}(\alpha,\beta)} \int s \mathrm{d}\pi - \varepsilon \, \mathrm{KL}(\pi | \alpha \otimes \beta).$$

The Fenchel–Young loss

$$\mathcal{L}_\varepsilon(c; \widehat{\pi}) = G_\varepsilon(-c) + G_\varepsilon^*(\widehat{\pi}) + \int c \mathrm{d}\widehat{\pi}$$

is nonnegative by Fenchel's inequality and vanishes exactly when $\widehat{\pi} \in \partial G_\varepsilon(-c)$, i.e. when $\widehat{\pi}$ satisfies the regularized optimality conditions for c. Sharpened Fenchel–Young losses for inverse problems over measures and inverse entropic/unbalanced OT are developed in [17]; curvature and identifiability of inverse OT with respect to the cost are studied in [432]. Entropic regularization is important here because it makes the forward map smoother and provides gradients with respect to c, at the price of a bias that must be controlled statistically.

In the discrete unregularized case, this loss reduces to the optimality gap of the observed coupling. For $\widehat{\mathrm{P}} \in \mathrm{U}(\mathrm{a}, \mathrm{b})$ and a cost matrix C, denote it by

$$\mathcal{L}_{\mathrm{iOT}}(\mathrm{C}; \widehat{\mathrm{P}}) = \langle \mathrm{C}, \widehat{\mathrm{P}} \rangle - \min_{\mathrm{P} \in \mathrm{U}(\mathrm{a},\mathrm{b})} \langle \mathrm{C}, \mathrm{P} \rangle.$$

This inverse-OT gap loss is nonnegative and vanishes exactly when $\widehat{\mathrm{P}}$ is optimal for C.

In practice, one restricts the cost to a finite-dimensional model class, often affine:

$$\mathrm{C}_\theta = \sum_{r=1}^{R} \theta_r \mathrm{C}^{(r)}, \qquad \theta \in \Theta,$$

where Θ is convex and the matrices $\mathrm{C}^{(r)}$ encode features, graph distances or a Mahalanobis parameterization. This viewpoint appears in low-rank and sparse inverse OT models [196, 16] and in convex formulations for learning OT costs from observed plans [359, 432].

A minimal finite-dimensional model is obtained by learning a bilinear cost on $\mathbb{R}^d$,

$$c_A(x, y) = \langle Ax,\, y\rangle, \qquad A \in \mathbb{R}^{d\times d}.$$

For empirical measures $\alpha = \frac{1}{n}\sum_i \delta_{x_i}$ and $\beta = \frac{1}{n}\sum_j \delta_{y_j}$, this gives the cost matrix

$$\mathrm{C}(A)_{i,j} = \langle Ax_i,\, y_j\rangle,$$

so both maps $A \mapsto \mathrm{C}(A)$ and $A \mapsto c_A$ are linear. Inverse OT within this model asks which matrix A makes an observed matching or coupling look optimal; learning the cost is thus reduced to estimating a linear parameter.

For a fixed matrix A, the forward prediction is the optimal face

$$\mathcal{P}_A \coloneqq \operatorname*{argmin}_{\mathrm{P}\in \mathrm{U}(\mathbb{1}_n/n, \mathbb{1}_n/n)} \langle \mathrm{C}(A),\, \mathrm{P}\rangle.$$

When this face is a singleton, write its element as P_A; otherwise P_A denotes a deterministic tie-broken selection. Although $A \mapsto \mathrm{C}(A)$ is linear, the solution correspondence $A \mapsto \mathcal{P}_A$ is polyhedral: changing A changes the direction in which the transport polytope is probed, and a tie-broken selection is constant on normal-cone cells. Figure 12.9 illustrates this correspondence on the OT4ML point clouds. The construction follows the visual idea of the Python Optimal Transport logo [217]: red source atoms, blue target atoms and straight segments show the selected optimal bijection. With $e = (1, 1)^\top$ and $\delta = 10^{-3}$, the first two rank-one matrices are $A_h = -e_1e^\top + \delta e_2e^\top$ and $A_v = \delta e_1e^\top - e_2e^\top$. These small transverse terms break the large ties of the pure horizontal or vertical scores while preserving a rank-one cost. The matrix $A = -\mathrm{Id}$ gives the usual quadratic $\mathcal{W}_2$ assignment, up to the marginal-only terms discussed below, while $A = +\mathrm{Id}$ reverses the correlation and produces an anti-$\mathcal{W}_2$ matching.

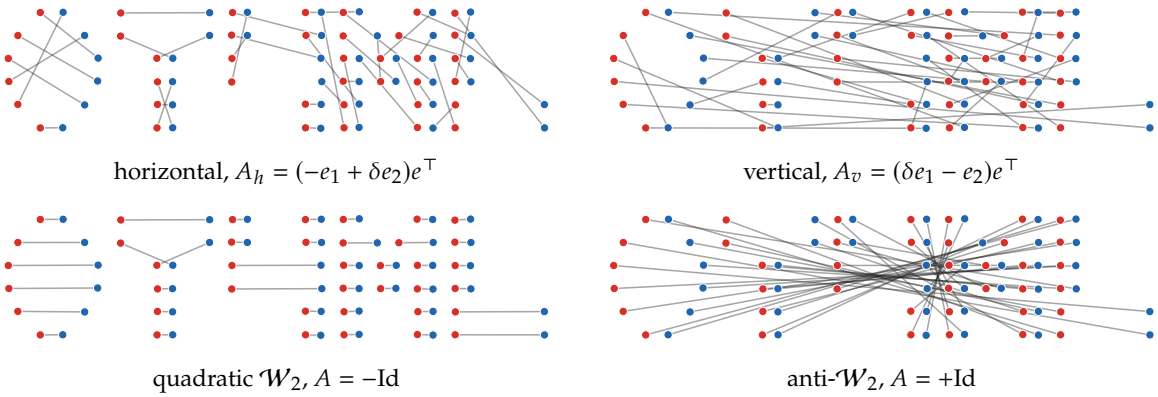


Figure 12.9: Forward solutions of the bilinear cost $c_A(x, y) = \langle Ax,\, y\rangle$ on the OT4ML logo point clouds. Each panel solves the equal-weight assignment problem with a different matrix A; the first two use $\delta = 10^{-3}$ to break rank-one ties. The source atoms are red, the target atoms are blue, and the gray segments give one deterministic optimal bijection.

This elementary model already contains the quadratic Wasserstein assignment. Adding to a cost matrix a term depending only on x_i or only on y_j shifts all feasible couplings by the same constant, and therefore does not change the optimizer. Since

$$\|x - y\|^2 = \|x\|^2 + \|y\|^2 - 2\langle x,\, y\rangle,$$

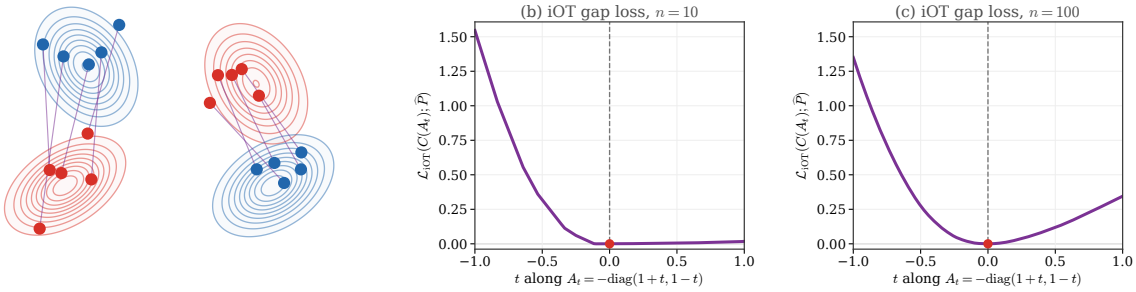
the usual quadratic Wasserstein assignment has the same optimizer as the bilinear cost with $A_\star = -\mathrm{Id}$, up to these marginal-only terms and an irrelevant positive factor. The inverse problem goes in the opposite direction: after observing a coupling, one asks which matrices A could have generated it. Figure 12.10 generates an observed coupling $\widehat{\mathrm{P}}$ from this cost on two empirical mixtures of Gaussians, and then evaluates $\mathcal{L}_{\mathrm{iOT}}(\mathrm{C}(A_t); \widehat{\mathrm{P}})$ along the anisotropic path

$$A_t = -\operatorname{diag}(1 + t, 1 - t), \qquad -1 \leqslant t \leqslant 1,$$

so that $t = 0$ recovers the matrix that generated the observed coupling. Equivalently, with equal weights, $\widehat{\mathrm{P}} \in \mathrm{U}(\mathbb{1}_n/n, \mathbb{1}_n/n) = \mathcal{B}_n/n$ and the plotted loss is the Kantorovich gap

$$\mathcal{L}_{\mathrm{iOT}}(\mathrm{C}(A_t); \widehat{\mathrm{P}}) = \langle \mathrm{C}(A_t), \widehat{\mathrm{P}} \rangle - \min_{\mathrm{P} \in \mathrm{U}(\mathbb{1}_n/n, \mathbb{1}_n/n)} \langle \mathrm{C}(A_t), \mathrm{P} \rangle, \qquad \mathrm{C}(A_t)_{i,j} = \langle A_t x_i, \, y_j \rangle.$$

Because $t \mapsto \mathrm{C}(A_t)$ is affine and the Kantorovich value is a minimum of affine functions over the fixed polytope $\mathrm{U}(\mathbb{1}_n/n, \mathbb{1}_n/n)$, this one-dimensional gap is convex and piecewise affine. Its zero set can contain an interval for a small sample, reflecting the fact that the same observed coupling remains optimal for a cone of nearby costs.



observed map, $n = 10$ loss curve, $n = 10$ loss curve, $n = 100$

Figure 12.10: Inverse-OT gap loss for a bilinear cost. Panel (a): two empirical mixtures of two Gaussians are matched with the cost $c_{A_\star}(x, y) = \langle A_\star x, \, y \rangle$ for $A_\star = -\mathrm{Id}$, which gives the same optimizer as the quadratic $\mathcal{W}_2$ cost; red and blue level sets display the two sampling densities. Panels (b,c): the unregularized Fenchel–Young Kantorovich gap $\mathcal{L}_{\mathrm{iOT}}(\mathrm{C}(A_t); \widehat{\mathrm{P}})$ along $A_t = -\mathrm{diag}(1 + t, 1 - t)$ for $n = 10$ and $n = 100$, using the same vertical scale. The red dot marks the generating parameter $t = 0$; the curves are convex and piecewise affine.

The comparison between $n = 10$ and $n = 100$ illustrates an important statistical effect: as the number of sampled points grows, the flat zero region can shrink to a sharper piecewise-affine minimum. A finite-sample linear-programming gap remains polyhedral and therefore has no classical second derivative inside its cells. Curvature is instead a population phenomenon. Peyré, Poon and Tron [432] prove local curvature and identifiability modulo the natural cost invariances for smooth positive marginals under a nondegeneracy condition. They also identify affine-map settings, including Gaussian or elliptical examples, as genuinely degenerate cases. Thus larger samples can reveal population curvature, but do not remove structural non-identifiability by themselves.

Proposition 12.30 (Convex dual-gap formulation of inverse OT)**.** *Let $\widehat{\mathrm{P}} \in \mathrm{U}(\mathrm{a}, \mathrm{b})$ be an observed coupling and let C_θ depend affinely on $\theta \in \Theta$, where Θ is convex. The condition that $\widehat{\mathrm{P}}$ is optimal for the cost C_θ is equivalent to the existence of dual potentials (f, g) such that*

$$f_i + g_j \leqslant (\mathrm{C}_\theta)_{i,j} \quad \text{and} \quad \sum_{i,j} \widehat{\mathrm{P}}_{i,j} \big((\mathrm{C}_\theta)_{i,j} - f_i - g_j \big) = 0.$$

Consequently, for a convex regularizer R and $\lambda \geqslant 0$, the dual-gap fitting problem

$$\min_{\theta \in \Theta, f, g} \; R(\theta) + \lambda \sum_{i,j} \widehat{\mathrm{P}}_{i,j} \big((\mathrm{C}_\theta)_{i,j} - f_i - g_j \big) \quad \text{subject to} \quad f_i + g_j \leqslant (\mathrm{C}_\theta)_{i,j} \quad \forall i, j. \tag{12.42}$$

is convex. It is nontrivial only if Θ imposes a cost normalization or otherwise excludes the zero-cost and marginal-gauge degeneracies.

Proof. For a fixed cost C_θ, Kantorovich duality gives

$$\min_{\mathrm{P} \in \mathrm{U}(\mathrm{a}, \mathrm{b})} \langle \mathrm{C}_\theta, \mathrm{P} \rangle = \max_{f_i + g_j \leqslant (\mathrm{C}_\theta)_{i,j}} \langle f, \mathrm{a} \rangle + \langle g, \mathrm{b} \rangle.$$

Since $\widehat{\mathrm{P}}$ has marginals (a, b), every dual feasible pair satisfies

$$\langle \mathrm{C}_\theta, \widehat{\mathrm{P}} \rangle - \langle f, \mathrm{a} \rangle - \langle g, \mathrm{b} \rangle = \sum_{i,j} \widehat{\mathrm{P}}_{i,j} \big((\mathrm{C}_\theta)_{i,j} - f_i - g_j \big) \geqslant 0.$$

This nonnegative quantity is exactly the primal-dual gap of $\widehat{\mathrm{P}}$. It vanishes if and only if $\widehat{\mathrm{P}}$ reaches the dual value and is therefore optimal. If C_θ is affine, $\lambda \geqslant 0$, and Θ and R are convex, the constraints and objective in (12.42) are convex. If $\mathrm{C}_\theta = 0$ is allowed and minimizes R, then $(\mathrm{C}, f, g) = (0, 0, 0)$ is a trivial minimizer; this proves the need for the normalization stated above. □

The formulation (12.42) avoids differentiating through a forward OT solver: it learns a normalized cost by making the observed plan nearly satisfy complementary slackness. In statistical settings, $\widehat{\mathrm{P}}$ is only partially observed or noisy, so one adds sparsity, low-rank, smoothness or metric constraints to select a meaningful representative [196, 16]. For entropic OT, the optimality condition becomes smoother:

$$\widehat{\mathrm{P}}_{i,j} \approx a_i b_j \exp\left(\frac{f_i + g_j - (\mathrm{C}_\theta)_{i,j}}{\varepsilon}\right),$$

which leads to likelihood-based or KL-based convex objectives when C_θ is affine, and connects inverse OT with generalized Sinkhorn iterations and transport-regularized inverse problems [298, 359]. Neural parameterizations of C_θ are more flexible but reintroduce non-convexity; the convex formulation above is the clean mathematical baseline.

12.6 Weak Optimal Transport

Weak OT relaxes the cost so that it depends on the conditional distribution of destinations rather than only on pointwise pairs. It is useful when a source point is allowed to choose a randomized response and the model only penalizes an aggregate of that response, such as its conditional mean.

Barycentric projection of a coupling. The first object to isolate is therefore the map obtained by collapsing each conditional law to its barycenter.

Definition 12.31 (Barycentric projection of a coupling)**.** Let $\alpha, \beta \in \mathcal{P}_1(\mathbb{R}^d)$ and let $\pi \in \mathcal{U}(\alpha, \beta)$. Disintegrate π with respect to its first marginal as $\pi(\mathrm{d}x, \mathrm{d}y) = \pi_x(\mathrm{d}y)\alpha(\mathrm{d}x)$. Since β has finite first moment, the conditional mean is finite for α-a.e. x, and the barycentric projection of π is the map

$$\bar{T}_\pi(x) := \int_{\mathbb{R}^d} y \mathrm{d}\pi_x(y), \qquad \bar{\beta}_\pi := (\bar{T}_\pi)_\sharp \alpha. \tag{12.43}$$

The projected target $\bar{\beta}_\pi$ records the distribution of conditional means, not the full second marginal. Thus it is generally different from β; if $\pi = (\mathrm{Id}, T)_\sharp \alpha$ is induced by a map, then $\bar{T}_\pi = T$ and $\bar{\beta}_\pi = \beta$. This projection is not an optimal map for an arbitrary coupling: a deterministic rotation of a radially symmetric source, for example, projects to the rotation itself, whereas the optimal map from the source to itself is the identity. The useful positive statement is attached to quadratic optimal plans, as in the tangent-space viewpoint on $\mathcal{W}_2$ developed by Ambrosio, Gigli and Savaré [12, Chap. 7].

Proposition 12.32 (Barycentric projection of a quadratic optimal plan)**.** *Let $\pi \in \mathcal{U}(\alpha, \beta)$ be optimal for the quadratic cost $\|x - y\|^2$ between $\alpha, \beta \in \mathcal{P}_2(\mathbb{R}^d)$, and define $\bar{T}_\pi$ and $\bar{\beta}_\pi$ by (12.43). Then $(\mathrm{Id}, \bar{T}_\pi)_\sharp \alpha$ is an optimal coupling between α and $\bar{\beta}_\pi$. Equivalently, $\bar{T}_\pi$ is a quadratic optimal transport map from α to the projected target $\bar{\beta}_\pi$.*

Proof. By Theorem 3.46, π is concentrated on a c-cyclically monotone set Γ for $c(x, y) = \|x - y\|^2$. For the quadratic cost, and since it is enough to check cyclic permutations, this means that every finite cycle $(x_i, y_i)_{i=1}^m \subset \Gamma$ satisfies

$$\sum_{i=1}^m \langle x_i, y_i \rangle \geqslant \sum_{i=1}^m \langle x_i, y_{i+1} \rangle, \qquad y_{m+1} = y_1.$$

After changing the disintegration on an α-negligible set, π_x is supported on the section $\Gamma_x = \{y \,;\, (x, y) \in \Gamma\}$ for α-a.e. x. Choose $x_1, \dots, x_m$ in this full-measure set and independently sample $Y_i \sim \pi_{x_i}$. Applying the cyclic inequality to (x_i, Y_i) and taking expectations gives

$$\sum_{i=1}^m \langle x_i, \bar{T}_\pi(x_i) \rangle \geqslant \sum_{i=1}^m \langle x_i, \bar{T}_\pi(x_{i+1}) \rangle.$$

Thus $(\mathrm{Id}, \bar{T}_\pi)_\sharp \alpha$ is concentrated on a cyclically monotone graph. By the cyclic-monotonicity characterization of quadratic optimality, this plan is optimal between its two marginals, namely α and $\bar{\beta}_\pi$. □

> **Remark 12.33 (Barycentric projection appears everywhere).** Barycentric projection is the operation that turns a conditional law into a mean. Definition 12.31 and (12.43) apply this to a coupling disintegrated as $\pi(\mathrm{d}x, \mathrm{d}y) = \pi_x(\mathrm{d}y)\alpha(\mathrm{d}x)$, and Proposition 12.32 shows that quadratic optimal plans are stable under this collapse. The same operation is the organizing summary in weak OT: the barycentric weak cost in Proposition 12.36 is obtained by inserting $\bar{T}_\pi$ into (12.44). Martingale OT fixes this summary instead of merely penalizing it, imposing $\int y \, \mathrm{d}\pi_x(y) = x$ in (12.45). Conditional Wasserstein distances in Definition 11.45 use the same disintegration language fiber by fiber, see (11.55). Later, the mean-shift and Gaussian-attention velocity (16.36) is again a barycentric average, now of a kernel-weighted neighborhood, and SVGD in Section 16.2 replaces this barycentric averaging by the RKHS steepest-descent average (16.22). Across these examples, the recurring move is to keep the conditional structure of a coupling but replace the full conditional law by a tractable first moment or kernelized moment.

Weak transport costs. Weak transport costs use the same disintegration but allow the objective to depend on the whole conditional law, or on summaries such as the barycentric projection (12.43). The framework was introduced through general transport costs and weak transport inequalities in [250]; existence, duality and optimality conditions on Polish spaces are developed in [26]. For a weak cost $C : \mathcal{X} \times \mathcal{P}(\mathcal{Y}) \to \mathbb{R} \cup \{+\infty\}$, the weak OT value is

Definition 12.34 (Weak optimal transport)**.** Let $\mathcal{X}$ and $\mathcal{Y}$ be Polish spaces, let $\alpha \in \mathcal{P}(\mathcal{X})$ and $\beta \in \mathcal{P}(\mathcal{Y})$, and let $C : \mathcal{X} \times \mathcal{P}(\mathcal{Y}) \to \mathbb{R} \cup \{+\infty\}$ be measurable. For each $\pi \in \mathcal{U}(\alpha, \beta)$, let $\pi(\mathrm{d}x, \mathrm{d}y) = \alpha(\mathrm{d}x)\pi_x(\mathrm{d}y)$ be a disintegration. Then

$$\mathrm{WOT}_C(\alpha, \beta) := \inf_{\pi \in \mathcal{U}(\alpha,\beta)} \int C(x, \pi_x) \mathrm{d}\alpha(x). \tag{12.44}$$

The classical Kantorovich problem is recovered when $C(x, \nu) = \int c(x, y) \mathrm{d}\nu(y)$, because the objective then becomes $\int c(x, y) \mathrm{d}\pi(x, y)$. The genuinely weak behavior starts when C is nonlinear in ν.

Proposition 12.35 (Weak Kantorovich duality)**.** *Let $\mathcal{X}$ and $\mathcal{Y}$ be compact metric spaces, and let $C : \mathcal{X} \times \mathcal{P}(\mathcal{Y}) \to \mathbb{R} \cup \{+\infty\}$ be proper, jointly lower semicontinuous, bounded from below and convex in its second argument. Fix $\alpha \in \mathcal{P}(\mathcal{X})$ and $\beta \in \mathcal{P}(\mathcal{Y})$, and assume that $\mathrm{WOT}_C(\alpha, \beta) < +\infty$. For $g \in C(\mathcal{Y})$, define the weak C-transform*

$$g^C(x) := \inf_{\nu \in \mathcal{P}(\mathcal{Y})} \left\{ C(x, \nu) - \int g(y) \mathrm{d}\nu(y) \right\}.$$

Then

$$\mathrm{WOT}_C(\alpha, \beta) = \sup_{g \in C(\mathcal{Y})} \left\{ \int g^C(x) \mathrm{d}\alpha(x) + \int g(y) \mathrm{d}\beta(y) \right\}.$$

When $C(x, \nu) = \int c(x, y) \mathrm{d}\nu(y)$, this reduces to the usual Kantorovich dual with $g^C(x) = \inf_y (c(x, y) - g(y))$.

Proof. For any coupling π and any $g \in C(\mathcal{Y})$, the definition of g^C gives $C(x, \pi_x) \geqslant g^C(x) + \int g(y) \mathrm{d}\pi_x(y)$. After integration with respect to α, the second term becomes $\int g \mathrm{d}\beta$ because the second marginal of π is β. This proves weak duality.

For the converse, lift a kernel $x \mapsto \pi_x$ to the measure $\alpha(\mathrm{d}x)\delta_{\pi_x}(\mathrm{d}\nu)$ on $\mathcal{X} \times \mathcal{P}(\mathcal{Y})$. Relaxing this graph constraint gives probability measures P whose first marginal is α and whose intensity satisfies $\int_{\mathcal{X} \times \mathcal{P}(\mathcal{Y})} \nu \, \mathrm{d}P(x, \nu) = \beta$. This relaxation does not change the value: disintegrating $P(\mathrm{d}x, \mathrm{d}\nu) = P_x(\mathrm{d}\nu)\alpha(\mathrm{d}x)$ and replacing P_x by its barycenter $\bar{\nu}_x = \int \nu \, \mathrm{d}P_x(\nu)$ preserves the intensity constraint and cannot increase the objective, by convexity of $C(x, \cdot)$. The relaxed feasible set is compact, and the integral functional is lower semicontinuous. Fenchel–Rockafellar duality for the affine intensity constraint therefore has no gap; its continuous multiplier is $g \in C(\mathcal{Y})$. Minimization in ν gives $g^C(x)$, while the constraint contributes $\int g \mathrm{d}\beta$. This is the weak-cost analogue of Proposition 5.2; see [26] for the Polish-space formulation. □

Proposition 12.36 (Barycentric weak transport is weaker than $\mathcal{W}_2$)**.** *Let $\alpha, \beta \in \mathcal{P}_2(\mathbb{R}^d)$ and define*

$$C_{\text{bar}}(x, \nu) = \|x - \int y \mathrm{d}\nu(y)\|^2.$$

Equivalently, for a coupling π, the integrand is $\|x - \bar{T}_\pi(x)\|^2$. Then

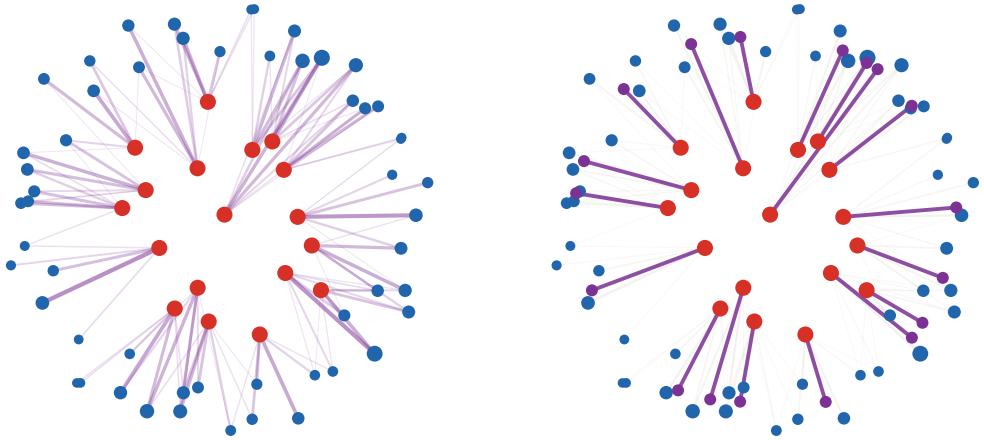
$$\mathrm{WOT}_{C_{\text{bar}}}(\alpha, \beta) \leqslant \mathcal{W}_2^2(\alpha, \beta).$$

Proof. Let π be any coupling and disintegrate it as $\pi_x \alpha$. By Jensen's inequality,

$$\|x - \bar{T}_\pi(x)\|^2 \leqslant \int \|x - y\|^2 \mathrm{d}\pi_x(y).$$

Integrating in x gives $\int C_{\text{bar}}(x, \pi_x)\mathrm{d}\alpha(x) \leqslant \int \|x - y\|^2 \mathrm{d}\pi(x, y)$. Taking the infimum over π proves the claim. □

Figure 12.11 separates the full conditional coupling from its barycentric projection, making explicit which part of each conditional law is retained by the weak cost.



conditionals π_x barycentric projection $\bar{T}_\pi$

Figure 12.11: Weak barycentric transport on a small disk-to-annulus coupling. The left panel shows the full conditional laws: each red source atom splits its mass among several blue target atoms, with segment thickness proportional to transported mass. The right panel collapses each conditional law π_x to its barycenter $\bar{T}_\pi(x) = \int y \mathrm{d}\pi_x(y)$, shown in violet. The barycentric weak cost only sees the red-to-violet displacement, and therefore ignores the conditional spread around each barycenter.

The barycentric cost is the canonical example to keep in mind: admissibility still constrains the full conditional laws to have second marginal β, but the objective only charges the displacement from x to $\bar{T}_\pi(x)$ and ignores the conditional variance around this barycenter. This connects weak OT with martingale transport, Strassen-type convex-order constraints, barycentric projections and learning problems where conditional averages are meaningful objects.

12.6.1 Martingale Optimal Transport

Martingale OT is the extreme barycentric version of the weak viewpoint: a source point may split randomly, but the average destination must remain equal to the source point. This turns the barycentric projection from an object used in the cost into a hard constraint.

Definition 12.37 (Martingale couplings and martingale OT)**.** Let $\alpha, \beta \in \mathcal{P}_1(\mathbb{R}^d)$. A coupling $\pi \in \mathcal{U}(\alpha, \beta)$ is a martingale coupling if, for the disintegration $\pi(\mathrm{d}x, \mathrm{d}y) = \pi_x(\mathrm{d}y)\alpha(\mathrm{d}x)$,

$$\int_{\mathbb{R}^d} y \mathrm{d}\pi_x(y) = x \qquad \text{for } \alpha\text{-a.e. } x. \tag{12.45}$$

Equivalently, $\bar{T}_\pi = \mathrm{Id}$ in (12.43). The set of such couplings is denoted

$$\mathcal{U}_{\mathrm{mart}}(\alpha,\beta) := \left\{\pi \in \mathcal{U}(\alpha,\beta)\ ;\ \bar{T}_\pi = \mathrm{Id} \quad \alpha\text{-a.e.}\right\}.$$

For a cost c, the martingale OT problem minimizes $\int c\mathrm{d}\pi$ over $\mathcal{U}_{\mathrm{mart}}(\alpha,\beta)$, with value $+\infty$ when the admissible set is empty.

The terminology comes from probability: if $(X,Y)\sim\pi$, then (12.45) is exactly $\mathbb{E}[Y|X]=X$. Hence martingale OT is a Kantorovich problem with the usual two marginal constraints plus a barycentric constraint on the conditional laws. Equivalently, the barycentric projected coupling

$$(\mathrm{Id},\bar{T}_\pi)_\sharp\alpha$$

must be the diagonal coupling $(\mathrm{Id},\mathrm{Id})_\sharp\alpha$. This is stronger than merely asking the projected target $(\bar{T}_\pi)_\sharp\alpha$ to equal α, since a nontrivial measure-preserving map could have the same projected marginal without satisfying $\bar{T}_\pi(x)=x$ pointwise. Martingale OT is central in robust finance, where one transports today prices to tomorrow prices without introducing drift, and has led to a rich martingale transport theory [32, 226, 190, 257].

Stochastic orders. The admissibility of constrained couplings is governed by stochastic order. The basic principle is that inequalities tested against a class of functions are equivalent to the existence of couplings satisfying a pointwise or conditional constraint. Strassen's theorem is the canonical result of this kind [506].

The one-dimensional stochastic order tests whether every increasing observable is larger under the second law.

Definition 12.38 (Classical stochastic order)**.** For $\alpha,\beta\in\mathcal{P}(\mathbb{R})$,

$$\alpha \preceq_{\mathrm{st}} \beta \quad\Longleftrightarrow\quad \int \varphi\,\mathrm{d}\alpha \leqslant \int \varphi\,\mathrm{d}\beta \quad \text{for every bounded increasing } \varphi.$$

Proposition 12.39 (Strassen's theorem for stochastic order)**.** *Let $\alpha,\beta\in\mathcal{P}(\mathbb{R})$. Then $\alpha\preceq_{\mathrm{st}}\beta$ if and only if there exists $\pi\in\mathcal{U}(\alpha,\beta)$ concentrated on $\{(x,y)\ ;\ x\leqslant y\}$.*

Proof. A coupling supported on $x\leqslant y$ immediately gives the integral inequality for every increasing φ. Conversely, testing approximate indicators of intervals $(t,+\infty)$ gives $F_\alpha(t)\geqslant F_\beta(t)$ for all continuity points of the distribution functions. If U is uniform on $(0,1)$ and $F_\alpha^{-1}, F_\beta^{-1}$ are the generalized quantiles, then $X=F_\alpha^{-1}(U)$ and $Y=F_\beta^{-1}(U)$ have laws α and β, and the inequality between distribution functions gives $X\leqslant Y$ almost surely. □

Convex order and martingale feasibility. For martingale OT, the pointwise order constraint is replaced by a barycentric constraint on conditional laws. The corresponding order is the convex order.

Definition 12.40 (Convex order)**.** For $\alpha,\beta\in\mathcal{P}_1(\mathbb{R}^d)$, write $\alpha\preceq_{\mathrm{cx}}\beta$ when

$$\int \varphi\,\mathrm{d}\alpha \leqslant \int \varphi\,\mathrm{d}\beta$$

for every continuous convex $\varphi:\mathbb{R}^d\to\mathbb{R}$ with at most linear growth.

Testing affine functions gives equality of means, while the remaining convex tests say that β is more spread out than α. Strassen's martingale theorem says that this spread condition is exactly what is needed to realize β from α by mean-preserving randomization.

Theorem 12.41 (Strassen's martingale theorem)**.** *Let $\alpha,\beta\in\mathcal{P}_1(\mathbb{R}^d)$. Then*

$$\alpha\preceq_{\mathrm{cx}}\beta \quad\Longleftrightarrow\quad \mathcal{U}_{\mathrm{mart}}(\alpha,\beta)\neq\emptyset.$$

Equivalently, β is above α in convex order if and only if there exists a coupling $\pi(\mathrm{d}x,\mathrm{d}y)=\pi_x(\mathrm{d}y)\alpha(\mathrm{d}x)$ such that $\int y\mathrm{d}\pi_x(y)=x$ for α-almost every x.

Proof. If π is a martingale coupling, then Jensen's inequality gives, for every convex φ,

$$\varphi(x) = \varphi\left(\int y \mathrm{d}\pi_x(y)\right) \leqslant \int \varphi(y)\mathrm{d}\pi_x(y),$$

and integration in x gives $\int \varphi \mathrm{d}\alpha \leqslant \int \varphi \mathrm{d}\beta$.

Conversely, assume $\alpha \preceq_{\mathrm{cx}} \beta$. First suppose that both measures are supported in a compact convex set K. Separate β from the closed convex set of probability measures on K that are reachable from α by martingale kernels supported in K. If β were not in this set, a continuous separating function $\psi : K \to \mathbb{R}$ would satisfy

$$\int \psi \mathrm{d}\beta > \sup\left\{\int \psi \mathrm{d}\eta \ : \ \eta \in \mathcal{P}(K),\ \mathcal{U}_{\mathrm{mart}}(\alpha,\eta) \neq \emptyset\right\}.$$

For a fixed $x \in K$, optimizing over probability measures on K with barycenter x gives the concave envelope $\mathrm{conc}_K \psi(x)$; hence the right-hand side is $\int \mathrm{conc}_K \psi \mathrm{d}\alpha$. Since convex order is equivalently the reverse inequality for concave functions,

$$\int \mathrm{conc}_K \psi \mathrm{d}\alpha \geqslant \int \mathrm{conc}_K \psi \mathrm{d}\beta \geqslant \int \psi \mathrm{d}\beta,$$

a contradiction. For general measures in $\mathcal{P}_1(\mathbb{R}^d)$, the same separation argument is carried out in the $\mathcal{W}_1$ topology, whose continuous test functions have at most linear growth. Compactness is replaced by tightness and uniform integrability of first moments; these properties also make the set of attainable second marginals closed and preserve the martingale constraint under limits. This is the standard extension in Strassen's theorem [506]. □

The same theorem yields an exact geometric description of the barycentric weak cost. Rather than transporting all the way to β, weak OT transports optimally to the closest measure that lies below β in convex order.

Proposition 12.42 (Brenier–Strassen projection formula)**.** *Let $\alpha, \beta \in \mathcal{P}_2(\mathbb{R}^d)$ and let C_{bar} be the barycentric quadratic cost of Proposition 12.36. Then*

$$\mathrm{WOT}_{C_{\mathrm{bar}}}(\alpha,\beta) = \inf_{\eta \in \mathcal{P}_2(\mathbb{R}^d):\, \eta \preceq_{\mathrm{cx}} \beta} \mathcal{W}_2^2(\alpha,\eta). \tag{12.46}$$

Proof. Let (X, Y) have law $\pi \in \mathcal{U}(\alpha,\beta)$, set $Z = \mathbb{E}[Y \mid X]$, and denote its law by η. Conditional Jensen gives $\mathbb{E}[\varphi(Z)] \leqslant \mathbb{E}[\varphi(Y)]$ for every convex φ, so $\eta \preceq_{\mathrm{cx}} \beta$. Since (X, Z) couples α and η,

$$\mathcal{W}_2^2(\alpha,\eta) \leqslant \mathbb{E}\|X - Z\|^2 = \int C_{\mathrm{bar}}(x, \pi_x)\mathrm{d}\alpha(x).$$

Taking the infimum over π proves that the left-hand side of (12.46) is no smaller than the right-hand side.

Conversely, fix $\eta \preceq_{\mathrm{cx}} \beta$. By Theorem 12.41, there is a martingale coupling from η to β. Glue it to an optimal coupling (X, Z) between α and η, conditionally independently given Z. The resulting variables satisfy $\mathbb{E}[Y \mid Z] = Z$, hence $\mathbb{E}[Y \mid X] = \mathbb{E}[Z \mid X]$. Conditional Jensen therefore gives

$$\mathbb{E}\|X - \mathbb{E}[Y \mid X]\|^2 = \mathbb{E}\|X - \mathbb{E}[Z \mid X]\|^2 \leqslant \mathbb{E}\|X - Z\|^2 = \mathcal{W}_2^2(\alpha,\eta).$$

Taking the infimum over η proves the reverse inequality. Existence and finer structure of the projected measure are developed in [26]. □

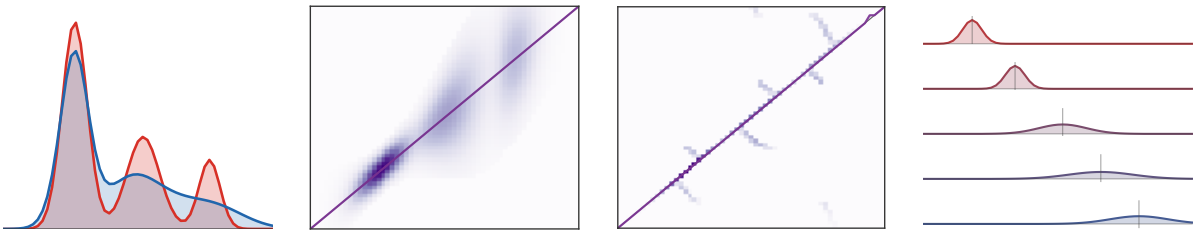
Theorem 12.41 explains why convex order is the right admissibility notion for martingale OT: it is exactly the feasibility condition for the constraint (12.45). If $\alpha \not\preceq_{\mathrm{cx}} \beta$, then $\mathcal{U}_{\mathrm{mart}}(\alpha,\beta) = \emptyset$ and the martingale OT value is $+\infty$, independently of the cost. If $\alpha \preceq_{\mathrm{cx}} \beta$, then the optimization problem is nonempty and the cost selects, among all mean-preserving splittings of each source point, the martingale coupling best adapted to the application. This is the probabilistic meaning of the barycentric constraint: mass may branch, but it cannot drift on average.

For Gaussian measures with the same mean, convex order reduces to the Loewner order on covariance matrices:

$$\mathcal{N}(m,\Sigma_0) \preceq_{\mathrm{cx}} \mathcal{N}(m,\Sigma_1) \quad \Longleftrightarrow \quad \Sigma_1 - \Sigma_0 \succeq 0.$$

Indeed, if $\Sigma_1 - \Sigma_0 \succeq 0$, then $\mathcal{N}(m, \Sigma_1)$ is obtained from $\mathcal{N}(m, \Sigma_0)$ by adding independent centered Gaussian noise; the converse follows by testing the convex quadratic functions $x \mapsto \langle u, x\rangle^2$.

Figure 12.12 gives a discrete non-Gaussian counterpart: centered conditional kernels provide a feasible martingale plan, while optimizing the transport cost selects a much sparser plan with the same marginals and barycentric constraint.



marginals generated plan optimal plan centered kernels

Figure 12.12: A discrete one-dimensional martingale OT example. The source α is a red Gaussian mixture on a grid. A first feasible plan is generated by centered kernels $K_i(y_j) = \kappa_i(y_j - x_i)$, whose discrete barycenter is x_i, so $\mathrm{P}_{i,j} = a_i K_i(y_j)$ satisfies $\bar{T}_{\mathrm{P}}(x_i) = x_i$. Keeping the same two marginals, the third panel solves the finite-dimensional martingale OT linear program with objective $\sum_{i,j} |x_i - y_j| \mathrm{P}_{i,j}$, row and column constraints, and martingale constraints $\sum_j (y_j - x_i)\mathrm{P}_{i,j} = 0$. The optimized plan is much sparser, while both plans have the identity barycentric projection.

CHAPTER 13

Beyond Comparing Measures

This chapter leaves the setting of scalar measures on a common ambient space. Vector- and matrix-valued OT transports mass with internal degrees of freedom, Gromov–Wasserstein compares metric-measure spaces without a prescribed correspondence, and quantum OT replaces scalar couplings by positive operators. In each case, the transport plan must also encode structure carried by the support, the fibers or the non-commutative state space.

13.1 Vector and Matrix-Valued Measures

Scalar OT transports a nonnegative density. In imaging, color processing, spectral analysis, diffusion tensor imaging and quantum-inspired models, the object attached to a point can instead have several nonnegative components or a positive semidefinite matrix. The first step beyond scalar OT is the positive vector-valued case: the fiber remains linear and commutative, but the transport cost may couple its channels.

Positive vector-valued measures. The simplest way to keep internal structure in transport is to attach several nonnegative masses to each spatial point and to decide whether these channels move independently or interact through the cost.

Definition 13.1 (Positive vector-valued measure)**.** A positive $\mathbb{R}^m_+$-valued measure on $\mathcal{X}$ is a tuple

$$\alpha = (\alpha^1, \dots, \alpha^m) \in \mathcal{M}_+(\mathcal{X}; \mathbb{R}^m_+),$$

where each component α^k is a nonnegative finite measure.

This models multi-channel densities such as color histograms, spectral bins or several species transported on the same domain. In a conservative model the mass of each channel is preserved, so one assumes $\alpha_0^k(\mathcal{X}) = \alpha_1^k(\mathcal{X})$ for every k. The natural vector-valued extension of OT therefore starts from the positive cone $\mathbb{R}^m_+$.

For the dynamic formulas below, assume that $\mathcal{X}$ is either $\mathbb{R}^d$, the flat torus, or a bounded convex domain in $\mathbb{R}^d$ equipped with no-flux boundary conditions. First suppose that the endpoints and the curve have densities. The direct analogue of Benamou–Brenier fixes a vector density $u_t(x) \in \mathbb{R}^m_+$ and a spatial flux $V_t(x) = (V_{t,1}, \dots, V_{t,d}) \in (\mathbb{R}^m)^d$, where $V^k_{t,\ell}$ is the momentum of channel k in spatial direction ℓ. The conservative vector transport cost associated with an action density Φ is

Definition 13.2 (Vector-valued dynamic transport)**.** For a nonnegative extended-valued action density $\Phi : \mathbb{R}^m_+ \times (\mathbb{R}^m)^d \to [0, +\infty]$, define

$$\mathcal{W}^2_\Phi(\alpha_0, \alpha_1) := \inf_{u,V} \int_0^1 \int_{\mathcal{X}} \Phi(u_t(x), V_t(x))\, \mathrm{d}x\, \mathrm{d}t \tag{13.1}$$

subject to the endpoint constraints $u_0 \mathrm{d}x = \alpha_0$, $u_1 \mathrm{d}x = \alpha_1$ and the componentwise continuity equation

$$\partial_t u_t + \nabla_x \cdot V_t = 0, \qquad (\nabla_x \cdot V_t)^k = \sum_{\ell=1}^{d} \partial_{x_\ell} V^k_{t,\ell}. \tag{13.2}$$

Thus each component satisfies its own continuity equation, but the cost may still couple the components. Singular curves are handled as in scalar dynamic OT by replacing densities and fluxes by measures and using the lower semicontinuous perspective recession convention.

The following elementary family separates independent channel motion from genuinely coupled vector transport.

Example 13.3 (Diagonal and coupled positive mobilities). Choose a mobility matrix $\mathsf{M}(u) \in \mathbb{S}^m_+$, where $\mathbb{S}^m_+$ denotes the cone of real symmetric positive semidefinite matrices, and set

$$\Phi_{\mathsf{M}}(u, V) = \sum_{\ell=1}^{d} V_\ell^\top \mathsf{M}(u)^\dagger V_\ell,$$

with the usual convention that the value is finite only when each V_ℓ belongs to the range of $\mathsf{M}(u)$. If $u \mapsto \mathsf{M}(u)$ is linear and takes values in $\mathbb{S}^m_+$, then this action is jointly convex and positively one-homogeneous in (u, V). Indeed, the matrix fractional map $(M, z) \mapsto z^\top M^\dagger z$ is jointly convex on its effective domain, and linearity gives $\mathsf{M}(su) = s\mathsf{M}(u)$ for $s > 0$. For $m = 1$ and $\mathsf{M}(u) = u$, one recovers exactly the scalar Benamou–Brenier action. For

$$\mathsf{M}_{\mathrm{diag}}(u) = \mathrm{diag}(u_1, \dots, u_m),$$

the channels move independently. Non-diagonal mobilities are the simplest way to couple the coordinates while keeping the same componentwise conservation law. For instance, with $q = m^{-1/2}(1, \dots, 1)$ and $\kappa \geqslant 0$,

$$\mathsf{M}_\kappa(u) = \mathrm{diag}(u) + \kappa\Big(\sum_{k=1}^{m} u_k\Big) q q^\top$$

increases the mobility in the common channel direction q while leaving transverse directions controlled by the diagonal part. The local cost of moving one component can therefore depend on the densities and momenta of the other components, even though each component mass remains conserved.

Proposition 13.4 (Diagonal positive vector Benamou–Brenier). *Assume that $\alpha_0^k, \alpha_1^k \in \mathcal{M}_+(\mathcal{X})$ have finite second moments and the same mass m_k for every k. For the diagonal mobility $\mathsf{M}_{\mathrm{diag}}$, the value of (13.1) is*

$$\mathcal{W}^2_{\mathrm{diag}}(\alpha_0, \alpha_1) = \sum_{k:m_k>0} m_k \, \mathcal{W}_2^2\left(\frac{\alpha_0^k}{m_k}, \frac{\alpha_1^k}{m_k}\right),$$

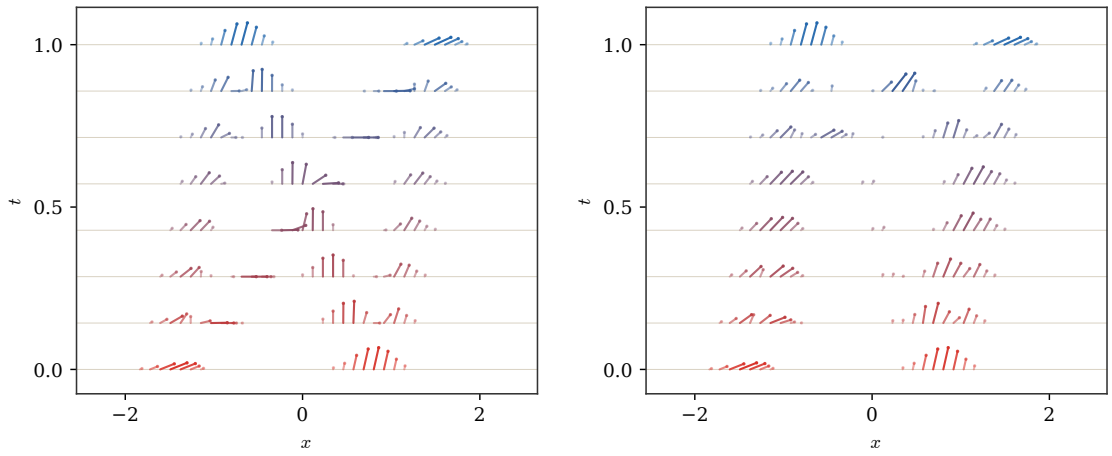
with the convention that zero-mass channels contribute zero.

Proof. For $\mathsf{M}_{\mathrm{diag}}(u) = \mathrm{diag}(u_1, \dots, u_m)$, the action separates as

$$\sum_{\ell=1}^{d} V_\ell^\top \mathsf{M}_{\mathrm{diag}}(u)^\dagger V_\ell = \sum_{k=1}^{m} \frac{|V^k|^2}{u^k},$$

where $V^k = (V_1^k, \dots, V_d^k)$ is the spatial momentum of channel k, and the scalar perspective convention is used. The constraint (13.2) also separates into $\partial_t u^k + \nabla \cdot V^k = 0$. The minimization therefore splits into m independent scalar Benamou–Brenier problems. If $m_k = 0$, nonnegativity and conservation force the whole channel to vanish. If $m_k > 0$, normalizing $\rho_t^k = u_t^k / m_k$ and $p_t^k = V_t^k / m_k$ factors the channel action as $m_k \int |p_t^k|^2 / \rho_t^k$, hence the scalar value is $m_k \, \mathcal{W}_2^2(\alpha_0^k / m_k, \alpha_1^k / m_k)$. Summing over the channels proves the claim. □

The conservative positive-cone model above is the basic extension of Benamou–Brenier. Adding a source term $\partial_t u + \nabla \cdot V = S$ and a convex perspective penalty in S gives unbalanced or reaction–transport variants. Such generalized transport models with dissipation and density modulation were developed by Maas, Rumpf, Schönlieb and Simon [362, 363]; related nonlinear mobility distances and gradient structures appear in [189, 384]. Figure 13.1 contrasts the exact diagonal case $\kappa = 0$, where each positive channel is transported by its quantile map, with a large-κ illustrative common-mode interpolation in which the channels move more coherently. The endpoints are two-mode mixtures: at each spatial mode the two channels have Gaussian profiles with the same center but different amplitudes.



independent channels ($\kappa = 0$) strongly coupled channels ($\kappa \gg 1$)

Figure 13.1: One-dimensional positive $\mathbb{R}^2_+$-valued transport displayed by arrow glyphs at eight time levels. Each endpoint is a mixture of two localized Gaussian modes, and, inside each mode, both channel profiles have the same center. Each arrow is proportional to the local fiber value $(u_t^1(x), u_t^2(x))$, and time runs vertically from the red source to the blue target. Left: for $\kappa = 0$, the diagonal mobility of Proposition 13.4 gives two independent scalar quantile geodesics. Right: a large-κ common-mode interpolation bends the display toward the direction $q = 2^{-1/2}(1, 1)$, illustrating the qualitative effect of a mobility that favors coherent channel motion while keeping the same componentwise continuity equation.

Positive matrix-valued measures. The next simplest fiber is the positive matrix cone. This is the simplest tensor-valued model beyond vectors: the diagonal entries behave like positive channels, while the eigenvectors encode local orientations.

Definition 13.5 (Positive matrix-valued measure)**.** Write $\mathbb{S}^m$ for real symmetric matrices and $\mathbb{S}^m_+$ for the positive semidefinite cone. A positive $\mathbb{S}^m_+$-valued measure is a countably additive map from Borel sets of $\mathcal{X}$ to $\mathbb{S}^m_+$; we denote the cone of such measures by

$$\mathcal{A} \in \mathcal{M}_+(\mathcal{X}; \mathbb{S}^m_+).$$

Equivalently, $\mathcal{A}$ is a symmetric matrix of finite signed measures such that $\mathcal{A}(E) \succeq 0$ for every Borel set E. If $\mathcal{A}$ has density $A(x) \succeq 0$, then $\operatorname{tr} A(x)$ is the scalar amount of mass at x, while, wherever $\operatorname{tr} A(x) > 0$, the normalized matrix $A(x)/\operatorname{tr} A(x)$ records an internal covariance or orientation. This is the matrix analogue of the positive vector case: diagonal matrices encode nonnegative vector components, and non-diagonal matrices add a local eigenbasis.

The conservative Benamou–Brenier model fixes a matrix density $A_t(x) \in \mathbb{S}^m_+$ and symmetric matrix fluxes $P_t(x) = (P_{t,1}, \dots, P_{t,d}) \in (\mathbb{S}^m)^d$. With no flux through the boundary of $\mathcal{X}$, the full matrix mass $\int_{\mathcal{X}} A_t(x)\mathrm{d}x$ is conserved, so the endpoints must have the same total matrix. The resulting matrix-perspective problem is the following.

Definition 13.6 (Matrix-valued dynamic transport)**.**

$$\mathcal{W}^2_{\mathrm{mat}}(\mathcal{A}_0, \mathcal{A}_1) := \inf_{A,P} \int_0^1 \int_{\mathcal{X}} \sum_{\ell=1}^{d} \operatorname{tr}\left(P_{t,\ell}^\top A_t^\dagger P_{t,\ell}\right)\mathrm{d}x\mathrm{d}t \tag{13.3}$$

subject to $A_0\mathrm{d}x = \mathcal{A}_0$, $A_1\mathrm{d}x = \mathcal{A}_1$ and to the matrix-valued continuity equation

$$\partial_t A_t + \nabla_x \cdot P_t = 0, \qquad \nabla_x \cdot P_t = \sum_{\ell=1}^{d} \partial_{x_\ell} P_{t,\ell}. \tag{13.4}$$

Here $A_t^\dagger$ is the Moore–Penrose inverse, and the integrand is $+\infty$ unless $\operatorname{Ran}(P_{t,\ell}) \subseteq \operatorname{Ran}(A_t)$ for every ℓ.

This is the usual lower-semicontinuous perspective convention. The matrix fractional map $(A, P) \mapsto$

$\operatorname{tr}(P^\top A^\dagger P)$ is jointly convex on this effective domain and positively one-homogeneous in (A, P). This gives the simplest non-trivial matrix-valued transport model: spatial motion is conservative, but the fiber carries orientation through the eigenvectors of $A_t(x)$.

Proposition 13.7 (Diagonal matrix subproblem)**.** *Assume that the endpoints are diagonal in a fixed orthonormal basis,*

$$\mathcal{A}_i = \operatorname{diag}(\alpha_i^1, \dots, \alpha_i^m), \qquad i = 0, 1,$$

and that $\alpha_0^k(\mathcal{X}) = \alpha_1^k(\mathcal{X}) = m_k$ for every k. If one restricts the admissible curves in (13.3) to remain diagonal in that basis,

$$A_t = \operatorname{diag}(u_t^1, \dots, u_t^m), \qquad P_{t,\ell} = \operatorname{diag}(V_{t,\ell}^1, \dots, V_{t,\ell}^m),$$

then the value of this restricted matrix problem is

$$\sum_{k:m_k>0} m_k \, \mathcal{W}_2^2\left(\frac{\alpha_0^k}{m_k}, \frac{\alpha_1^k}{m_k}\right),$$

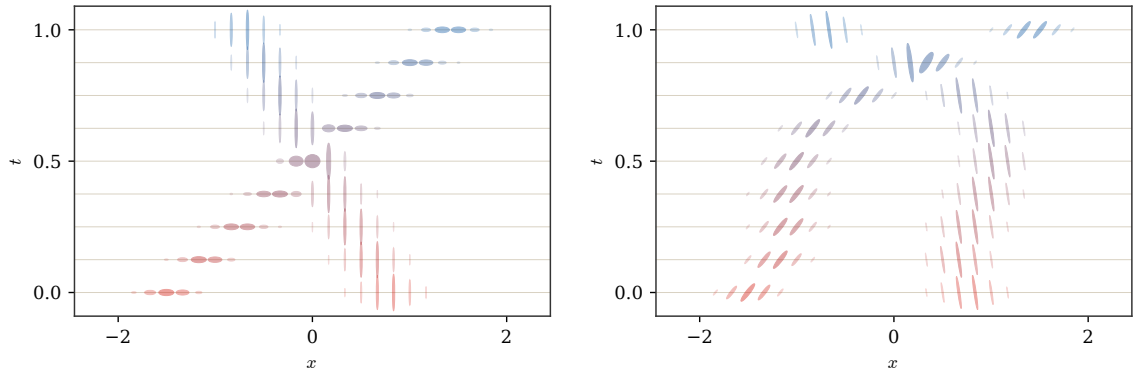
with zero contribution from zero-mass channels. Thus the commuting matrix submodel is exactly the diagonal positive vector-valued Benamou–Brenier model of Proposition 13.4.

Proof. The continuity equation (13.4) is diagonal entry by diagonal entry and gives $\partial_t u^k + \nabla \cdot V^k = 0$. Moreover,

$$\sum_{\ell=1}^{d} \operatorname{tr}\left(P_{t,\ell}^\top A_t^\dagger P_{t,\ell}\right) = \sum_{k=1}^{m} \frac{|V_t^k|^2}{u_t^k}$$

with the same scalar perspective convention as before. The admissible set and the action are therefore exactly those of the diagonal vector model. □

The restriction to a fixed diagonal basis gives eigenvalue transport; it should be read as a commuting submodel, not as a claim that non-diagonal excursions can never change the unrestricted value. The genuinely matrix-valued case starts when the eigenspaces vary with x or along the interpolation, so that the transported object carries both mass and orientation. Static matrix-valued Monge–Kantorovich problems and dual test-function metrics were developed in [402, 292, 403]; dynamic versions and related non-commutative geometries appear in [132, 130, 103, 429]. Figure 13.2 shows the analogous independent/coupled contrast for positive 2×2 matrix fibers, using two localized matrix modes whose eigenvalue profiles share a common center at each mode.



commuting tensor channels strongly coupled tensor fibers

Figure 13.2: Positive 2×2 matrix-valued transport on a one-dimensional base. Each endpoint is a mixture of two localized matrix modes; within one mode, both eigenvalue profiles are Gaussian bumps with the same center. Each ellipse is the glyph of a positive semidefinite matrix $A_t(x)$, with axes given by eigenvectors and eigenvalues. Left: the matrices are diagonal in a fixed basis, giving the commuting tensor analogue of independent vector channels. Right: a coupled illustrative interpolation bends packet motion toward the trace-density transport and uses non-commuting eigendirections; the superposition remains positive semidefinite and produces spatially varying orientations.

13.2 Wasserstein over Wasserstein

The construction can be iterated. Once $(\mathcal{X}, d)$ is a metric space, the set of probability measures on $\mathcal{X}$ becomes a metric space through $\mathcal{W}_p$. It can therefore serve as a new ground space. This is useful whenever the objects to compare are themselves random probability measures, or mixtures whose components are meaningful objects rather than only a collapsed density.

The standard setting is that of Polish spaces, introduced in Definition 2.3. These assumptions rule out many measure-theoretic pathologies: probability laws can be approximated by countable objects, tightness gives compactness criteria, regular conditional probabilities exist on the associated Borel spaces, and weak convergence is stable. The next proposition records that Wasserstein spaces preserve this structure, so the construction can be iterated without leaving the same well-behaved category.

Proposition 13.8 (Wasserstein spaces as ground spaces)**.** *Let $1 \leqslant p < \infty$. If $(\mathcal{X}, d)$ is Polish, then $(\mathcal{P}_p(\mathcal{X}), \mathcal{W}_p)$ is Polish. If $\mathcal{X}$ is compact, then $\mathcal{P}_p(\mathcal{X}) = \mathcal{P}(\mathcal{X})$ and this space is compact for $\mathcal{W}_p$. Consequently, one may iterate the construction, with the corresponding finite-moment condition at every level.*

Proof. This is a standard structural theorem for Wasserstein spaces [539, 469, 12]. For completeness, extract from a $\mathcal{W}_p$-Cauchy sequence a subsequence $(\alpha_k)_k$ such that $\sum_k \mathcal{W}_p(\alpha_k, \alpha_{k+1}) < \infty$. Gluing optimal couplings of consecutive terms produces random variables $(X_k)_k$ with laws α_k and $\sum_k \|d(X_k, X_{k+1})\|_{L^p} < \infty$. Hence X_k converges almost surely and in L^p to an $\mathcal{X}$-valued random variable; its law is the $\mathcal{W}_p$ limit of the subsequence, and the Cauchy property gives convergence of the original sequence. Separability follows by approximating measures with finitely supported measures on a countable dense subset and rational weights. If $\mathcal{X}$ is compact, Prokhorov's theorem gives compactness of $\mathcal{P}(\mathcal{X})$ for weak convergence, and Proposition 4.25 identifies this topology with the $\mathcal{W}_p$ topology. □

Fix $1 \leqslant p < \infty$. We denote elements of $\mathcal{P}_p(\mathcal{X})$ by $\alpha, \beta, \ldots$. Elements of $\mathcal{P}_p(\mathcal{P}_p(\mathcal{X}))$ are denoted by fraktur letters, for instance $\mathfrak{A}, \mathfrak{B}$; they are probability laws over probability measures, or random probability measures. A basic parametric example is obtained from a measurable family $(\alpha_\zeta)_{\zeta \in Z}$ and a probability law γ on the parameter space:

$$\mathfrak{A} = (\zeta \mapsto \alpha_\zeta)_\sharp \gamma. \tag{13.5}$$

It belongs to $\mathcal{P}_p(\mathcal{P}_p(\mathcal{X}))$ precisely when, for one and hence every $x_0 \in \mathcal{X}$,

$$\int_Z \mathcal{W}_p(\alpha_\zeta, \delta_{x_0})^p \, \mathrm{d}\gamma(\zeta) < \infty.$$

If $\gamma = \sum_{i=1}^K a_i \delta_{\zeta_i}$, then

$$\mathfrak{A} = \sum_{i=1}^K a_i \delta_{\alpha_{\zeta_i}}.$$

Definition 13.9 (Collapsed, or barycentric, mixture)**.** For $\mathfrak{A} \in \mathcal{P}(\mathcal{P}(\mathcal{X}))$, the collapsed, or barycentric, mixture associated with $\mathfrak{A}$ is the measure $\bar{\alpha}_{\mathfrak{A}}$ defined by

$$\int_{\mathcal{X}} f(x) \mathrm{d}\bar{\alpha}_{\mathfrak{A}}(x) = \int_{\mathcal{P}(\mathcal{X})} \left(\int_{\mathcal{X}} f(x) \mathrm{d}\alpha(x) \right) \mathrm{d}\mathfrak{A}(\alpha), \tag{13.6}$$

for bounded continuous f.

In the finite case, $\bar{\alpha}_{\mathfrak{A}} = \sum_i a_i \alpha_{\zeta_i}$. Moreover, if $\mathfrak{A} \in \mathcal{P}_p(\mathcal{P}_p(\mathcal{X}))$, then $\bar{\alpha}_{\mathfrak{A}} \in \mathcal{P}_p(\mathcal{X})$ because

$$\int_{\mathcal{X}} d(x, x_0)^p \, \mathrm{d}\bar{\alpha}_{\mathfrak{A}}(x) = \int_{\mathcal{P}_p(\mathcal{X})} \mathcal{W}_p(\alpha, \delta_{x_0})^p \, \mathrm{d}\mathfrak{A}(\alpha).$$

The Wasserstein construction can now be applied once more on this ground space.

Definition 13.10 (Wasserstein-over-Wasserstein distance)**.** Let $1 \leqslant p < +\infty$ and let $\mathcal{X}$ be a Polish metric space. For $\mathfrak{A}, \mathfrak{B} \in \mathcal{P}_p(\mathcal{P}_p(\mathcal{X}))$, define

$$\mathbb{W}_p^p(\mathfrak{A}, \mathfrak{B}) := \inf_{\Pi \in \mathcal{U}(\mathfrak{A}, \mathfrak{B})} \int_{\mathcal{P}_p(\mathcal{X}) \times \mathcal{P}_p(\mathcal{X})} \mathcal{W}_p^p(\alpha, \beta) \mathrm{d}\Pi(\alpha, \beta). \tag{13.7}$$

For $p = 2$, Gaussian mixtures provide a particularly explicit example with two levels of geometry. A mixture $\sum_i a_i \mathcal{N}(m_i, \Sigma_i)$ can either be viewed as the collapsed measure on $\mathcal{X}$, or as the component law

$$\mathfrak{A} = \sum_i a_i \delta_{\mathcal{N}(m_i, \Sigma_i)}$$

on the Bures-Wasserstein space of Gaussian components. Given two component laws

$$\mathfrak{A} = \sum_i a_i \delta_{\mathcal{N}(m_i, \Sigma_i)}, \qquad \mathfrak{B} = \sum_j b_j \delta_{\mathcal{N}(n_j, \Lambda_j)},$$

the discrete problem induced by (13.7) uses the component cost

$$\mathrm{C}_{i,j} = \|m_i - n_j\|^2 + \mathcal{B}(\Sigma_i, \Lambda_j)^2.$$

Let $\mathrm{P}^\star$ be an optimal coupling between the weights a and b. If $A_{i,j}$ denotes the Brenier linear part from Σ_i to Λ_j, then each active component pair is interpolated by

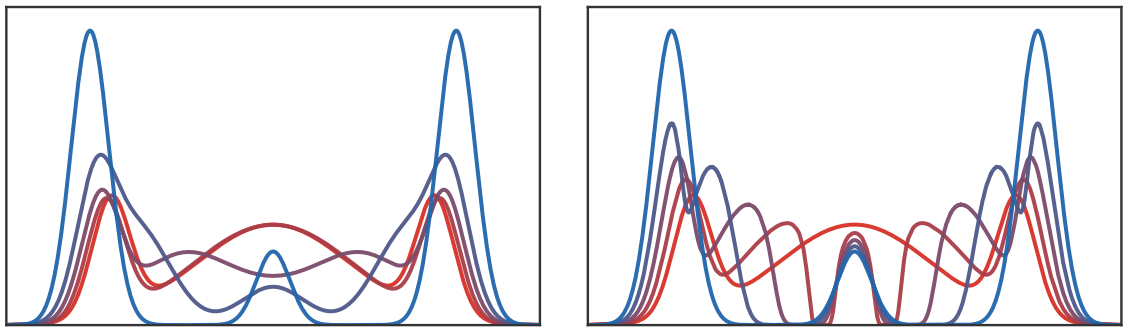
$$m_{ij,t} = (1-t)m_i + t n_j, \qquad \Sigma_{ij,t} = \big((1-t)\mathrm{Id} + tA_{i,j}\big)\Sigma_i\big((1-t)\mathrm{Id} + tA_{i,j}\big)^\top,$$

and collapsing these component geodesics gives

$$\bar{\alpha}_t = \sum_{i,j} \mathrm{P}^\star_{i,j} \mathcal{N}(m_{ij,t}, \Sigma_{ij,t}).$$

This component-level interpolation transports Gaussian components as atoms of the Wasserstein space before returning to measures on $\mathcal{X}$. It is generally not the same as the true $\mathcal{W}_2$ interpolation between the collapsed mixture densities, which can split and recombine mass inside and across components.

Figure 13.3 contrasts this component-level geodesic with the true one-dimensional Wasserstein interpolation of the collapsed mixtures, making the internal mass splitting absent from the former visible.



component law on Gaussian atoms collapsed mixture law

Figure 13.3: Two interpolations between the same asymmetric three-component one-dimensional Gaussian mixtures. The red endpoint has a broad central component carrying most of the mass, while the blue endpoint has two dominant sharp side modes. Left: Gaussian components are transported as atoms using their Bures-Wasserstein distance. Right: the collapsed densities are interpolated by the true one-dimensional quantile formula for $\mathcal{W}_2$. The central mass is split and recombined in the collapsed geometry, making the two paths visibly distinct.

Proposition 13.11 (Collapsing is non-expansive)**.** *Let $(\mathcal{X}, d)$ be Polish, let $1 \leqslant p < \infty$, and let $\mathfrak{A}, \mathfrak{B} \in \mathcal{P}_p(\mathcal{P}_p(\mathcal{X}))$. For the collapsed mixtures defined by* (13.6)*,*

$$\mathcal{W}_p(\bar{\alpha}_{\mathfrak{A}}, \bar{\beta}_{\mathfrak{B}}) \leqslant \mathbb{W}_p(\mathfrak{A}, \mathfrak{B}).$$

Proof. Fix $\Pi \in \mathcal{U}(\mathfrak{A}, \mathfrak{B})$ and $\eta > 0$. A standard measurable-selection theorem for optimal transport on Polish spaces gives a measurable kernel $(\alpha, \beta) \mapsto \pi_{\alpha,\beta} \in \mathcal{U}(\alpha, \beta)$ such that

$$\int_{\mathcal{X}\times\mathcal{X}} d(x,y)^p \, \mathrm{d}\pi_{\alpha,\beta}(x,y) \leqslant \mathcal{W}_p(\alpha,\beta)^p + \eta.$$

Integrating this kernel against Π gives a coupling $\bar{\pi}$ between $\bar{\alpha}_{\mathfrak{A}}$ and $\bar{\beta}_{\mathfrak{B}}$. Its cost satisfies

$$\int_{\mathcal{X}\times\mathcal{X}} d(x,y)^p \mathrm{d}\bar{\pi}(x,y) \leqslant \int_{\mathcal{P}_p(\mathcal{X})^2} \mathcal{W}_p^p(\alpha,\beta)\mathrm{d}\Pi(\alpha,\beta) + \eta.$$

Letting $\eta \downarrow 0$, taking the infimum over Π, and then taking the p-th root proves the claim. □

The following remark records a useful way in which this iterated construction reappears in the Gromov–Wasserstein lower bounds of the next section.

Remark 13.12 (Local profiles as Wasserstein-over-Wasserstein laws). Given a metric-measure space $\mathbb{X} = (\mathcal{X}, d_{\mathcal{X}}, \alpha_{\mathcal{X}})$, each point defines a local distance distribution

$$\alpha_x = (d_{\mathcal{X}}(x,\cdot))_\sharp \alpha_{\mathcal{X}} \in \mathcal{P}(\mathbb{R}_+), \qquad \mathfrak{D}_{\mathbb{X}} = (x \mapsto \alpha_x)_\sharp \alpha_{\mathcal{X}} \in \mathcal{P}(\mathcal{P}(\mathbb{R}_+)).$$

The Mémoli profile lower bound in Proposition 13.25 is precisely a Wasserstein-over-Wasserstein comparison of these laws of local profiles. It replaces the full pairwise distortion by an ordinary OT problem whose ground cost is itself a one-dimensional Wasserstein distance.

13.3 Gromov–Wasserstein

Gromov–Wasserstein compares spaces through their internal distance structures rather than through a fixed ambient ground cost. This is the right extension for graphs, shapes and point clouds whose points are not pre-aligned.

Discrete formulation. Optimal transport needs a ground cost C to compare histograms (a, b), and thus cannot be used directly if the histograms are not defined on the same underlying space, or if one cannot pre-register these spaces to define a ground cost. To address this issue, one can instead use a weaker requirement: two matrices $D \in \mathbb{R}^{n\times n}$ and $D' \in \mathbb{R}^{m\times m}$ are available and represent relationships between the points on which the histograms are defined. A typical scenario is when these matrices are powers of distance matrices. Define the quadratic distortion energy

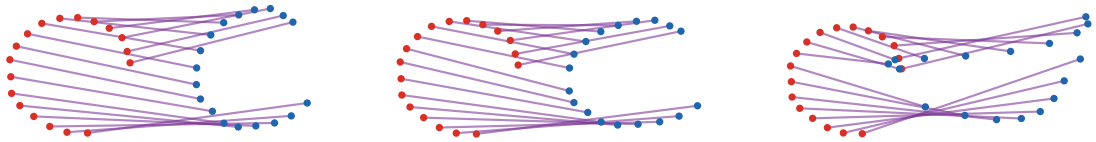
$$\mathcal{E}_{D,D'}(P) := \sum_{i,j,i',j'} \Delta(D_{i,i'}, D'_{j,j'})^p P_{i,j} P_{i',j'}, \qquad \mathrm{GW}((a,D),(b,D'))^p := \min_{P\in U(a,b)} \mathcal{E}_{D,D'}(P). \tag{13.8}$$

where $p \geqslant 1$ and Δ is a distance on $\mathbb{R}$, typically $\Delta(u,v) = |u-v|$. The objective seeks a correspondence that is as close as possible to an isometry. Indeed, if two matched pairs (i,j) and (i',j') carry significant coupling mass, then their within-space distances should satisfy $D_{i,i'} \simeq D'_{j,j'}$. Equivalently, $\mathcal{E}_{D,D'}(P)$ is the expected distortion between two pairs drawn independently from P. In the hard-assignment case $j = \sigma(i)$, minimizing it tries to enforce $D_{i,i'} \simeq D'_{\sigma(i),\sigma(i')}$ simultaneously for all pairs of matched points. This is a quadratic optimization problem over the transport polytope, and its restriction to the feasible affine directions need not be convex. In the uniform case with the additional hard-assignment constraint $m = n$ and P a permutation matrix, the corresponding generalized structural-matching problem is a Quadratic Assignment Problem (QAP) [352]. By contrast, for the relaxed finite-metric problem (13.8), neither NP-hardness nor a polynomial-time exact algorithm is currently known; the precise complexity status and references are given after (13.21). The relaxed coupling formulation used in GW can therefore be read as a soft graph-matching model [358].

Example 13.13 (Application to structured objects). Graphs, molecules, meshes and point clouds often do not come with a common coordinate system or a reliable node-to-node correspondence. GW compares them through their internal distances, so the coupling $P_{i,j}$ becomes a soft structural correspondence rather than a spatial displacement. This is useful for graph matching, node embedding, shape comparison and molecule or mesh alignment [431, 535]. When node attributes are also available, fused GW below adds a feature transport term to the intrinsic distortion, separating what should be

preserved as structure from what should be matched as signal.

Figure 13.4 shows this intrinsic matching principle under progressively stronger deformations: correspondences are selected from within-space distance patterns rather than an ambient cross-space cost.



isometric copy mild deformation strong deformation

Figure 13.4: Gromov–Wasserstein correspondences under increasing deformation. The red and blue point clouds are not compared through an ambient Euclidean cross-cost; instead, the GW coupling compares their internal pairwise distances. A perfectly isometric copy admits a clean structural match, while mild and deliberately stronger deformations progressively bend the correspondence.

When D and D′ satisfy the metric axioms on $\{1,\dots,m\}$ and $\{1,\dots,n\}$, respectively, these weighted sets define finite metric-measure spaces with measures $\sum_i a_i\delta_i$ and $\sum_j b_j\delta_j$. Theorem 13.20 therefore applies directly to (13.8): GW satisfies the triangle inequality and defines a distance modulo measure-preserving isometries. This distance was introduced and studied in detail by Memoli in [376]. An in-depth mathematical exposition (in particular, its geodesic structure and gradient flows) is given in [509]. See also [477] for applications in computer vision. Its relation to Hausdorff and Gromov–Hausdorff distances is discussed at the end of this section.

General setting. The continuous formulation abstracts the discrete distance matrices into metric-measure spaces, so that GW compares intrinsic geometries independently of labels, parametrizations or ambient coordinates.

Definition 13.14 (Metric-measure space)**.** A metric-measure space is a triple

$$\mathbb{X} = (X, d_X, \alpha),$$

where (X, d_X) is Polish and α is a Borel probability measure on X. For a finite p-Gromov–Wasserstein distance, one additionally assumes the finite-size condition

$$\int_{X\times X} d_X(x,x')^p \,\mathrm{d}\alpha(x)\mathrm{d}\alpha(x') < \infty.$$

The general setting corresponds to computing couplings between metric-measure spaces $\mathbb{X} = (X, d_X, \alpha)$ and $\mathbb{Y} = (Y, d_Y, \beta)$, where the distance and the measure are both part of the data. Compactness is often assumed below to avoid additional tightness and integrability arguments. The distortion between two metric-measure spaces is minimized over couplings of their mass distributions.

Definition 13.15 (Gromov–Wasserstein distance)**.** Fix $p \geqslant 1$ and a distance Δ on $[0,+\infty)$. Then

$$\mathcal{GW}(\mathbb{X},\mathbb{Y})^p := \min_{\pi\in\mathcal{U}(\alpha,\beta)} \int_{X^2\times Y^2} \Delta(d_X(x,x'), d_Y(y,y'))^p \,\mathrm{d}\pi(x,y)\mathrm{d}\pi(x',y'). \tag{13.9}$$

Definition 13.16 (Infinite-order Gromov–Wasserstein distance)**.** For a distance Δ on $[0,+\infty)$, the infinite-order Gromov–Wasserstein distance is

$$\mathcal{GW}_\infty(\mathbb{X},\mathbb{Y}) := \inf_{\pi\in\mathcal{U}(\alpha,\beta)} \operatorname*{ess\,sup}_{\pi\otimes\pi} \Delta\big(d_X(x,x'), d_Y(y,y')\big). \tag{13.10}$$

On compact spaces, the infinite-order value is the genuine endpoint of the finite-order family. The support formula below will also be useful when comparing GW with Gromov–Hausdorff distance.

Proposition 13.17 (Infinite-order limit of Gromov–Wasserstein)**.** *Let $\mathbb{X}$ and $\mathbb{Y}$ be compact metric-measure spaces, and assume that Δ is continuous. Then*

$$\mathcal{GW}_\infty(\mathbb{X},\mathbb{Y}) = \inf_{\pi\in\mathcal{U}(\alpha,\beta)} \sup_{\big((x,y),(x',y')\big)\in\operatorname{supp}(\pi)^2} \Delta\big(d_X(x,x'),d_Y(y,y')\big). \tag{13.11}$$

Moreover, $\mathcal{GW}_p(\mathbb{X},\mathbb{Y})$ is nondecreasing in p and

$$\lim_{p\to\infty} \mathcal{GW}_p(\mathbb{X},\mathbb{Y}) = \mathcal{GW}_\infty(\mathbb{X},\mathbb{Y}). \tag{13.12}$$

Proof. For each coupling π, continuity of the distortion integrand identifies its essential supremum under $\pi\otimes\pi$ with its supremum over $\operatorname{supp}(\pi)^2$, which proves (13.11). Monotonicity follows from monotonicity of probability-space L^p norms, and gives a limit $L \leqslant \mathcal{GW}_\infty(\mathbb{X},\mathbb{Y})$.

Let $p_k\to\infty$ and let π_k minimize the order-p_k problem. Compactness of $\mathcal{U}(\alpha,\beta)$ gives, after extraction, $\pi_k \rightharpoonup \pi_\star$. If the distortion vanishes on $\operatorname{supp}(\pi_\star)^2$, the result is immediate. Otherwise, fix $(x,y),(x',y')\in\operatorname{supp}(\pi_\star)$ with positive distortion and a number

$$0 < r < \Delta\big(d_X(x,x'),d_Y(y,y')\big).$$

By continuity, there are neighborhoods U and V of these points on whose product the distortion exceeds r. The Portmanteau theorem gives $\liminf_k \pi_k(U) > 0$ and $\liminf_k \pi_k(V) > 0$. Hence, for some $c>0$ and all sufficiently large k,

$$\mathcal{GW}_{p_k}(\mathbb{X},\mathbb{Y}) \geqslant r\big(\pi_k(U)\pi_k(V)\big)^{1/p_k} \geqslant r c^{1/p_k}.$$

Letting first $k\to\infty$ and then r increase to the distortion of the selected pair shows that L dominates every distortion on $\operatorname{supp}(\pi_\star)^2$. Therefore

$$L \geqslant \sup_{\big((x,y),(x',y')\big)\in\operatorname{supp}(\pi_\star)^2} \Delta\big(d_X(x,x'),d_Y(y,y')\big) \geqslant \mathcal{GW}_\infty(\mathbb{X},\mathbb{Y}),$$

which proves (13.12). □

Taking $\mathcal{X}=\{1,\dots,m\}$, $\mathcal{Y}=\{1,\dots,n\}$, $\alpha=\sum_i \mathrm{a}_i\delta_i$ and $\beta=\sum_j \mathrm{b}_j\delta_j$ reduces (13.9) exactly to the discrete problem (13.8), since every coupling is represented by a matrix $\mathrm{P}\in\mathrm{U}(\mathrm{a},\mathrm{b})$. Thus the construction above is precisely the finite-space, or equivalently finitely supported, specialization of the metric-measure formulation.

Comparison with Wasserstein and Wasserstein–Procrustes. The three distances retain different amounts of geometric information. Ordinary Wasserstein transport compares absolute positions, Wasserstein–Procrustes first removes ambient rigid motions, and GW retains only the metric-measure structure. On a common Euclidean ambient space, these nested invariances yield a useful two-sided estimate.

Proposition 13.18 (GW, Wasserstein–Procrustes and Wasserstein)**.** *Let $p\geqslant 1$, let $\alpha,\beta\in\mathcal{P}_p(\mathcal{X})$ be probability measures on the same metric space $(\mathcal{X},d_\mathcal{X})$, and take $\Delta(u,v)=|u-v|$ in (13.9). Then*

$$\mathcal{GW}((\mathcal{X},d_\mathcal{X},\alpha),(\mathcal{X},d_\mathcal{X},\beta)) \leqslant 2\,\mathcal{W}_p(\alpha,\beta), \tag{13.13}$$

where $\mathcal{W}_p$ uses the same ground metric $d_\mathcal{X}$. In particular, when $\mathcal{X}=\mathbb{R}^d$ is equipped with the Euclidean distance, write $\mathbb{X}_\alpha=(\mathbb{R}^d,\|\cdot\|,\alpha)$ and $\mathbb{X}_\beta=(\mathbb{R}^d,\|\cdot\|,\beta)$. Then

$$\frac{1}{2}\,\mathcal{GW}(\mathbb{X}_\alpha,\mathbb{X}_\beta) \leqslant \mathcal{W}_{p,\mathrm{E}(d)}([\alpha],[\beta]) \leqslant \mathcal{W}_p(\alpha,\beta). \tag{13.14}$$

For $p=2$, the middle term is the Wasserstein–Procrustes distance of Definition 11.36. The factor $1/2$ reflects the normalization in (13.9); it disappears under the alternative convention $\Delta(u,v)=|u-v|/2$.

Proof. Fix $\pi\in\mathcal{U}(\alpha,\beta)$, and draw (x,y) and (x',y') independently from π. The reverse triangle inequality gives

$$|d_\mathcal{X}(x,x') - d_\mathcal{X}(y,y')| \leqslant d_\mathcal{X}(x,y) + d_\mathcal{X}(x',y').$$

Taking the $L^p(\pi\otimes\pi)$ norm, applying Minkowski's inequality, and optimizing over π proves (13.13).

Now specialize to $\mathbb{R}^d$. Choosing the identity in Definition 11.34 proves the right inequality in (13.14). Rigid motions preserve pairwise Euclidean distances, so the GW value is unchanged when α and β are pushed forward by arbitrary $g, h \in \mathrm{E}(d)$. Applying (13.13) to these aligned measures gives

$$\mathcal{GW}(\mathbb{X}_\alpha, \mathbb{X}_\beta) \leqslant 2\,\mathcal{W}_p(g_\sharp\alpha, h_\sharp\beta).$$

Taking the infimum over g, h proves the left inequality. □

To turn GW from a distortion score into a metric statement, one must quotient out the relabelings that preserve both distances and mass.

Definition 13.19 (Isometric metric-measure spaces)**.** Two metric-measure spaces $\mathbb{X} = (\mathcal{X}, d_\mathcal{X}, \alpha)$ and $\mathbb{Y} = (\mathcal{Y}, d_\mathcal{Y}, \beta)$ are isometric if there exists a bijection $\varphi : \mathrm{supp}(\alpha) \to \mathrm{supp}(\beta)$ such that $\varphi_\sharp\alpha = \beta$ and

$$d_\mathcal{Y}(\varphi(x), \varphi(x')) = d_\mathcal{X}(x, x')$$

for all $x, x' \in \mathrm{supp}(\alpha)$.

The next theorem explains why the averaged distortion above is not merely a matching score. Once one quotients out measure-preserving isometries, it defines a genuine distance between metric-measure spaces.

Theorem 13.20 (Gromov–Wasserstein metric modulo isometries)**.** *For compact metric-measure spaces, $p \geqslant 1$ and $\Delta(u, v) = |u - v|$, $\mathcal{GW}$ defines a distance up to measure-preserving isometries [376].*

Proof. Compactness of the supports makes the coupling set compact and the distortion functional continuous, so an optimal coupling exists. If φ is a measure-preserving isometry, the graph coupling $(\mathrm{Id}, \varphi)_\sharp\alpha$ has zero distortion; in particular, $\mathcal{GW}(\mathbb{X}, \mathbb{X}) = 0$. Symmetry and non-negativity are immediate.

If $\mathcal{GW}(\mathbb{X}, \mathbb{Y}) = 0$ and π is an optimal plan, then $d_\mathcal{X}(x, x') = d_\mathcal{Y}(y, y')$ holds $\pi \otimes \pi$-almost everywhere. By continuity, this equality holds on $\mathrm{supp}(\pi)^2$. We show that $\mathbb{X}$ and $\mathbb{Y}$ are isometric by showing that both are isometric to the support space $(\mathrm{supp}(\pi), d_\pi, \pi)$, where

$$d_\pi((x, y), (x', y')) \coloneqq \frac{1}{2} d_\mathcal{X}(x, x') + \frac{1}{2} d_\mathcal{Y}(y, y').$$

The first projection $\psi : (x, y) \mapsto x$ is measure-preserving. For $((x, y), (x', y')) \in \mathrm{supp}(\pi)^2$,

$$d_\mathcal{X}(\psi(x, y), \psi(x', y')) = d_\mathcal{X}(x, x') = d_\mathcal{Y}(y, y') = d_\pi((x, y), (x', y')),$$

so ψ is an isometry and therefore injective. To see surjectivity onto $\mathrm{supp}(\alpha)$, take $x \in \mathrm{supp}(\alpha)$. Since $\psi_\sharp\pi = \alpha$, there is a sequence $(x_k, y_k) \in \mathrm{supp}(\pi)$ with $x_k \to x$. The equality of distances on $\mathrm{supp}(\pi)$ makes $(y_k)_k$ Cauchy, and compactness gives a convergent subsequence with limit $(x, y) \in \mathrm{supp}(\pi)$. The same argument for the second projection shows that the support space is also isometric to $\mathbb{Y}$.

For the triangle inequality, let π be an optimal coupling between $\mathbb{X}$ and $\mathbb{Y}$, and ξ an optimal coupling between $\mathbb{Y}$ and $\mathbb{Z} = (Z, d_Z, \gamma)$. By the gluing lemma, take σ on $\mathcal{X} \times \mathcal{Y} \times Z$ whose $(\mathcal{X}, \mathcal{Y})$ and $(\mathcal{Y}, Z)$ marginals are π and ξ. Let $\rho = (P_{\mathcal{X},Z})_\sharp\sigma$, and write $\bar\sigma = \sigma \otimes \sigma$ for the product law of two independent triples (x, y, z) and (x', y', z'). Then ρ is feasible between $\mathbb{X}$ and $\mathbb{Z}$, and

$$\begin{aligned}
\mathcal{GW}(\mathbb{X}, \mathbb{Z}) &\leqslant \left(\int |d_\mathcal{X}(x, x') - d_Z(z, z')|^p \, \mathrm{d}\bar\sigma \right)^{1/p} \\
&\leqslant \left(\int |d_\mathcal{X}(x, x') - d_\mathcal{Y}(y, y')|^p \, \mathrm{d}\bar\sigma \right)^{1/p} + \left(\int |d_\mathcal{Y}(y, y') - d_Z(z, z')|^p \, \mathrm{d}\bar\sigma \right)^{1/p} \\
&= \mathcal{GW}(\mathbb{X}, \mathbb{Y}) + \mathcal{GW}(\mathbb{Y}, \mathbb{Z}),
\end{aligned}$$

where the second inequality uses the pointwise triangle inequality followed by Minkowski's inequality. This proves the triangle inequality and completes the metric proof. □

Remark 13.21 (Completeness and compactness). The metric conclusion of Theorem 13.20 does not make the full moduli space of compact metric-measure spaces complete or compact. A standard counterexample is the uniform measure on the n-point equilateral space Δ_n, whose distinct points are

all at distance one. Under the normalization (13.9),

$$\mathcal{GW}(\Delta_n, \Delta_m) \leqslant n^{-1/p} + m^{-1/p},$$

so $(\Delta_n)_n$ is GW-Cauchy. Indeed, couple Δ_{nm} to Δ_n by the block projection: the two pairwise distances disagree with probability at most $1/n$, and similarly for Δ_m, after which the triangle inequality applies. The formal limit would be an atomless probability space in which two independent points are at distance one almost surely. Such a limit has no realization by a compact, or even separable, metric support. Thus the class of compact metric-measure spaces is not complete for $\mathcal{GW}$, and consequently is not compact; its completion requires generalized pseudo-metric or gauged measure spaces [377, 509].

This contrasts with Wasserstein geometry over one *fixed* compact ground space Ω. There, $\mathcal{P}(\Omega)$ is weakly compact by Prokhorov's theorem, and Proposition 4.25 identifies weak convergence with $\mathcal{W}_p$ convergence; hence $(\mathcal{P}(\Omega), \mathcal{W}_p)$ is compact. For GW the ground space itself varies, and compactness of every individual space gives no uniform control of the family. GW precompactness requires additional uniform geometric hypotheses, such as diameter and covering-number bounds.

Remark 13.22 (Distance hierarchy and topology). The comparison (13.14) immediately gives

$$\mathcal{W}_p\text{-convergence} \implies \mathcal{W}_{p,\mathrm{E}(d)}\text{-convergence} \implies \mathcal{GW}\text{-convergence}.$$

The first implication is strict when measures, rather than their rigid-motion orbits, are compared. For example, let $\alpha = \frac{2}{3}\delta_0 + \frac{1}{3}\delta_{e_1}$ and $\alpha_n = (x \mapsto x + ne_1)_\sharp \alpha$. Then $\mathcal{W}_{p,\mathrm{E}(d)}([\alpha_n], [\alpha]) = \mathcal{GW}(\mathbb{X}_{\alpha_n}, \mathbb{X}_\alpha) = 0$ for every n, whereas $\mathcal{W}_p(\alpha_n, \alpha) = n$. Thus rigid registration removes a genuine obstruction to ordinary Wasserstein convergence.

The second implication is subtler. On families of compact Euclidean metric-measure spaces with fixed ambient dimension and uniformly bounded diameters, there is no sequence that converges in GW but not in Wasserstein–Procrustes. Indeed, a distance-preserving bijection between Euclidean supports extends from their affine hulls to an ambient rigid motion, so the two distances have the same zero classes. More quantitatively, Mémoli [376] proves that distortion GW and Sturm's common-embedding transport distance induce the same topology, while [375] gives a dimension-dependent Hölder comparison between the latter and isometry-invariant Wasserstein transport on uniformly bounded Euclidean sets. Consequently, GW and Wasserstein–Procrustes induce the same quotient topology on this class, although they are not uniformly Lipschitz equivalent. Thus the three distances are genuinely different numerically, but they do not define three strictly nested topologies in this Euclidean setting.

Marginal stability and geodesics. The GW metric structure controls changes in the marginal measures and also provides intrinsic displacement interpolations between metric-measure spaces.

A useful consequence is that, when the geometry is fixed and only the sampling of the measures changes, GW inherits ordinary Wasserstein sampling rates. This is often the right way to read empirical GW: the non-convexity of the coupling problem is algorithmic, but the value is Lipschitz with respect to marginal perturbations measured in $\mathcal{W}_p$.

Proposition 13.23 (Marginal stability and empirical GW rates)**.** *Let $\mathbb{X} = (X, d_X, \alpha)$ and $\mathbb{Y} = (Y, d_Y, \beta)$ be compact metric-measure spaces, with $\Delta(u, v) = |u - v|$ and exponent $p \geqslant 1$. For any probability measures $\tilde{\alpha}$ on X and $\tilde{\beta}$ on Y, set*

$$\widetilde{\mathbb{X}} = (X, d_X, \tilde{\alpha}), \qquad \widetilde{\mathbb{Y}} = (Y, d_Y, \tilde{\beta}).$$

Then, deterministically,

$$\left|\mathcal{GW}(\widetilde{\mathbb{X}}, \widetilde{\mathbb{Y}}) - \mathcal{GW}(\mathbb{X}, \mathbb{Y})\right| \leqslant 2\,\mathcal{W}_p^X(\tilde{\alpha}, \alpha) + 2\,\mathcal{W}_p^Y(\tilde{\beta}, \beta),$$

where $\mathcal{W}_p^X$ and $\mathcal{W}_p^Y$ denote Wasserstein distances computed with d_X and d_Y. If $\hat{\alpha}_n$ and $\hat{\beta}_m$ are independent empirical measures built from iid samples with laws α and β, write $\widehat{\mathbb{X}}_n = (X, d_X, \hat{\alpha}_n)$ and $\widehat{\mathbb{Y}}_m = (Y, d_Y, \hat{\beta}_m)$. If, in addition, X and Y are regular compact subsets of Euclidean spaces of intrinsic dimensions d_X and d_Y, and the high-dimensional Wasserstein regime $d_X, d_Y > 2p$ applies, then

$$\mathbb{E}\left|\mathcal{GW}(\widehat{\mathbb{X}}_n, \widehat{\mathbb{Y}}_m) - \mathcal{GW}(\mathbb{X}, \mathbb{Y})\right| = O(n^{-1/d_X} + m^{-1/d_Y}).$$

For $p = 1$ on subsets of $[0,1]^d$, the rates, including the low-dimensional logarithmic cases, are those of Proposition 10.7.

Proof. The triangle inequality of Theorem 13.20 gives

$$\mathcal{GW}(\widetilde{\mathbb{X}}, \widetilde{\mathbb{Y}}) - \mathcal{GW}(\mathbb{X}, \mathbb{Y}) \leqslant \mathcal{GW}(\widetilde{\mathbb{X}}, \mathbb{X}) + \mathcal{GW}(\mathbb{Y}, \widetilde{\mathbb{Y}}).$$

Exchanging the two pairs gives the same bound for the opposite difference, hence the absolute value. The two terms on the right compare metric-measure spaces with the same underlying metric and only different measures, so Proposition 13.18 bounds them by $2\mathcal{W}_p^{\mathcal{X}}(\tilde{\alpha}, \alpha)$ and $2\mathcal{W}_p^{\mathcal{Y}}(\tilde{\beta}, \beta)$. Substituting $\tilde{\alpha} = \hat{\alpha}_n$ and $\tilde{\beta} = \hat{\beta}_m$, taking expectations and applying the empirical Wasserstein rates recalled in Section 10.2 gives the stated Euclidean rate. □

The metric structure also gives geodesics. Sturm's construction is useful conceptually because it makes it possible to discuss interpolation, barycenters and gradient flows directly on the space of metric-measure spaces, even though the intermediate space lives on a product support and is therefore expensive numerically [509].

Proposition 13.24 (Gromov–Wasserstein geodesics)**.** *Let $\mathbb{X}_0 = (\mathcal{X}_0, d_{\mathcal{X}_0}, \alpha_0)$ and $\mathbb{X}_1 = (\mathcal{X}_1, d_{\mathcal{X}_1}, \alpha_1)$ be compact metric-measure spaces, take $\Delta(u, v) = |u - v|$ in (13.9), and let $\pi^\star$ be an optimal coupling. Define, on $Z = \mathcal{X}_0 \times \mathcal{X}_1$,*

$$d_t((x_0, x_1), (x_0', x_1')) := (1-t) d_{\mathcal{X}_0}(x_0, x_0') + t\, d_{\mathcal{X}_1}(x_1, x_1'), \qquad \mathbb{X}_t = (Z, d_t, \pi^\star).$$

For $0 < t < 1$, d_t is a metric. At $t = 0$ and $t = 1$, one quotients Z by the zero-distance relation associated with d_t. Then $t \mapsto \mathbb{X}_t$ is a constant-speed geodesic:

$$\mathcal{GW}(\mathbb{X}_s, \mathbb{X}_t) = |t - s|\, \mathcal{GW}(\mathbb{X}_0, \mathbb{X}_1) \qquad \forall s, t \in [0,1].$$

Proof. Write $D = \mathcal{GW}(\mathbb{X}_0, \mathbb{X}_1)$. For $s < t$, couple $\mathbb{X}_s$ and $\mathbb{X}_t$ by the diagonal coupling induced by the identity on Z and the measure $\pi^\star$. For two independent points $z = (x_0, x_1)$ and $z' = (x_0', x_1')$ sampled from $\pi^\star$,

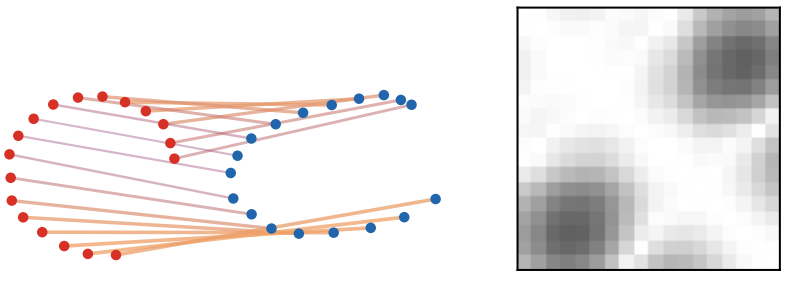
$$d_t(z, z') - d_s(z, z') = (t - s)\big(d_{\mathcal{X}_1}(x_1, x_1') - d_{\mathcal{X}_0}(x_0, x_0')\big).$$

Using this feasible coupling gives $\mathcal{GW}(\mathbb{X}_s, \mathbb{X}_t) \leqslant (t-s)D$. The same construction with the projections from Z to $\mathcal{X}_0$ and $\mathcal{X}_1$ gives $\mathcal{GW}(\mathbb{X}_0, \mathbb{X}_t) \leqslant tD$ and $\mathcal{GW}(\mathbb{X}_t, \mathbb{X}_1) \leqslant (1-t)D$. The triangle inequality for $\mathcal{GW}$ then yields, for $0 \leqslant s \leqslant t \leqslant 1$,

$$D \leqslant \mathcal{GW}(\mathbb{X}_0, \mathbb{X}_s) + \mathcal{GW}(\mathbb{X}_s, \mathbb{X}_t) + \mathcal{GW}(\mathbb{X}_t, \mathbb{X}_1) \leqslant sD + \mathcal{GW}(\mathbb{X}_s, \mathbb{X}_t) + (1-t)D,$$

so $\mathcal{GW}(\mathbb{X}_s, \mathbb{X}_t) \geqslant (t-s)D$. This proves equality. □

Figure 13.5 complements the global GW objective with local diagnostics, displaying where a mildly non-isometric correspondence creates the largest pairwise-distance residuals.



hard correspondence σ residual $|d_{\mathcal{X}} - d_{\mathcal{Y}} \circ \sigma|$

Figure 13.5: Local distortion in a mildly non-isometric GW match. The left panel colors transport segments by the average residual induced by the displayed hard correspondence. The right panel shows the pairwise-distance residual matrix $|d_{\mathcal{X}}(x_i, x_{i'}) - d_{\mathcal{Y}}(y_{\sigma(i)}, y_{\sigma(i')})|$ in white-to-black scale, with darker entries marking larger local distortion. This matrix is the local contribution minimized by the discrete GW objective for the displayed correspondence.

Distance-profile lower bound. Local distance distributions provide an inexpensive intrinsic certificate and a useful initialization principle for the non-convex solver.

Proposition 13.25 (Mémoli profile lower bound). *Let $\mathbb{X} = (X, d_X, \alpha)$ and $\mathbb{Y} = (Y, d_Y, \beta)$ be compact metric-measure spaces and take $\Delta(u, v) = |u - v|$ in (13.9), with the same exponent $p \geqslant 1$. For each $x \in X$ and $y \in Y$, define the distance-profile measures on $\mathbb{R}_+$ by*

$$\alpha_x := (d_X(x, \cdot))_\sharp \alpha, \qquad \beta_y := (d_Y(y, \cdot))_\sharp \beta.$$

Let $\mathfrak{D}_\mathbb{X} = (x \mapsto \alpha_x)_\sharp \alpha$ and $\mathfrak{D}_\mathbb{Y} = (y \mapsto \beta_y)_\sharp \beta$, which are probability measures on $\mathcal{P}_p(\mathbb{R}_+)$. Then

$$\mathbb{W}_p\big(\mathfrak{D}_\mathbb{X}, \mathfrak{D}_\mathbb{Y}\big) \leqslant \mathcal{GW}(\mathbb{X}, \mathbb{Y}).$$

Here $\mathbb{W}_p$ is the Wasserstein-over-Wasserstein distance (13.7): its ground space is $(\mathcal{P}_p(\mathbb{R}_+), \mathcal{W}_p)$.

Proof. Fix any $\pi \in \mathcal{U}(\alpha, \beta)$. It induces a coupling $(x, y) \mapsto (\alpha_x, \beta_y)$ between $\mathfrak{D}_\mathbb{X}$ and $\mathfrak{D}_\mathbb{Y}$, hence

$$\mathbb{W}_p\big(\mathfrak{D}_\mathbb{X}, \mathfrak{D}_\mathbb{Y}\big)^p \leqslant \int_{X \times Y} \mathcal{W}_p(\alpha_x, \beta_y)^p \, \mathrm{d}\pi(x, y).$$

For fixed (x, y), the map $(x', y') \mapsto (d_X(x, x'), d_Y(y, y'))$ pushes the same coupling π to a coupling between α_x and β_y. Therefore

$$\mathcal{W}_p(\alpha_x, \beta_y)^p \leqslant \int_{X \times Y} |d_X(x, x') - d_Y(y, y')|^p \, \mathrm{d}\pi(x', y').$$

Integrating in (x, y) gives

$$\mathbb{W}_p\big(\mathfrak{D}_\mathbb{X}, \mathfrak{D}_\mathbb{Y}\big)^p \leqslant \int_{X^2 \times Y^2} |d_X(x, x') - d_Y(y, y')|^p \, \mathrm{d}\pi(x, y) \mathrm{d}\pi(x', y').$$

Taking the infimum over π and then the p-th root proves the claim. □

Figure 13.6 makes the two transport levels explicit for the planar shapes $\mathbb{X}$ and $\mathbb{Y}$: sorting each distance profile computes the one-dimensional costs, then an outer assignment couples the resulting profile laws.

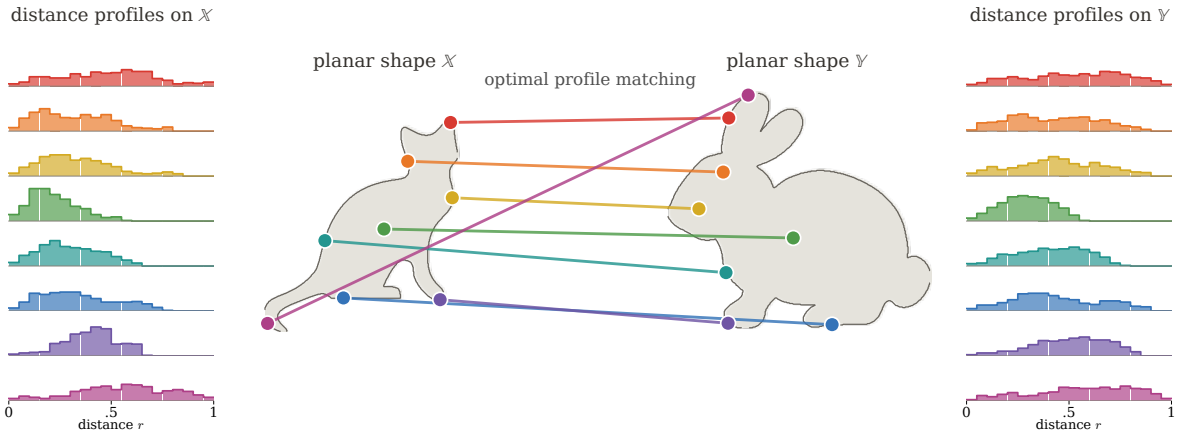


Figure 13.6: Mémoli distance profiles expose a computable lower bound for intrinsic GW comparison. The planar shapes $\mathbb{X}$ and $\mathbb{Y}$ are represented by cat and bunny silhouettes, centered and normalized to unit diameter. Matching colors identify representative anchor pairs selected by the optimal assignment for the profile cost $C_{i,j} = \mathcal{W}_2^2(\alpha_{x_i}, \beta_{y_j})$, their connecting segments, and the corresponding distance-profile histograms in the side columns. The histograms are display summaries only: the profile costs are computed by sorting the complete distance profiles. The resulting outer assignment realizes the Mémoli profile lower bound.

This lower bound is useful computationally because the profile cost matrix $C_{i,j} = \mathcal{W}_p(\alpha_{x_i}, \beta_{y_j})^p$ is an ordinary OT cost between points. Solving this easier OT problem gives a geometry-aware initialization for the non-convex GW iterations below, before the full pairwise-distance distortion is optimized.

The profile lower bound is intrinsic. In Euclidean applications, Proposition 13.18 pairs it with an extrinsic upper certificate obtained by registering the measures before applying ordinary Wasserstein transport. Combining the two propositions gives the extended Euclidean sandwich

$$\mathbb{W}_p(\mathfrak{D}_{\mathbb{X}}, \mathfrak{D}_{\mathbb{Y}}) \leqslant \mathcal{GW}(\mathbb{X}, \mathbb{Y}) \leqslant 2\, \mathcal{W}_{p,\mathrm{E}(d)}([\alpha], [\beta]) \leqslant 2\, \mathcal{W}_p(\alpha, \beta), \tag{13.15}$$

where $\mathbb{X} = (\mathbb{R}^d, \|\cdot\|, \alpha)$ and $\mathbb{Y} = (\mathbb{R}^d, \|\cdot\|, \beta)$. The profile term is an inexpensive intrinsic certificate, while the Wasserstein–Procrustes term gives an extrinsic certificate after optimal rigid registration. The final inequality simply omits registration and is therefore generally looser.

First-order optimality condition as self-consistent OT. Unlike linear OT, GW minimizes a quadratic functional over the convex set of couplings. Its first-order linearization is nevertheless an ordinary OT problem, which provides the basic stationarity condition used by essentially all GW solvers. The limitation is fundamental: solving this self-consistent linearized problem is necessary at a local minimum, but does not by itself certify either local or global optimality.

To isolate the relevant curvature, set $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ and

$$L\big((x, y), (x', y')\big) := \Delta\big(d_{\mathcal{X}}(x, x'), d_{\mathcal{Y}}(y, y')\big)^p.$$

For finite signed measures $\xi, \eta \in \mathcal{M}(\mathcal{Z})$ for which the integrals are absolutely convergent, define the symmetric bilinear form and its associated quadratic form by

$$\mathcal{B}_L(\xi, \eta) := \iint_{\mathcal{Z}^2} L(z, z')\mathrm{d}\xi(z)\mathrm{d}\eta(z'), \qquad \mathcal{Q}_L(\xi) := \mathcal{B}_L(\xi, \xi). \tag{13.16}$$

Thus (13.9) minimizes $\mathcal{Q}_L(\pi)$ over $\pi \in \mathcal{U}(\alpha, \beta)$. At a coupling π, define the linearized ground cost

$$c_\pi(z) := 2 \int_{\mathcal{Z}} L(z, z')\mathrm{d}\pi(z'). \tag{13.17}$$

For any $\widehat{\pi} \in \mathcal{U}(\alpha, \beta)$ and $\xi = \widehat{\pi} - \pi$, the exact quadratic expansion is

$$\mathcal{Q}_L(\pi + t\xi) = \mathcal{Q}_L(\pi) + t \int_{\mathcal{Z}} c_\pi \mathrm{d}\xi + t^2 \mathcal{Q}_L(\xi). \tag{13.18}$$

Definition 13.26 (First-order stationary GW coupling)**.** A coupling $\pi \in \mathcal{U}(\alpha, \beta)$ is *first-order stationary for GW* if it satisfies the variational inequality

$$\int_{\mathcal{Z}} c_\pi \mathrm{d}(\widehat{\pi} - \pi) \geqslant 0 \qquad \text{for every } \widehat{\pi} \in \mathcal{U}(\alpha, \beta). \tag{13.19}$$

This is stationarity for constrained minimization, not the unconstrained equation $c_\pi = 0$: in finite dimensions it is the usual variational inequality $-\nabla \mathcal{Q}_L(\pi) \in N_{\mathcal{U}(\alpha,\beta)}(\pi)$, with the normal-cone convention $N_K(x) = \{q : \langle q, y - x\rangle \leqslant 0 \ \forall y \in K\}$.

Proposition 13.27 (Frozen-OT characterization of stationarity)**.** *A coupling π is first-order stationary in the sense of Definition 13.26 if and only if*

$$\pi \in \operatorname{argmin}_{\widehat{\pi} \in \mathcal{U}(\alpha,\beta)} \int_{\mathcal{Z}} c_\pi \mathrm{d}\widehat{\pi}. \tag{13.20}$$

In particular, every local minimizer for the narrow topology, and hence every global minimizer, of $\mathcal{Q}_L$ is an optimal Kantorovich coupling for its own linearized cost c_π.

Proof. The variational inequality (13.19) is exactly the assertion that π minimizes the linear functional $\widehat{\pi} \mapsto \int c_\pi \mathrm{d}\widehat{\pi}$, proving the equivalence. If π is a local minimizer and $\widehat{\pi}$ is arbitrary, the segment $\pi_t = (1 - t)\pi + t\widehat{\pi}$ remains feasible. The right derivative at $t = 0$ in (13.18) must be nonnegative, which gives (13.19). □

The converse of the last implication in Proposition 13.27 is false: a coupling that solves its own frozen OT problem need not be a local, let alone global, minimizer of the quadratic GW objective. Frozen-OT self-consistency is exactly first-order stationarity. Even the concave quadratic $q(s) = -s^2$ on $[-1, 1]$ has the stationary point $s = 0$, which minimizes its constant linearization but is a strict maximum of q.

Conditional concavity of GW in Euclidean spaces. Self-consistency alone gives only a first-order condition. For several important Euclidean dissimilarities, however, the GW quadratic functional is concave on $\mathcal{U}(\alpha,\beta)$. This does not make its minimization convex: the problem remains non-convex and generally difficult. Concavity nevertheless turns the tangent model into a global majorizer. As formalized in Proposition 13.30, an exact frozen-OT step cannot increase the GW objective, and every optimizer of the frozen problem at a global GW solution is itself globally optimal. The same structure yields the learned-cost representation of Proposition 13.31: GW becomes an infimum of ordinary Kantorovich transport values plus a convex penalty on the ground cost. This is closely related to ground-cost robust Wasserstein transport, although the optimization over costs is here an infimum rather than an adversarial supremum. We first formulate the required sign condition abstractly on compact metric spaces and then verify it for Euclidean distance powers.

To make these statements precise, define the space of signed zero-marginal directions

$$\mathcal{T}_{\mathcal{X},\mathcal{Y}} := \left\{\xi \in \mathcal{M}(\mathcal{X}\times\mathcal{Y}) \;;\; (P_{\mathcal{X}})_\sharp \xi = 0,\ (P_{\mathcal{Y}})_\sharp \xi = 0\right\}. \tag{13.21}$$

The space $\mathcal{T}_{\mathcal{X},\mathcal{Y}}$ depends only on the two ambient spaces, not on the values of α and β: it is the direction space parallel to the affine marginal constraints. Every difference of two feasible couplings belongs to it. For finite spaces, identifying ξ with an $n\times m$ matrix Ξ gives

$$\mathcal{T}_{\mathcal{X},\mathcal{Y}} = \left\{\Xi \in \mathbb{R}^{n\times m} \;;\; \Xi \mathbb{1}_m = 0,\ \Xi^\top \mathbb{1}_n = 0\right\}, \qquad \dim \mathcal{T}_{\mathcal{X},\mathcal{Y}} = (m-1)(n-1).$$

Curvature of $\mathcal{Q}_L$ on the coupling set is determined entirely by its restriction to this space. Indeed, for any $\pi_0,\pi_1 \in \mathcal{U}(\alpha,\beta)$, the chord direction $\xi = \pi_1 - \pi_0$ belongs to $\mathcal{T}_{\mathcal{X},\mathcal{Y}}$, and along $\pi_t = (1-t)\pi_0 + t\pi_1$ one has

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2}\mathcal{Q}_L(\pi_t) = 2\mathcal{Q}_L(\xi).$$

The GW objective is generally highly non-convex, so one should not expect $\mathcal{Q}_L$ to be positive on $\mathcal{T}_{\mathcal{X},\mathcal{Y}}$. Perhaps surprisingly, in several cases of practical importance it has the opposite sign, $\mathcal{Q}_L(\xi) \leqslant 0$ for every $\xi \in \mathcal{T}_{\mathcal{X},\mathcal{Y}}$, and is therefore concave along every feasible chord. We refer to this property as *conditional concavity*.

The next result identifies a broad regime in which the sign of the quadratic form is known on the only directions that matter. Recall the conditionally positive and conditionally negative kernels of Definition 7.7.

Proposition 13.28 (Conditional concavity of squared GW)**.** *Let $\mathcal{X}$ and $\mathcal{Y}$ be compact metric spaces, and let $\alpha \in \mathcal{M}_+^1(\mathcal{X})$ and $\beta \in \mathcal{M}_+^1(\mathcal{Y})$. In (13.9), take $p = 2$ and $\Delta(u,v) = |u-v|$. Assume that the two distance kernels $d_{\mathcal{X}}$ and $d_{\mathcal{Y}}$ are either both conditionally positive definite or both conditionally negative definite in the sense of Definition 7.7. Then*

$$\mathcal{Q}_L(\xi) \leqslant 0 \qquad \text{for every } \xi \in \mathcal{T}_{\mathcal{X},\mathcal{Y}}. \tag{13.22}$$

Consequently, $\pi \mapsto \mathcal{Q}_L(\pi)$ is concave on $\mathcal{U}(\alpha,\beta)$.

Proof. The zero marginal conditions make the two pure terms in the squared loss vanish, and therefore

$$\mathcal{Q}_L(\xi) = -2\iint d_{\mathcal{X}}(x,x')d_{\mathcal{Y}}(y,y')\mathrm{d}\xi(x,y)\mathrm{d}\xi(x',y').$$

Choose $s \in \{-1,1\}$ so that $sd_{\mathcal{X}}$ and $sd_{\mathcal{Y}}$ are conditionally positive definite. Fix base points $x_0 \in \mathcal{X}$ and $y_0 \in \mathcal{Y}$, and define the centered kernels

$$\widetilde{d}_{\mathcal{X}}(x,x') = s\big(d_{\mathcal{X}}(x,x') - d_{\mathcal{X}}(x,x_0) - d_{\mathcal{X}}(x_0,x') + d_{\mathcal{X}}(x_0,x_0)\big),$$
$$\widetilde{d}_{\mathcal{Y}}(y,y') = s\big(d_{\mathcal{Y}}(y,y') - d_{\mathcal{Y}}(y,y_0) - d_{\mathcal{Y}}(y_0,y') + d_{\mathcal{Y}}(y_0,y_0)\big).$$

These kernels are positive definite. Since ξ has zero marginals and $s^2 = 1$, every centering term vanishes after integration. Let $\Phi_{\mathcal{X}}$ and $\Phi_{\mathcal{Y}}$ be Hilbert feature maps for $\widetilde{d}_{\mathcal{X}}$ and $\widetilde{d}_{\mathcal{Y}}$. Compactness and continuity make these maps bounded. Writing $z = (x,y)$ and $z' = (x',y')$, one obtains

$$\iint_{\mathcal{Z}^2} d_{\mathcal{X}}(x,x')d_{\mathcal{Y}}(y,y')\mathrm{d}\xi(z)\mathrm{d}\xi(z') = \Big\| \int_{\mathcal{Z}} \Phi_{\mathcal{X}}(x)\otimes\Phi_{\mathcal{Y}}(y)\mathrm{d}\xi(z)\Big\|^2 \geqslant 0.$$

This proves (13.22); the second-derivative identity above then proves concavity on every feasible segment. □

Conditional concavity has an immediate structural consequence: although minimizing a concave functional is not a convex problem, at least one minimizer lies at an extreme point of the feasible set. For GW this yields sparse couplings, and permutation couplings in the uniform square case.

Proposition 13.29 (Extreme-point and permutation optimizers in the concave regime)**.** *Assume the hypotheses of Proposition 13.28. Then the squared GW problem has an optimal coupling that is an extreme point of the convex set $\mathcal{U}(\alpha,\beta)$. If, in addition, α and β are supported on respectively n and m points, such an optimizer can be chosen with at most $n+m-1$ nonzero entries. If $n=m$ and both measures are uniform, it can be chosen as a permutation coupling P_σ/n.*

Proof. Set $K=\mathcal{U}(\alpha,\beta)$. Because $\mathcal{X}$ and $\mathcal{Y}$ are compact metric spaces, K is weakly compact, metrizable and convex in the locally convex space of finite signed measures on $\mathcal{X}\times\mathcal{Y}$. The kernel L is continuous and bounded, so $\mathcal{Q}_L$ is weakly continuous and admits a minimizer $\pi_\star\in K$. The Choquet representation theorem for metrizable compact convex sets [434] provides a probability measure ν concentrated on the extreme points of K whose barycenter is $\pi_\star$. Concavity of $\mathcal{Q}_L$ and minimality of $\pi_\star$ give

$$\mathcal{Q}_L(\pi_\star)\geqslant\int_K\mathcal{Q}_L(\eta)\mathrm{d}\nu(\eta)\geqslant\mathcal{Q}_L(\pi_\star).$$

Both inequalities are therefore equalities. Since $\mathcal{Q}_L(\eta)\geqslant\mathcal{Q}_L(\pi_\star)$ throughout K, equality of the integral implies that $\mathcal{Q}_L(\eta)=\mathcal{Q}_L(\pi_\star)$ for ν-almost every η. Because ν is concentrated on extreme couplings, at least one extreme coupling is optimal. This proves directly the instance of the Bauer maximum principle needed here.

For finitely supported measures, the coupling set is a transport polytope. Its marginal operator has rank $n+m-1$, as computed in the proof of Proposition 3.8. If an extreme coupling had more than $n+m-1$ positive entries, the restriction of this operator to its support would have a nonzero kernel direction H; for sufficiently small $t>0$, both $\mathrm{P}+tH$ and $\mathrm{P}-tH$ would be distinct feasible couplings with midpoint P, contradicting extremality. This proves the support bound. In the equal-size uniform case, Theorem 3.17 identifies the extreme points of the scaled Birkhoff polytope as P_σ/n. □

Conditional concavity strengthens the consequences of Proposition 13.27, but not its characterization: frozen-OT self-consistency is still only a first-order condition and need not imply local minimality. What concavity adds is a global tangent majorizer. It therefore gives descent after an exact frozen step and shows that, at a local minimizer, all frozen minimizers are flat replacements with the same GW value.

Proposition 13.30 (Tangent majorization and frozen-OT replacement)**.** *Assume that $\mathcal{Q}_L$ is concave on $\mathcal{U}(\alpha,\beta)$. Then, for every $\pi,\widehat{\pi}\in\mathcal{U}(\alpha,\beta)$,*

$$\mathcal{Q}_L(\widehat{\pi})\leqslant\mathcal{Q}_L(\pi)+\int_{\mathcal{Z}}c_\pi\mathrm{d}(\widehat{\pi}-\pi)=2\mathcal{B}_L(\pi,\widehat{\pi})-\mathcal{Q}_L(\pi).\tag{13.23}$$

Consequently:

1. *if $\pi^+\in\operatorname{argmin}_{\eta\in\mathcal{U}(\alpha,\beta)}\int c_\pi\mathrm{d}\eta$, then $\mathcal{Q}_L(\pi^+)\leqslant\mathcal{Q}_L(\pi)$;*
2. *if π is a local minimizer of $\mathcal{Q}_L$ and π^+ minimizes its frozen OT problem, then*
$$\mathcal{Q}_L\big((1-t)\pi+t\pi^+\big)=\mathcal{Q}_L(\pi)\qquad\text{for every } t\in[0,1];$$
3. *if $\pi_\star$ is globally GW-optimal, then every minimizer of the frozen OT problem with cost $c_{\pi_\star}$ is also globally GW-optimal.*

Proof. Apply (13.18) at $t=1$ with $\xi=\widehat{\pi}-\pi$. Concavity gives $\mathcal{Q}_L(\xi)\leqslant 0$, which proves (13.23). For the first conclusion, frozen optimality gives $\int c_\pi\mathrm{d}(\pi^+-\pi)\leqslant 0$.

Now assume that π is a local minimizer. Proposition 13.27 shows that π itself minimizes its frozen objective. Hence every other frozen minimizer π^+ satisfies $\int c_\pi\mathrm{d}(\pi^+-\pi)=0$. With $\xi=\pi^+-\pi$, the exact expansion (13.18) becomes

$$\mathcal{Q}_L(\pi+t\xi)=\mathcal{Q}_L(\pi)+t^2\mathcal{Q}_L(\xi).$$

Local minimality for sufficiently small $t>0$ gives $\mathcal{Q}_L(\xi)\geqslant 0$, while concavity gives the reverse inequality. Thus $\mathcal{Q}_L(\xi)=0$, proving the second conclusion for every $t\in[0,1]$. If $\pi=\pi_\star$ is globally optimal, the common value is the global minimum, which proves the third conclusion. □

Thus freezing one occurrence of the coupling produces a global upper bound, not merely a formal linearization. The second conclusion is a flat-replacement theorem for a coupling already known to be locally minimizing; it is not a converse to Proposition 13.27. A generic stationary point can still be a maximum or a non-global local minimum.

Schoenberg's theorem gives the principal Euclidean example: on compact subsets of Euclidean spaces, $d_{\mathcal{X}}(x,x') = \|x-x'\|^q$ and $d_{\mathcal{Y}}(y,y') = \|y-y'\|^q$ are conditionally negative definite for $0 < q \leqslant 2$ [479]. Proposition 13.28 therefore covers the usual squared GW distortion between Euclidean distances when $q = 1$. Although $q > 1$ no longer defines a metric, the proof only uses conditional definiteness and applies unchanged to these dissimilarities, including squared Euclidean dissimilarities when $q = 2$. The conditional-concavity mechanism and its graph-matching and GW consequences are developed in [369, 484, 534].

GW as cost-regularized robust Wasserstein transport. Conditional concavity also turns GW into an ordinary Kantorovich problem with a learned ground cost. The cost is not chosen freely: a convex penalty records the original quadratic distortion. This viewpoint, introduced for ground-cost adversarial transport in [420] and developed for cross-space structured transforms in [482], separates the linear transport step from the nonlinear choice of geometry.

For this paragraph, extend $\mathcal{L}_c(\alpha,\beta)$ by the same formula as (3.7) to bounded continuous costs of arbitrary sign. On compact spaces this is harmless: adding a constant to c makes it nonnegative and only adds that constant to the transport value.

Proposition 13.31 (Concave coupling energies as cost-regularized OT)**.** *Let $\mathcal{X}$ and $\mathcal{Y}$ be compact, set $\mathcal{Z} = \mathcal{X}\times\mathcal{Y}$, and let $Q : \mathcal{U}(\alpha,\beta) \to \mathbb{R}\cup\{-\infty\}$ be proper, upper semicontinuous and concave. Let $\mathfrak{F}_Q$ be any proper convex weak-* lower-semicontinuous extension of $-Q$ from $\mathcal{U}(\alpha,\beta)$ to finite signed measures on $\mathcal{Z}$. Such an extension always exists: one possible choice is the canonical constrained extension*

$$\mathfrak{F}_Q^{\mathrm{can}}(\xi) := \begin{cases} -Q(\xi), & \xi \in \mathcal{U}(\alpha,\beta), \\ +\infty, & \text{otherwise}, \end{cases}.$$

Define the convex conjugate of the chosen extension by

$$\mathfrak{F}_Q^*(h) := \sup_{\xi\in\mathcal{M}(\mathcal{Z})} \left\{ \int_{\mathcal{Z}} h\,\mathrm{d}\xi - \mathfrak{F}_Q(\xi) \right\}, \qquad h \in C(\mathcal{Z}).$$

Then the convex cost penalty $\mathfrak{R}_Q(c) := \mathfrak{F}_Q^*(-c)$ satisfies

$$\inf_{\pi\in\mathcal{U}(\alpha,\beta)} Q(\pi) = \inf_{c\in C(\mathcal{X}\times\mathcal{Y})} \left\{ \mathcal{L}_c(\alpha,\beta) + \mathfrak{R}_Q(c) \right\}. \tag{13.24}$$

Proof. Fenchel–Moreau gives, for every feasible π,

$$Q(\pi) = \inf_{c\in C(\mathcal{Z})} \left\{ \int_{\mathcal{Z}} c\,\mathrm{d}\pi + \mathfrak{F}_Q^*(-c) \right\}.$$

Taking the infimum in π and exchanging the two infima yields (13.24); the inner infimum in π is exactly $\mathcal{L}_c(\alpha,\beta)$. □

Remark 13.32 (Toland duality and DC algorithms). When Q is finite on the coupling set, as for the bounded continuous GW kernels considered here, Proposition 13.31 is exactly Toland–Singer duality specialized to measures. Write $\mathcal{K} = \mathcal{U}(\alpha,\beta)$, and let $\iota_{\mathcal{K}}$ be its convex indicator. Since $\mathfrak{F}_Q = -Q$ on $\mathcal{K}$, the original problem is the difference of the two convex functionals $\iota_{\mathcal{K}}$ and $\mathfrak{F}_Q$. Evaluating this difference on the effective domain of $\iota_{\mathcal{K}}$, Toland duality reads

$$\inf_{\xi\in\mathcal{K}} Q(\xi) = \inf_{\xi\in\mathcal{K}} \left\{ \iota_{\mathcal{K}}(\xi) - \mathfrak{F}_Q(\xi) \right\} = \inf_{h\in C(\mathcal{Z})} \left\{ \mathfrak{F}_Q^*(h) - \iota_{\mathcal{K}}^*(h) \right\}. \tag{13.25}$$

Setting $h = -c$ and observing that $\iota_{\mathcal{K}}^*(-c) = \sup_{\pi\in\mathcal{K}}(-\int c\,\mathrm{d}\pi) = -\mathcal{L}_c(\alpha,\beta)$ turns the last problem into (13.24). Toland introduced this non-convex conjugate duality in [519]; Carlier developed its relation to convexity constraints and optimal transport in [104].

The same decomposition identifies the associated algorithm. For $Q = Q_L$, choosing $\mathfrak{F}_Q = \mathfrak{F}_Q^{\mathrm{can}}$ ensures that the linearized cost (13.17) satisfies $-c_{\pi^{(\ell)}} \in \partial\mathfrak{F}_Q^{\mathrm{can}}(\pi^{(\ell)})$. Applying the classical difference-of-convex

algorithm (DCA) to $(\iota_{\mathcal{K}} + \mathcal{R}_\varepsilon) - \mathfrak{F}_Q$ therefore gives the successive-linearization step

$$\pi^{(\ell+1)} \in \operatorname{argmin}_{\pi\in\mathcal{K}} \left\{ \int c_{\pi^{(\ell)}} \mathrm{d}\pi + \mathcal{R}_\varepsilon(\pi) \right\}. \tag{13.26}$$

For $\varepsilon = 0$, this is exactly the frozen-OT step of Proposition 13.30, whose descent inequality is the corresponding DCA descent property. More generally, in the conditionally concave regime, interleaving the two block updates in (13.46) produces (13.26); the displayed two-block algorithm merely groups two consecutive DCA steps. This classical method and its variants are developed in [433, 330]. Outside the conditionally concave regime, $\mathfrak{F}_Q$ need not be convex. The same block updates remain valid alternating-minimization steps for the biconvex lift, but they are no longer DCA steps for the diagonal GW objective and need not decrease it.

Applied to $Q = Q_L|_{\mathcal{U}(\alpha,\beta)}$, Proposition 13.31 is an exact representation of GW whenever the conditional-concavity hypothesis of Proposition 13.28 holds. Without concavity, a convex extension of $-Q$ need not exist. Biconjugating the canonical constrained functional instead replaces Q by its upper-semicontinuous concave envelope, so it generally changes both the value and the minimizers. The penalized infimum over Wasserstein costs in (13.24) should be contrasted with the robust representation of Section 11.5, especially Proposition 11.43, which takes a supremum of Wasserstein distances.

In the squared-GW setting of Proposition 13.28, the quadratic structure gives this cost penalty a Hilbertian interpretation. Its proof constructs a tensor-product feature map $\Phi(x,y) = \Phi_{\mathcal{X}}(x) \otimes \Phi_{\mathcal{Y}}(y)$ in a Hilbert space $\mathcal{H}_{\mathcal{Z}}$. After collecting the terms affine in π into a fixed cost c_0, one can write

$$Q_L(\pi) = C(\alpha,\beta) + \int_{\mathcal{Z}} c_0 \mathrm{d}\pi - 2\|m_\pi\|^2_{\mathcal{H}_{\mathcal{Z}}}, \qquad m_\pi := \int_{\mathcal{Z}} \Phi(z)\mathrm{d}\pi(z).$$

Completing the square therefore gives the equivalent learned-cost representation

$$Q_L(\pi) = C(\alpha,\beta) + \inf_{c\in c_0+\mathcal{H}_{\mathcal{Z}}} \left\{ \int_{\mathcal{Z}} c\mathrm{d}\pi + \frac{1}{8}\|c-c_0\|^2_{\mathcal{H}_{\mathcal{Z}}} \right\}. \tag{13.27}$$

Thus, after quotienting out separable costs that integrate to constants on $\mathcal{U}(\alpha,\beta)$, the quadratic extension in Proposition 13.31 makes $\mathfrak{R}_Q$, up to the additive marginal constant, a squared RKHS norm, with value $+\infty$ outside the admissible affine cost space. This is the cost-side counterpart of the kernel norms and mean embeddings of Section 7.2 [44, 482]. On Euclidean domains, Riesz and Matérn constructions identify such RKHSs with homogeneous Sobolev or Bessel-potential spaces, so the penalty imposes smoothness on the learned ground cost [43, 549].

For the bilinear and squared-distance Euclidean dissimilarities considered next, this RKHS is finite dimensional: its features are low-degree polynomials, and the map $\pi \mapsto c_\pi$ remains in a fixed finite-dimensional affine family. This is a finite-rank, hence comparatively degenerate, instance of the preceding construction because the learned cost loses the expressivity of a generic infinite-dimensional RKHS. Its advantage is that both the theory and the numerical optimization reduce to matrix algebra. This reduction appears in Vayer's analysis of GW on incomparable Euclidean spaces [533, Chapter 4]; its cost-regularized interpretation is developed systematically in [482]. The next normalization gives equality of objective values; formulations that only seek the same coupling minimizers may absorb positive factors and marginal-only constants into the objective.

Proposition 13.33 (Finite-dimensional cost learning for Euclidean GW)**.** *Let $\mathcal{X} \subset \mathbb{R}^{d_x}$ and $\mathcal{Y} \subset \mathbb{R}^{d_y}$ be compact, and use $p = 2$ and $\Delta(u,v) = |u-v|$ throughout. Retain the notation $\mathcal{GW}$ of (13.9) when the within-space functions are symmetric dissimilarities rather than metrics.*

For each of the two choices below, set $\mathbb{X} = (\mathcal{X}, d_{\mathcal{X}}, \alpha)$ and $\mathbb{Y} = (\mathcal{Y}, d_{\mathcal{Y}}, \beta)$. There is a constant $C(\alpha,\beta)$, depending only on the marginals and on the chosen dissimilarities, such that

$$\mathcal{GW}(\mathbb{X},\mathbb{Y})^2 = C(\alpha,\beta) + \inf_{M\in\mathbb{R}^{d_y\times d_x}} \left\{ \mathcal{L}_{c_M}(\alpha,\beta) + \frac{1}{2}\|M\|^2_{\mathrm{F}} \right\}. \tag{13.28}$$

The constant $C(\alpha,\beta)$ is different in the two cases, while the learned cost is as follows.

1. *For the inner-product dissimilarities*

$$d_{\mathcal{X}}(x,x') = -\langle x, x'\rangle, \qquad d_{\mathcal{Y}}(y,y') = -\langle y, y'\rangle, \tag{13.29}$$

the learned costs form the linear family

$$c_M(x, y) = -2\langle Mx,\, y\rangle.$$

2. *For the squared-distance dissimilarities*

$$d_{\mathcal{X}}(x, x') = \|x - x'\|^2, \qquad d_{\mathcal{Y}}(y, y') = \|y - y'\|^2, \tag{13.30}$$

translate α and β independently so that both have zero mean, which leaves $\mathcal{GW}$ unchanged. The learned costs then form the affine family

$$c_M(x, y) = -4\|x\|^2\|y\|^2 - 4\langle Mx,\, y\rangle.$$

Proof. For $\pi \in \mathcal{U}(\alpha, \beta)$, write $S_\pi := \int yx^\top \mathrm{d}\pi(x, y)$. For two independent pairs $(x, y), (x', y') \sim \pi$,

$$\iint \langle x,\, x'\rangle\langle y,\, y'\rangle \mathrm{d}\pi\mathrm{d}\pi = \|S_\pi\|_{\mathrm{F}}^2.$$

For (13.29), expanding the square therefore gives $\mathcal{Q}_L(\pi) = C(\alpha, \beta) - 2\|S_\pi\|_{\mathrm{F}}^2$. For fixed π, completing the square gives

$$\inf_M \left\{-2\langle M,\, S_\pi\rangle_{\mathrm{F}} + \frac{1}{2}\|M\|_{\mathrm{F}}^2\right\} = -2\|S_\pi\|_{\mathrm{F}}^2, \qquad M = 2S_\pi,$$

which proves (13.28) in the first case after minimizing over π.

For centered marginals, direct expansion yields

$$\iint \|x - x'\|^2\|y - y'\|^2 \mathrm{d}\pi\mathrm{d}\pi = 2\int \|x\|^2\|y\|^2 \mathrm{d}\pi + 2\left(\int_{\mathcal{X}} \|x\|^2 \mathrm{d}\alpha(x)\right)\left(\int_{\mathcal{Y}} \|y\|^2 \mathrm{d}\beta(y)\right) + 4\|S_\pi\|_{\mathrm{F}}^2.$$

Substitution into the squared-distance plan objective gives

$$\mathcal{Q}_L(\pi) = C(\alpha, \beta) - 4\int \|x\|^2\|y\|^2 \mathrm{d}\pi(x, y) - 8\|S_\pi\|_{\mathrm{F}}^2. \tag{13.31}$$

Finally,

$$\inf_M \left\{-4\langle M,\, S_\pi\rangle_{\mathrm{F}} + \frac{1}{2}\|M\|_{\mathrm{F}}^2\right\} = -8\|S_\pi\|_{\mathrm{F}}^2, \qquad M = 4S_\pi,$$

which proves (13.28) in the second case. □

In both Euclidean models, Formula (13.28) shows that the GW loss can be understood as learning the finite-dimensional cost parameter M while solving an ordinary OT problem with ground cost c_M. The alternating algorithm developed below follows exactly this interpretation: it alternates between updating the transport coupling and updating M from that coupling. Eliminating the coupling leaves the objective $M \mapsto \mathcal{L}_{c_M}(\alpha, \beta) + \frac{1}{2}\|M\|_{\mathrm{F}}^2$. It is generally non-convex—indeed, the first term is concave in M, whereas the second is convex—so the finite-dimensional reduction simplifies each block without removing non-global stationary points.

GW between Gaussian measures. Gaussian inputs expose a sharp difference between the two Euclidean models. GW with the inner-product dissimilarities (13.29) is closed under Gaussian couplings: Salmona, Delon and Desolneux prove this for centered inputs, and Dandapanthula et al. extend it to arbitrary means [465, 175]. For the squared-distance dissimilarities (13.30), by contrast, the available closed form solves the problem *constrained* to jointly Gaussian couplings; whether this constraint changes the optimum remains open in general [465]. The next result records both formulas with the normalization of Proposition 13.33.

Let

$$\alpha = \mathcal{N}(m_\alpha, \Sigma_\alpha) \in \mathcal{P}_2(\mathbb{R}^{d_x}), \qquad \beta = \mathcal{N}(m_\beta, \Sigma_\beta) \in \mathcal{P}_2(\mathbb{R}^{d_y}),$$

and write $\lambda_1 \geqslant \cdots \geqslant \lambda_{d_x} \geqslant 0$ and $\eta_1 \geqslant \cdots \geqslant \eta_{d_y} \geqslant 0$ for the eigenvalues of Σ_α and Σ_β. Pad both lists by zeros up to $D = \max(d_x, d_y)$. For each dissimilarity below, write $\mathbb{X} = (\mathbb{R}^{d_x}, d_{\mathcal{X}}, \alpha)$ and $\mathbb{Y} = (\mathbb{R}^{d_y}, d_{\mathcal{Y}}, \beta)$.

Proposition 13.34 (Exact inner-product and Gaussian-constrained squared GW)**.** *For the inner-product dissimilarities* (13.29)*, the unrestricted problem admits a jointly Gaussian minimizer for arbitrary m_α and m_β. If both means vanish, then*

$$\mathcal{GW}(\mathbb{X},\mathbb{Y})^2 = \sum_{i=1}^{D}(\lambda_i-\eta_i)^2. \tag{13.32}$$

For the squared-distance dissimilarities (13.30)*, define the Gaussian-constrained squared distortion*

$$\mathcal{GW}^{\mathrm{Gau}}(\mathbb{X},\mathbb{Y})^2 := \inf\{\mathcal{Q}_L(\pi) : \pi\in\mathcal{U}(\alpha,\beta), \pi \text{ is jointly Gaussian}\}. \tag{13.33}$$

Then

$$\mathcal{GW}^{\mathrm{Gau}}(\mathbb{X},\mathbb{Y})^2 = 4\left(\sum_{i=1}^{D}\lambda_i-\sum_{i=1}^{D}\eta_i\right)^2 + 8\sum_{i=1}^{D}(\lambda_i-\eta_i)^2. \tag{13.34}$$

If $d_x \geqslant d_y$ and $\Sigma_\alpha \succ 0$, a minimizer in (13.33) *is induced by a deterministic affine map that aligns the principal covariance directions, rescales the first d_y coordinates and discards the remaining ones.*

Proof. For a coupling π, set

$$K_\pi := \int (y-m_\beta)(x-m_\alpha)^\top \mathrm{d}\pi(x,y).$$

Its block covariance is positive semidefinite, $\left(\begin{smallmatrix}\Sigma_\alpha & K_\pi^\top\\ K_\pi & \Sigma_\beta\end{smallmatrix}\right)\succeq 0$. Conversely, every matrix K satisfying this constraint is the cross-covariance of a jointly Gaussian coupling. The expansion in the proof of Proposition 13.33 shows that the inner-product objective $\mathcal{Q}_L(\pi)$ depends on π only through the cross-moment $\int yx^\top \mathrm{d}\pi = m_\beta m_\alpha^\top + K_\pi$. Replacing any coupling by this Gaussian coupling therefore leaves the objective unchanged. This proves Gaussian optimality, including for nonzero means.

When the means vanish, the positive-semidefinite block constraint is equivalent to the exact factorization $K = \Sigma_\beta^{1/2} R^\top \Sigma_\alpha^{1/2}$ with $R\in\mathbb{R}^{d_x\times d_y}$ and $\|R\|_{\mathrm{op}}\leqslant 1$; this remains valid for singular covariances. Von Neumann's trace inequality then gives

$$\max_K \|K\|_{\mathrm{F}}^2 = \sum_{i=1}^{\min(d_x,d_y)}\lambda_i\eta_i.$$

Substitution into the inner-product instance of (13.28) yields (13.32).

For a centered jointly Gaussian pair (X,Y) with cross-covariance K, Isserlis' formula gives

$$\mathcal{Q}_L(\pi) = 4\big(\operatorname{tr}\Sigma_\alpha - \operatorname{tr}\Sigma_\beta\big)^2 + 8\left(\operatorname{tr}\Sigma_\alpha^2 + \operatorname{tr}\Sigma_\beta^2 - 2\|K\|_{\mathrm{F}}^2\right).$$

Squared distances are translation invariant, so the same expression applies for arbitrary means. Maximizing $\|K\|_{\mathrm{F}}^2$ as above proves (13.34). More explicitly, write $\Sigma_\alpha = U\operatorname{diag}(\lambda_i)U^\top$ and $\Sigma_\beta = V\operatorname{diag}(\eta_j)V^\top$. For arbitrary signs $s_j\in\{-1,1\}$, the map

$$T(x) = m_\beta + V\left[\operatorname{diag}\left(s_j\sqrt{\eta_j/\lambda_j}\right)_{j=1}^{d_y} \quad 0\right]U^\top(x-m_\alpha)$$

pushes α to β and attains the maximizing cross-covariance. This is the claimed principal-component alignment. □

For the squared-distance dissimilarities (13.30), the unrestricted problem is subtler because (13.31) also contains the mixed fourth moment $\int \|x\|^2\|y\|^2\mathrm{d}\pi$, which is not fixed by the cross-moment $\int yx^\top\mathrm{d}\pi$ outside the Gaussian class. Equivalently, the frozen cost contains the non-quadratic radial product $-\|x\|^2\|y\|^2$. Hence $\mathcal{GW}(\mathbb{X},\mathbb{Y})^2 \leqslant \mathcal{GW}^{\mathrm{Gau}}(\mathbb{X},\mathbb{Y})^2$, and the next proposition shows that this inequality can be strict.

Proposition 13.35 (Failure of Gaussian closure for squared-distance GW)**.** *For all sufficiently large integers n, consider*

$$\alpha = \mathcal{N}(0,1) \quad \text{on } \mathbb{R}, \qquad \beta_n = \mathcal{N}(0,\mathrm{Id}_n) \quad \text{on } \mathbb{R}^n,$$

with the squared-distance dissimilarities (13.30)*, and write $\mathbb{X}_n = (\mathbb{R}, d_\mathcal{X}, \alpha)$ and $\mathbb{Y}_n = (\mathbb{R}^n, d_\mathcal{Y}, \beta_n)$. No jointly Gaussian coupling is optimal for the unrestricted GW problem between α and β_n. In particular,*

$$\mathcal{GW}(\mathbb{X}_n,\mathbb{Y}_n)^2 < \mathcal{GW}^{\mathrm{Gau}}(\mathbb{X}_n,\mathbb{Y}_n)^2.$$

Proof. For $(X,Y) \sim \pi$, equation (13.31) shows that minimizing the GW objective is equivalent to maximizing

$$\mathcal{G}(\pi) := \mathbb{E}[X^2\|Y\|^2] + 2\|\mathbb{E}[YX]\|^2.$$

If (X,Y) is jointly Gaussian, write $k = \mathbb{E}[YX] \in \mathbb{R}^n$. Positivity of its block covariance gives $\|k\|^2 \leqslant 1$, while Isserlis' formula gives

$$\mathbb{E}[X^2\|Y\|^2] = n + 2\|k\|^2.$$

Consequently, every jointly Gaussian coupling satisfies $\mathcal{G}(\pi) = n + 4\|k\|^2 \leqslant n+4$.

We construct a better non-Gaussian coupling. Let $Y \sim \mathcal{N}(0, \mathrm{Id}_n)$, set $R = \|Y\|^2$, and denote by G_j the cumulative distribution function of a χ^2_j variable. Then $U = G_n(R)$ is uniform on $(0,1)$ and $A = G_1^{-1}(U)$ has law χ^2_1. If ε is an independent Rademacher variable and $X = \varepsilon\sqrt{A}$, then $X \sim \mathcal{N}(0,1)$ and $\mathbb{E}[YX] = 0$. This coupling therefore satisfies

$$\mathcal{G}(\pi) = \mathbb{E}[AR] = \int_0^1 G_1^{-1}(u) G_n^{-1}(u)\,\mathrm{d}u.$$

The central limit theorem gives weak convergence of the standardized χ^2_n variable to a standard Gaussian. Since their second moments also converge, this convergence holds in $\mathcal{W}_2$; the quantile formula of Proposition 2.44 therefore gives

$$\frac{G_n^{-1}(u) - n}{\sqrt{2n}} \longrightarrow \Phi^{-1}(u) \qquad \text{in } L^2(0,1),$$

where Φ is the standard Gaussian cumulative distribution function. Since $G_1^{-1} \in L^2(0,1)$,

$$\mathcal{G}(\pi) = n + \sqrt{2n}\big(c + o(1)\big), \qquad c := \int_0^1 G_1^{-1}(u)\Phi^{-1}(u)\,\mathrm{d}u > 0.$$

Here c is the covariance of the two strictly increasing, nonconstant functions G_1^{-1} and Φ^{-1} of a uniform random variable, so $c > 0$. Thus $\mathcal{G}(\pi) > n+4$ for all sufficiently large n, which is strictly better than every jointly Gaussian coupling. □

Salmona, Delon and Desolneux derive explicit lower and upper bounds for the unrestricted squared-distance problem. They prove equality with the Gaussian-constrained value when $d_x = d_y$, $\Sigma_\alpha \succ 0$, and the ordered covariance spectra are proportional, $\eta_i = s\lambda_i$ for some $s \geqslant 0$; this includes all nondegenerate one-dimensional Gaussian pairs, for which the exact value is $12(\lambda_1 - \eta_1)^2$ [465, Theorem 4.1, Proposition 5.2 and Corollary 5.1]. Proposition 13.35 disproves their general Gaussian-optimality conjecture [465, Section 6]: the non-quadratic radial term can favor dependence that is invisible to covariance optimization. Their Gaussian-constrained affine optimizer is nevertheless the exact PCA alignment in Proposition 13.34 and extends Vayer's linear Gromov–Monge calculation [533, Theorem 4.2.6]. For centered Gaussian inputs, the entropy-regularized problem with (13.29) also has a jointly Gaussian optimizer and a closed form; the KL-unbalanced extension remains Gaussian and reduces to scalar polynomial equations [327].

Monge–GW maps. A GW optimizer need not generally be induced by a map. Restricting the admissible couplings to graphs gives the Gromov–Monge problem.

Definition 13.36 (Gromov–Monge problem)**.** Fix $p \geqslant 1$ and the discrepancy Δ used in Definition 13.15. The directed Gromov–Monge value is

$$\mathcal{GW}_p^{\mathrm{M}}(\mathbb{X},\mathbb{Y}) := \inf_{\substack{T:\mathcal{X}\to\mathcal{Y}\ \text{measurable}\\ T_\sharp\alpha=\beta}} \left(\iint_{\mathcal{X}^2} \Delta\big(d_{\mathcal{X}}(x,x'), d_{\mathcal{Y}}(T(x),T(x'))\big)^p \mathrm{d}\alpha(x)\mathrm{d}\alpha(x')\right)^{1/p}, \tag{13.35}$$

with value $+\infty$ when no such measure-preserving map exists.

Its infinite-order extension is

$$\mathcal{GW}_\infty^{\mathrm{M}}(\mathbb{X},\mathbb{Y}) := \inf_{\substack{T:\mathcal{X}\to\mathcal{Y}\ \text{measurable}\\ T_\sharp\alpha=\beta}} \operatorname*{ess\,sup}_{\alpha\otimes\alpha} \Delta\big(d_{\mathcal{X}}(x,x'), d_{\mathcal{Y}}(T(x),T(x'))\big). \tag{13.36}$$

These quantities are directed and can be infinite because a measure-preserving map need not exist. For $1 \leqslant p \leqslant q < \infty$, monotonicity of L^p norms and the restriction from couplings to graph couplings give

$$\begin{aligned} &\mathcal{GW}_p^{\mathrm{M}}(\mathbb{X},\mathbb{Y}) \leqslant \mathcal{GW}_q^{\mathrm{M}}(\mathbb{X},\mathbb{Y}) \leqslant \mathcal{GW}_\infty^{\mathrm{M}}(\mathbb{X},\mathbb{Y}), \\ &\mathcal{GW}_p(\mathbb{X},\mathbb{Y}) \leqslant \mathcal{GW}_p^{\mathrm{M}}(\mathbb{X},\mathbb{Y}), \qquad \mathcal{GW}_\infty(\mathbb{X},\mathbb{Y}) \leqslant \mathcal{GW}_\infty^{\mathrm{M}}(\mathbb{X},\mathbb{Y}). \end{aligned} \tag{13.37}$$

Unlike its coupling counterpart in Proposition 13.17, the limit of $\mathcal{GW}_p^{\mathrm{M}}$ need not commute with the infimum over maps without additional compactness. It does commute in the basic finite uniform setting.

Proposition 13.37 (Finite uniform Gromov–Monge limits)**.** *Let $\mathcal{X} = \{x_i\}_{i=1}^n$ and $\mathcal{Y} = \{y_j\}_{j=1}^n$ carry uniform measures. Then*

$$\mathcal{GW}_p^{\mathrm{M}}(\mathbb{X},\mathbb{Y}) = \min_{\sigma\in\mathfrak{S}_n} \left(\frac{1}{n^2} \sum_{i,i'=1}^n \Delta\big(d_{\mathcal{X}}(x_i,x_{i'}), d_{\mathcal{Y}}(y_{\sigma(i)},y_{\sigma(i')})\big)^p \right)^{1/p}, \tag{13.38}$$

$$\lim_{p\to\infty} \mathcal{GW}_p^{\mathrm{M}}(\mathbb{X},\mathbb{Y}) = \mathcal{GW}_\infty^{\mathrm{M}}(\mathbb{X},\mathbb{Y}) = \mathcal{GW}_\infty(\mathbb{X},\mathbb{Y}). \tag{13.39}$$

Proof. A measure-preserving map between equal-size uniform spaces is a permutation; this proves (13.38). Since $\mathfrak{S}_n$ is finite, the minimum commutes with $p \to \infty$. This gives the first equality in (13.39).

For the last equality, let P_∞ minimize $\mathcal{GW}_\infty$. The scaled matrix $n\mathrm{P}_\infty$ is doubly stochastic, so Theorem 3.17 provides a permutation whose graph is contained in $\operatorname{supp}(\mathrm{P}_\infty)$. Its distortion cannot exceed the maximal distortion on $\operatorname{supp}(\mathrm{P}_\infty)^2$, and hence $\mathcal{GW}_\infty^{\mathrm{M}} \leqslant \mathcal{GW}_\infty$. The reverse inequality is (13.37). □

For the inner-product dissimilarities (13.29), the learned-cost representation yields a Monge optimizer; the stronger transformed-Brenier description obtained from the twist condition has a precise rank limitation. Vayer first proved this description under surjectivity of the optimal cross-moment matrix [533, Theorem 4.2.3]. Dumont, Lacombe and Vialard subsequently removed the rank assumption for existence of a Monge optimizer [193].

Proposition 13.38 (Monge–GW map for inner-product dissimilarities)**.** *Let $\mathcal{X} \subset \mathbb{R}^{d_x}$ and $\mathcal{Y} \subset \mathbb{R}^{d_y}$ be compact, assume $d_x \geqslant d_y$, and let α be absolutely continuous with respect to Lebesgue measure. Then the GW problem with dissimilarities* (13.29) *admits an optimal coupling of the form $(\mathrm{Id}, T)_\sharp\alpha$. In the inner-product instance of* (13.28)*, let $M^\star$ be an optimal learned matrix. If $\operatorname{rank}(M^\star) = d_y$, then $c_{M^\star}$ satisfies the twist condition of Definition 2.35; the optimal coupling for this fixed cost is unique and*

$$T(x) = \nabla\varphi(M^\star x) \tag{13.40}$$

for the Brenier potential φ transporting $(M^\star)_\sharp\alpha$ to β.

Proof. Compactness of the coupling set and coercivity of the quadratic penalty in M give an optimal pair $(\pi^\star, M^\star)$ in the inner-product instance of (13.28). With $M^\star$ frozen, $\pi^\star$ solves linear OT for $c_{M^\star}$. The linear-cost Monge-existence theorem of Dumont, Lacombe and Vialard [193], also used in the cost-regularized setting in [482], gives a graph-supported optimizer for every matrix $M^\star$ under the stated dimension and absolute-continuity assumptions. Its rank-deficient case combines transport on the image of $M^\star$ with measurable transport along its fibers. Replacing $\pi^\star$ by this optimizer leaves the joint objective unchanged.

If $M^\star$ has full row rank, then

$$y \longmapsto \nabla_x c_{M^\star}(x,y) = -2(M^\star)^\top y$$

is injective, so the cost is twisted. Moreover, $(M^\star)_\sharp\alpha$ is absolutely continuous on $\mathbb{R}^{d_y}$, and minimizing $-2\langle z,\, y\rangle$ is equivalent, up to marginal-only terms, to quadratic transport between $z = M^\star x$ and y. Brenier's theorem, Theorem 2.23, therefore gives the unique map $z \mapsto \nabla\varphi(z)$, proving (13.40). This is precisely the full-rank construction of [533, Theorem 4.2.3]. □

If $M^\star$ is rank deficient, the elementary twist argument fails and uniqueness need not hold; the existence assertion nevertheless remains valid by the linear-cost analysis cited in the proof. This distinction is essential: a Monge optimizer always exists here under the hypotheses of Proposition 13.38,

but a literal full-rank Brenier characterization does not always apply. For two general nonatomic metric-measure spaces, Mémoli and Needham prove equality of the Gromov–Monge and Gromov–Wasserstein infimum values, but equality of infima alone does not ensure that the GW optimum is attained by a map [378].

GW with the squared-distance dissimilarities (13.30) behaves differently. Dumont, Lacombe and Vialard prove existence only of an optimal 2-map, namely a coupling supported on the union of two graphs; their one-dimensional computations give evidence that a single optimal graph need not exist [193]. A universal Brenier theorem therefore remains unavailable beyond the inner-product model (13.29).

Entropic regularization and biconvex alternating minimization. The frozen-OT characterization of Proposition 13.27 suggests a natural fixed-point iteration: given a coupling $\pi^{(\ell)}$, form its linearized cost $c_{\pi^{(\ell)}}$, then replace $\pi^{(\ell)}$ by an optimal coupling for this cost. Rather than viewing this procedure only as an ad hoc successive linearization, we now derive it from a symmetric biconvex relaxation obtained by duplicating the coupling. Before regularization, the diagonal of this lift recovers the original GW objective, optimization in either block is an ordinary OT problem, and minimizing over both blocks gives a computable lower bound [431]. Adding entropy symmetrically preserves this structure: the diagonal becomes the entropic GW objective and each block becomes an entropic OT problem. Entropy does not make the joint problem convex, but it allows every coupling update to be computed efficiently by Sinkhorn's algorithm.

Assume in this paragraph that $\mathcal{X}$ and $\mathcal{Y}$ are compact metric spaces, that L in (13.16) is continuous, and set $\mathfrak{m} = \alpha \otimes \beta$. With the relative entropy of Definition 8.11, define

$$\mathcal{R}_\varepsilon(\pi) := \begin{cases} 0, & \varepsilon = 0, \\ \varepsilon\,\mathrm{KL}(\pi|\mathfrak{m}), & \varepsilon > 0. \end{cases}$$

The separate convention at $\varepsilon = 0$ avoids the ambiguous product $0 \cdot (+\infty)$. The KL penalty is convenient but not essential to the lifted construction: it can be replaced by any proper convex lower-semicontinuous regularizer on couplings. Diagonal consistency, the lower bound and the polarization argument below continue to hold after splitting that regularizer equally between the two blocks. Strict convexity is needed when concluding that the two optimal blocks coincide. KL is singled out because it makes each block an entropic OT problem amenable to Sinkhorn scaling.

Definition 13.39 (Entropic GW and its biconvex relaxation)**.** For $\varepsilon \geqslant 0$, set

$$\mathcal{F}_\varepsilon(\pi) := \mathcal{Q}_L(\pi) + \mathcal{R}_\varepsilon(\pi), \qquad \mathcal{GW}_\varepsilon(\mathbb{X},\mathbb{Y})^p := \min_{\pi\in\mathcal{U}(\alpha,\beta)} \mathcal{F}_\varepsilon(\pi), \tag{13.41}$$

$$\mathcal{F}^{\mathrm{bi}}_\varepsilon(\pi,\xi) := \mathcal{B}_L(\pi,\xi) + \frac{1}{2}\mathcal{R}_\varepsilon(\pi) + \frac{1}{2}\mathcal{R}_\varepsilon(\xi), \qquad \mathcal{GW}_{\varepsilon,\mathrm{bi}}(\mathbb{X},\mathbb{Y})^p := \min_{\pi,\xi\in\mathcal{U}(\alpha,\beta)} \mathcal{F}^{\mathrm{bi}}_\varepsilon(\pi,\xi). \tag{13.42}$$

The factors $1/2$ in (13.42) are fixed by the diagonal-consistency identity $\mathcal{F}^{\mathrm{bi}}_\varepsilon(\pi,\pi) = \mathcal{F}_\varepsilon(\pi)$; they are not an additional regularization choice. Both minima are attained. Indeed, $\mathcal{U}(\alpha,\beta)$ is narrowly compact, the continuous bounded kernel L makes $\mathcal{B}_L$ narrowly continuous, and relative entropy is narrowly lower semicontinuous.

Diagonal consistency also gives the lower bound

$$\mathcal{GW}_{\varepsilon,\mathrm{bi}}(\mathbb{X},\mathbb{Y})^p \leqslant \mathcal{GW}_\varepsilon(\mathbb{X},\mathbb{Y})^p. \tag{13.43}$$

Without an additional curvature hypothesis, this lower bound can be strict. Even the scalar analogue already shows the obstruction: on $[-1,1]$, the diagonal minimum of $x \mapsto x^2$ is 0, whereas the lifted bilinear minimum of $(x,y) \mapsto xy$ is -1.

Block optimality explains both the usefulness and the limitation of the relaxation. If $(\pi_\star,\xi_\star)$ minimizes (13.42), then each component minimizes an OT problem with the other component frozen. Indeed, for the linearized cost c_ξ defined in (13.17), one has $\mathcal{B}_L(\pi,\xi) = \frac{1}{2}\int c_\xi \mathrm{d}\pi$ and therefore

$$\pi_\star \in \operatorname*{argmin}_{\pi\in\mathcal{U}(\alpha,\beta)} \left\{ \int c_{\xi_\star}\mathrm{d}\pi + \mathcal{R}_\varepsilon(\pi) \right\},$$

and symmetrically for $\xi_\star$. Thus the relaxation replaces the quadratic problem by two OT problems, but their ground costs are themselves unknown.

In the unregularized discrete case, these block problems are linear programs. If α and β are both uniform on n points, at least one *global minimizer of the lifted problem* has the form $(\mathrm{P}_\sigma/n, \mathrm{P}_\tau/n)$. This conclusion does not use joint convexity. Start from a global minimizer (P, Q). With Q frozen, P must minimize the first block problem; replacing it by an extreme minimizer preserves the global value. With this new first block frozen, the same global-minimality argument shows that the second block can also be replaced by an extreme minimizer without decreasing the value below the global minimum. The Birkhoff–von Neumann theorem identifies both extreme blocks as permutation couplings. Hence a unique global minimizer must be a pair of permutation couplings, although a nonunique optimal face may also contain fractional couplings. This result neither proves tightness, because σ and τ need not coincide, nor guarantees that alternating minimization reaches such a global pair. The algorithm only enforces blockwise optimality and can become trapped at a non-global pair of mutually optimal permutations. Thus the abundance of permutation-valued blockwise fixed points is a potential obstruction to global convergence, rather than a global-optimality certificate for the iterates. No analogous permutation conclusion holds for general nonuniform marginals or for $\varepsilon > 0$, when positive weights and finite costs yield dense entropic block minimizers.

The following polarization result extends Konno's unregularized equivalence between quadratic and bilinear programs [314]; its entropic GW specialization appears in [484]. It identifies the regime in which the lower bound is exact.

Proposition 13.40 (Tightness of the biconvex GW relaxation)**.** *Assume that $\mathcal{Q}_L$ is concave on $\mathcal{U}(\alpha, \beta)$. Since $\mathcal{Q}_L$ is quadratic, this is equivalent to*

$$\mathcal{Q}_L(\pi - \xi) \leqslant 0 \qquad \text{for every } \pi, \xi \in \mathcal{U}(\alpha, \beta). \tag{13.44}$$

Then (13.43) is an equality for every $\varepsilon \geqslant 0$. Moreover, every minimizing pair $(\pi_\star, \xi_\star)$ satisfies

$$\mathcal{F}_\varepsilon(\pi_\star) = \mathcal{F}_\varepsilon(\xi_\star) = \mathcal{GW}_\varepsilon(\mathbb{X}, \mathbb{Y})^p, \qquad \mathcal{Q}_L(\pi_\star - \xi_\star) = 0.$$

If $\varepsilon > 0$, then $\pi_\star = \xi_\star$. The same conclusion holds for $\varepsilon = 0$ when the inequality in (13.44) is strict for every nonzero feasible difference.

Proof. Symmetry and bilinearity give the polarization identity

$$\mathcal{F}_\varepsilon^{\mathrm{bi}}(\pi, \xi) = \frac{1}{2}\mathcal{F}_\varepsilon(\pi) + \frac{1}{2}\mathcal{F}_\varepsilon(\xi) - \frac{1}{2}\mathcal{Q}_L(\pi - \xi). \tag{13.45}$$

Under (13.44), the right-hand side is at least $\mathcal{GW}_\varepsilon(\mathbb{X}, \mathbb{Y})^p$. Conversely, diagonal pairs recover $\mathcal{F}_\varepsilon$, proving equality of the optimal values. Equality in the preceding bounds gives the two asserted identities. Strict negativity immediately forces $\pi_\star = \xi_\star$. Suppose now that $\varepsilon > 0$. The two diagonal objective values are finite, so both relative entropies are finite. If $\pi_\star \neq \xi_\star$, then $\mathcal{Q}_L(\pi_\star - \xi_\star) = 0$ makes the quadratic term affine on their joining segment, while strict convexity of relative entropy makes its value at the midpoint strictly smaller than the average of its endpoint values. The midpoint would therefore have a smaller diagonal objective, a contradiction. □

Condition (13.44) is precisely the conditional concavity studied above. In particular, Proposition 13.28 verifies it for squared GW when the two within-space kernels have the same conditional sign. In this regime, the biconvex relaxation is not merely a lower bound: it is an exact reformulation.

Starting from $(\pi^{(0)}, \xi^{(0)}) \in \mathcal{U}(\alpha, \beta)^2$, exact alternating minimization of the general-measure relaxation reads

$$\pi^{(\ell+1)} \in \operatorname*{argmin}_{\pi \in \mathcal{U}(\alpha,\beta)} \left\{ \int c_{\xi^{(\ell)}} \mathrm{d}\pi + \mathcal{R}_\varepsilon(\pi) \right\}, \tag{13.46}$$

$$\xi^{(\ell+1)} \in \operatorname*{argmin}_{\xi \in \mathcal{U}(\alpha,\beta)} \left\{ \int c_{\pi^{(\ell+1)}} \mathrm{d}\xi + \mathcal{R}_\varepsilon(\xi) \right\}. \tag{13.47}$$

Each block is an ordinary Kantorovich problem when $\varepsilon = 0$ and an entropic OT problem with regularization ε when $\varepsilon > 0$. Equivalently, dividing both the cost and regularization by two gives the half-gradient normalization used in the discrete formulation below.

For discrete measures, let H denote the discrete Shannon–Boltzmann entropy of Definition 8.1. Given $P \in U(a, b)$, define its contracted GW cost by

$$C(P)_{i,j} := \sum_{i'=1}^{n} \sum_{j'=1}^{m} \Delta\big(D_{i,i'}, D'_{j,j'}\big)^p P_{i',j'}. \tag{13.48}$$

Symmetry of the distortion kernel gives $\mathcal{B}_L(P, Q) = \langle C(P), Q\rangle = \langle C(Q), P\rangle$, while $\mathcal{E}_{D,D'}(P) = \langle C(P), P\rangle$. The KL identity on $U(a, b)$ shows that, after removing the fixed additive constant $\varepsilon(H(a) + H(b))$, the discrete biconvex problem is

$$\min_{P,Q\in U(a,b)} \langle C(P), Q\rangle - \frac{\varepsilon}{2}\big(H(P) + H(Q)\big), \tag{13.49}$$

and the original lifted optimal value is obtained by adding this constant back. There is no factor $1/2$ in front of the bilinear term under the convention (13.48): $C(P)$ is half of the full linearized cost in (13.17). Equivalently, if the full cost $2C(P)$ is used instead, the same term reads $\frac{1}{2}\langle 2C(P), Q\rangle$. Restricting (13.49) to the diagonal $P = Q$ recovers the usual entropic GW problem

$$\min_{P\in U(a,b)} \mathcal{E}_{D,D'}(P) - \varepsilon H(P). \tag{13.50}$$

For an unstructured distortion kernel, evaluating one entry of (13.48) requires nm terms. Forming all nm entries therefore costs $O(n^2m^2)$ dense arithmetic operations; explicitly storing the kernel would likewise require $O(n^2m^2)$ memory. The squared loss admits a substantially cheaper contraction without forming this four-index tensor.

Proposition 13.41 (Fast contracted cost for squared GW)**.** *Assume that*

$$\Delta\big(D_{i,i'}, D'_{j,j'}\big)^p = \big(D_{i,i'} - D'_{j,j'}\big)^2.$$

Then, for every $P \in U(a, b)$,

$$C(P) = D^{\odot 2} a\, \mathbb{1}_m^\top + \mathbb{1}_n (D'^{\odot 2} b)^\top - 2D\,P\,D'^\top. \tag{13.51}$$

Its ambient Frobenius gradient satisfies $\nabla\mathcal{E}_{D,D'}(P) = 2C(P)$. *Consequently,* $C(P)$ *can be computed in* $O(n^2m + nm^2)$ *dense arithmetic operations and* $O(nm)$ *working memory beyond the input distance matrices.*

Proof. Expanding the square in (13.48) gives

$$C(P)_{i,j} = \sum_{i',j'} D^2_{i,i'} P_{i',j'} + \sum_{i',j'} (D'_{j,j'})^2 P_{i',j'} - 2\sum_{i',j'} D_{i,i'} D'_{j,j'} P_{i',j'}.$$

The marginal identities $P\mathbb{1}_m = a$ and $P^\top \mathbb{1}_n = b$ turn the first two sums into $(D^{\odot 2}a)_i$ and $(D'^{\odot 2}b)_j$, while the last sum is $(DPD'^\top)_{i,j}$. This proves (13.51). The underlying four-index kernel is symmetric, so differentiating $\mathcal{E}_{D,D'}(P) = \langle C(P), P\rangle$ gives the factor two in the gradient. Finally, computing DP and then $(DP)D'^\top$ costs respectively $O(n^2m)$ and $O(nm^2)$ operations; the marginal terms are cheaper. □

Exact alternating minimization of (13.49) therefore consists of ordinary entropic OT solves with regularization $\varepsilon/2$. Even with (13.51), however, refreshing the cost costs $O(N^3)$ when $n = m = N$, whereas one dense Sinkhorn scaling step with a fixed cost costs $O(N^2)$. The GW cost update can thus become the dominant bottleneck. If $P = UV^\top$ has rank r, then $DPD'^\top = (DU)(D'V)^\top$, which lowers this contraction to $O(n^2r + m^2r + nmr)$. This is the mechanism behind the factored-coupling construction of Definition 12.24. Scetbon, Peyré and Cuturi additionally exploit low-rank distance matrices to obtain a complexity linear in $n + m$ when all approximation ranks are fixed [475].

Each half-step decreases the lifted objective $\mathcal{F}^{\mathrm{bi}}_\varepsilon$. Under (13.44), the same update is also the tangent majorization step (13.23) for the diagonal objective. Outside this regime, it minimizes only the biconvex lower bound and need not decrease (13.50). Since the lifted functional is symmetric in its two blocks, interleaving

$$Q^{(0)},\ P^{(1)},\ Q^{(1)},\ P^{(2)}, \dots$$

recovers the usual one-sequence GW linearization; the algorithm below merely groups two successive half-steps into one cycle. In the conditionally concave regime, Remark 13.32 identifies this interleaved

sequence with the classical DC algorithm. In finite dimensions, standard block-coordinate theory applies under its additional regularity hypotheses. In particular, for $\varepsilon > 0$, positive weights and finite costs make each entropic block minimizer unique; together with compactness, this places the iteration in the setting of the stationary-point results of Tseng [524]. For $\varepsilon = 0$, block minimizers can be nonunique and an appropriate convergence theorem requires further assumptions or a selection rule. In either case, block-coordinate results are local statements, not certificates of global optimality or of tightness of the relaxation.

Algorithm 13.1: Alternating minimization of the biconvex GW relaxation

Input: Symmetric metric matrices D, D', weights a, b, exponent $p \geqslant 1$, distortion Δ, regularization $\varepsilon \geqslant 0$, tolerance tol, iteration budget $\ell_{\max} \geqslant 1$.
Output: Approximate relaxed pair $(P, Q) \in U(a, b)^2$.
Initialize: Set $P^{(0)} = Q^{(0)} = a \otimes b$.
For $\ell = 0, \dots, \ell_{\max} - 1$ **do**:
  $P^{(\ell+1)} = \operatorname*{argmin}_{P \in U(a,b)} \langle C(Q^{(\ell)}), P\rangle - (\varepsilon/2)H(P)$.
  $Q^{(\ell+1)} = \operatorname*{argmin}_{Q \in U(a,b)} \langle C(P^{(\ell+1)}), Q\rangle - (\varepsilon/2)H(Q)$.
  If $\max\{\|P^{(\ell+1)} - P^{(\ell)}\|_F, \|Q^{(\ell+1)} - Q^{(\ell)}\|_F\} \leqslant$ tol **then**:
    Return $(P^{(\ell+1)}, Q^{(\ell+1)})$.
Return $(P^{(\ell_{\max})}, Q^{(\ell_{\max})})$.

Fused GW: adding features. Pure GW deliberately ignores any direct comparison between $x \in \mathcal{X}$ and $y \in \mathcal{Y}$: it only compares the two within-space geometries. Fused Gromov–Wasserstein (FGW) retains this intrinsic term while adding a cross-space feature cost [535]. In machine-learning applications, points in the two spaces often carry features in a common metric space $(\mathcal{Z}_f, d_{\mathcal{Z}_f})$ through embeddings

$$\varphi_{\mathcal{X}} : \mathcal{X} \to \mathcal{Z}_f, \qquad \varphi_{\mathcal{Y}} : \mathcal{Y} \to \mathcal{Z}_f,$$

in which case a standard choice is $c_{\mathcal{X},\mathcal{Y}}(x, y) = d_{\mathcal{Z}_f}(\varphi_{\mathcal{X}}(x), \varphi_{\mathcal{Y}}(y))^p$.

Definition 13.42 (Fused Gromov–Wasserstein problem)**.** Under the notation of (13.16), let $c_{\mathcal{X},\mathcal{Y}} : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$ be a measurable cross-space cost and let $0 \leqslant \lambda \leqslant 1$. The fused GW value is

$$\mathrm{FGW}_{\lambda,p}(\mathbb{X}, \mathbb{Y}; c_{\mathcal{X},\mathcal{Y}})^p := \min_{\pi \in \mathcal{U}(\alpha,\beta)} (1-\lambda) \int_{\mathcal{X}\times\mathcal{Y}} c_{\mathcal{X},\mathcal{Y}} \mathrm{d}\pi + \lambda Q_L(\pi). \tag{13.52}$$

The first term compares features by ordinary Kantorovich transport, while the second compares intrinsic geometry. Thus $\lambda = 0$ gives feature-only OT and $\lambda = 1$ gives pure GW; intermediate values balance pointwise feature fidelity against pairwise geometric fidelity. This balance is meaningful only after calibrating the scales of $c_{\mathcal{X},\mathcal{Y}}$ and L, since rescaling either term changes the interpretation of λ.

For discrete measures, setting $M_{i,j} = c_{\mathcal{X},\mathcal{Y}}(x_i, y_j)$ reduces (13.52) to

$$\min_{P \in U(a,b)} (1-\lambda)\langle M, P\rangle + \lambda \mathcal{E}_{D,D'}(P),$$

which is the usual discrete fused-GW problem.

The algorithms above extend by the same linearization principle. With the coupling regularizer $\mathcal{R}_\varepsilon$ defined before (13.41), the regularized fused objective is

$$\mathcal{F}^{\mathrm{FGW}}_{\varepsilon,\lambda}(\pi) := (1-\lambda) \int_{\mathcal{X}\times\mathcal{Y}} c_{\mathcal{X},\mathcal{Y}} \mathrm{d}\pi + \lambda Q_L(\pi) + \mathcal{R}_\varepsilon(\pi).$$

Its first variation is an ordinary OT objective with ground cost

$$c^{\mathrm{FGW}}_{\pi,\lambda}(x, y) := (1-\lambda) c_{\mathcal{X},\mathcal{Y}}(x, y) + 2\lambda \int_{\mathcal{X}\times\mathcal{Y}} L\big((x, y), (x', y')\big) \mathrm{d}\pi(x', y'). \tag{13.53}$$

Indeed, the feature term is linear in π, whereas differentiating Q_L produces the factor two already present in (13.17).

Equivalently, diagonal consistency gives the symmetric biconvex lift

$$\mathcal{F}^{\mathrm{FGW,bi}}_{\varepsilon,\lambda}(\pi,\xi) := \frac{1-\lambda}{2}\int_{\mathcal{X}\times\mathcal{Y}} c_{\mathcal{X},\mathcal{Y}}\,\mathrm{d}(\pi+\xi) + \lambda\mathcal{B}_L(\pi,\xi) + \frac{1}{2}\mathcal{R}_\varepsilon(\pi) + \frac{1}{2}\mathcal{R}_\varepsilon(\xi).$$

Multiplying a block objective by two shows that, with ξ fixed, its π-update is precisely

$$\operatorname*{argmin}_{\pi\in\mathcal{U}(\alpha,\beta)} \left\{\int_{\mathcal{X}\times\mathcal{Y}} c^{\mathrm{FGW}}_{\xi,\lambda}\,\mathrm{d}\pi + \mathcal{R}_\varepsilon(\pi)\right\}.$$

Every block is therefore an ordinary OT problem, or an entropic OT problem solvable by Sinkhorn when $\mathcal{R}_\varepsilon$ is the KL penalty. Moreover, the added feature term is affine in each coupling, so the tightness result of Proposition 13.40 remains valid whenever Q_L is concave on $\mathcal{U}(\alpha,\beta)$.

Figure 13.7 isolates the resulting tradeoff on two Euclidean point clouds.

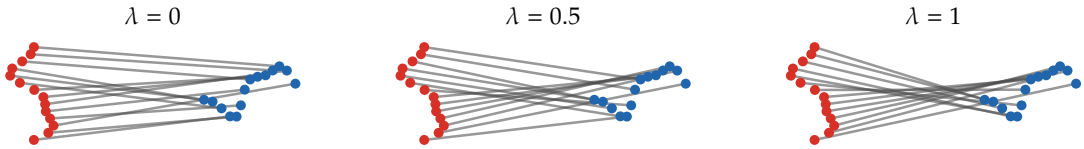


Figure 13.7: Feature information and intrinsic geometry in fused Gromov–Wasserstein. Here $\mathcal{X} = \mathcal{Y} = \mathbb{R}^2$, $c_{\mathcal{X},\mathcal{Y}}(x,y) = \|x-y\|^2$, and both intrinsic distances are Euclidean. The target is a rotated and translated, slightly perturbed copy of the source; the additional horizontal separation in the drawing is only for readability and is not included in the cost. Thus $\lambda = 0$ follows ambient proximity, $\lambda = 1$ follows intrinsic geometry, and $\lambda = 0.5$ balances the two criteria.

Example 13.43 (Application to multi-omics alignment). Multi-omics experiments can produce unmatched point clouds in different feature spaces, for instance RNA and ATAC profiles measured on different cells. The GW term preserves within-modality neighborhood structure, while any available cross-modality information enters FGW through $c_{\mathcal{X},\mathcal{Y}}$, or through the matrix M after discretization. Unbalanced and co-optimal variants further allow the number and relative importance of cells or features to differ across modalities. The resulting template is precise: preserve relational geometry within each modality while learning a coupling across modalities [521, 463, 505].

Hausdorff and Gromov–Hausdorff viewpoints. In geometry, one often wants to study a sequence of metric spaces without first embedding them into a common ambient space: the object of interest is then convergence of the spaces themselves, rather than convergence of distinguished points or measures carried by them. This requires a metric on a "space of spaces." The Gromov–Hausdorff distance was designed for precisely this purpose [255, 87], and its use for intrinsic shape and object comparison was developed notably by Mémoli [374]. Within the matching hierarchy developed here, it is the measure-free, worst-case counterpart of Gromov–Monge and Gromov–Wasserstein comparison. To make this relation precise, one must distinguish three feasible objects: arbitrary correspondences for Gromov–Hausdorff, measure-preserving maps for Gromov–Monge, and probability couplings for Gromov–Wasserstein.

Definition 13.44 (Hausdorff distance)**.** If A, B are nonempty compact subsets of a metric space (Z, d_Z), their Hausdorff distance is

$$d^Z_{\mathrm{H}}(A,B) := \max\left\{\sup_{a\in A}\inf_{b\in B} d_Z(a,b), \sup_{b\in B}\inf_{a\in A} d_Z(a,b)\right\}.$$

Writing $\mathcal{K}(Z)$ for the set of nonempty compact subsets of Z, d^Z_{H} is a metric on $\mathcal{K}(Z)$. If Z is complete, then $\mathcal{K}(Z)$ is complete; if Z is compact, then $\mathcal{K}(Z)$ is itself compact. Hausdorff distance is therefore the natural tool for studying convergence of compact sets inside one fixed ambient space [87, Section 7.3].

Gromov–Hausdorff distance removes the need to prescribe this ambient space.

Definition 13.45 (Gromov–Hausdorff distance)**.** For compact metric spaces $(\mathcal{X}, d_{\mathcal{X}})$ and $(\mathcal{Y}, d_{\mathcal{Y}})$, their Gromov–Hausdorff distance is

$$d_{\mathrm{GH}}(\mathcal{X}, \mathcal{Y}) := \inf_{Z, \varphi, \psi} d_{\mathrm{H}}^{Z}(\varphi(\mathcal{X}), \psi(\mathcal{Y})).$$

One has $d_{\mathrm{GH}}(\mathcal{X}, \mathcal{Y}) = 0$ exactly when $\mathcal{X}$ and $\mathcal{Y}$ are isometric. Hence d_{GH} is a genuine metric on isometry classes of compact metric spaces. This metric space is complete and separable, but it is *not* compact: for example, two-point spaces with diverging diameters have no convergent subsequence. Gromov's compactness theorem gives the precise replacement. A family is relatively compact for d_{GH} if and only if its diameters are uniformly bounded and, for every $\varepsilon > 0$, all its members can be covered by at most $N(\varepsilon) < +\infty$ balls of radius ε [87, Section 7.4]; see also [255, 422].

Now equip $\mathcal{X}$ and $\mathcal{Y}$ with full-support probability measures α and β, and write $\mathbb{X} = (\mathcal{X}, d_{\mathcal{X}}, \alpha)$ and $\mathbb{Y} = (\mathcal{Y}, d_{\mathcal{Y}}, \beta)$. Proposition 13.17 already identifies $\mathcal{GW}_\infty$ as the endpoint of finite-order GW. The remaining question is how this measure-dependent comparison relates to the measure-free distance d_{GH}.

Specialize from now on to $\Delta(r, s) = |r - s|$. A correspondence is a set $R \subset \mathcal{X} \times \mathcal{Y}$ whose two coordinate projections are surjective, and its distortion is

$$\mathrm{dis}(R) := \sup_{(x,y),(x',y') \in R} \left| d_{\mathcal{X}}(x, x') - d_{\mathcal{Y}}(y, y') \right|.$$

Writing $\mathfrak{R}(\mathcal{X}, \mathcal{Y})$ for the set of correspondences, compact metric spaces satisfy the classical identity [255, 87, 374]

$$2 d_{\mathrm{GH}}(\mathcal{X}, \mathcal{Y}) = \inf_{R \in \mathfrak{R}(\mathcal{X}, \mathcal{Y})} \mathrm{dis}(R). \tag{13.54}$$

Together with the support representation (13.11), this identity gives the exact connection between Gromov–Hausdorff and infinite-order GW. The factor 2 below reflects the convention (13.9); references that insert a factor $1/2$ in GW do not display it.

Proposition 13.46 (Gromov–Hausdorff as an infinite-order intrinsic matching)**.** *Let $\mathcal{X}$ and $\mathcal{Y}$ be compact metric spaces and let α, β have full support. Then*

$$2 d_{\mathrm{GH}}(\mathcal{X}, \mathcal{Y}) \leqslant \mathcal{GW}_\infty(\mathbb{X}, \mathbb{Y}). \tag{13.55}$$

Moreover, removing the prescribed weights recovers Gromov–Hausdorff exactly:

$$2 d_{\mathrm{GH}}(\mathcal{X}, \mathcal{Y}) = \inf_{\substack{\alpha \in \mathcal{M}_+^1(\mathcal{X),\ \mathrm{supp}(\alpha) = \mathcal{X}} \\ \beta \in \mathcal{M}_+^1(\mathcal{Y}),\ \mathrm{supp}(\beta) = \mathcal{Y}}} \mathcal{GW}_\infty\big((\mathcal{X}, d_{\mathcal{X}}, \alpha), (\mathcal{Y}, d_{\mathcal{Y}}, \beta)\big). \tag{13.56}$$

Proof. The support of any $\pi \in \mathcal{U}(\alpha, \beta)$ is compact. The support of each marginal is the closure of the corresponding coordinate projection of $\mathrm{supp}(\pi)$. These projections are compact, hence already closed; because α and β have full support, they are therefore equal to $\mathcal{X}$ and $\mathcal{Y}$. Thus $\mathrm{supp}(\pi)$ is a correspondence, and (13.54) gives (13.55).

For (13.56), the preceding lower bound applies to every pair of full-support weights. Conversely, choose a correspondence R with distortion arbitrarily close to the right-hand side of (13.54). Replacing R by its closure preserves both the correspondence property and its distortion, so R may be assumed compact. It then carries a probability π with support exactly R. Let α and β be its marginals. They have full support, while (13.11) gives $\mathcal{GW}_\infty(\mathbb{X}, \mathbb{Y}) \leqslant \mathrm{dis}(R)$. Taking the infimum over R proves the identity. □

Thus Gromov–Hausdorff is the mass-free, hard-support counterpart of GW: it minimizes worst-case distortion over correspondences between the two supports. By contrast, finite-p GW fixes the masses and averages distortion under a coupling; optimizing these masses and taking $p = \infty$ recovers $2 d_{\mathrm{GH}}$ through (13.56).

13.4 Quantum Optimal Transport

Quantum optimal transport replaces probability vectors by density matrices and scalar couplings by positive operators on a tensor product space. This is the right language when the transported objects are matrix-valued signals, covariance-like descriptors or quantum states, and it exposes a precise bridge between OT, non-commutative entropy and operator scaling. The finite-dimensional formulation below follows the semidefinite viewpoint developed in matrix-valued and quantum OT [402, 132, 130, 429, 99, 121].

Finite-dimensional states and couplings. The finite-dimensional case is the cleanest quantum analogue of the Kantorovich formulation. Probability vectors are replaced by positive semidefinite, trace-one matrices, and a coupling becomes a positive semidefinite operator on a tensor product whose partial traces prescribe the two marginals. This keeps the familiar "joint object with fixed marginals" geometry, while turning the transport problem into a semidefinite program.

Definition 13.47 (Hermitian and density matrices)**.** Let $\mathbb{H}_n$ be the real vector space of $n \times n$ Hermitian matrices,

$$\mathbb{H}_n^+ = \{A \in \mathbb{H}_n \,;\, A \succeq 0\}\,, \qquad \mathbb{H}_n^{+,1} = \left\{A \in \mathbb{H}_n^+ \,;\, \mathrm{tr}(A) = 1\right\}.$$

Elements of $\mathbb{H}_n^{+,1}$ are density matrices.

A joint quantum state between $\mathbb{C}^n$ and $\mathbb{C}^m$ is a matrix $T \in \mathbb{H}_{nm}^+$ acting on $\mathbb{C}^n \otimes \mathbb{C}^m$. Its marginals are the partial traces, defined by duality through

$$\mathrm{tr}(F\ \mathrm{Tr}_B\, T) = \mathrm{tr}((F \otimes \mathrm{Id}_m)T), \qquad \mathrm{tr}(G\ \mathrm{Tr}_A\, T) = \mathrm{tr}((\mathrm{Id}_n \otimes G)T), \tag{13.57}$$

for all $F \in \mathbb{H}_n$ and $G \in \mathbb{H}_m$. Thus $\mathrm{Tr}_B(T) \in \mathbb{H}_n^+$ and $\mathrm{Tr}_A(T) \in \mathbb{H}_m^+$ play exactly the role of the two marginals of a classical coupling. The feasible set is never empty, since $A \otimes B$ has marginals A and B.

Definition 13.48 (Finite-dimensional quantum OT)**.** Let $A \in \mathbb{H}_n^{+,1}$, $B \in \mathbb{H}_m^{+,1}$ and let $C \in \mathbb{H}_{nm}$ be a Hermitian cost observable. The quantum OT value is the semidefinite program

$$\mathrm{QOT}_C(A,B) \coloneqq \min_{T \in \mathbb{H}_{nm}^+} \{\mathrm{tr}(CT)\,;\ \mathrm{Tr}_B(T) = A,\ \mathrm{Tr}_A(T) = B\}\,. \tag{13.58}$$

Example 13.49 (Classical diagonal case). Suppose that A, B and C are diagonal in fixed product bases. Let Δ_A and Δ_B be the corresponding dephasing maps. For every feasible T, the matrix $\widetilde{T} = (\Delta_A \otimes \Delta_B)(T)$ is positive, has the same partial traces, and satisfies $\mathrm{tr}(C\widetilde{T}) = \mathrm{tr}(CT)$. Hence an optimal coupling may be chosen diagonal. Its diagonal entries form a nonnegative matrix with row sums given by A and column sums given by B, so (13.58) reduces exactly to classical Kantorovich OT. The genuinely quantum feature is that non-commuting data may require off-diagonal coherences or entanglement.

Proposition 13.50 (Quantum Kantorovich duality)**.** *For $A \in \mathbb{H}_n^{+,1}$ and $B \in \mathbb{H}_m^{+,1}$, the dual of (13.58) is*

$$\mathrm{QOT}_C(A,B) = \sup_{F \in \mathbb{H}_n,\ G \in \mathbb{H}_m} \{\mathrm{tr}(FA) + \mathrm{tr}(GB) :\ F \otimes \mathrm{Id}_m + \mathrm{Id}_n \otimes G \preceq C\}\,. \tag{13.59}$$

There is no duality gap. If A and B are positive definite, the supremum is attained. For singular marginals, the primal reduces to the tensor product of their supports; on this reduced space the corresponding dual maximum is attained, while the full-space formula above is always valid as a supremum.

Proof. Introduce Hermitian Lagrange multipliers F and G for the two marginal constraints. The Lagrangian is

$$\mathrm{tr}(CT) + \mathrm{tr}(F(A - \mathrm{Tr}_B\, T)) + \mathrm{tr}(G(B - \mathrm{Tr}_A\, T)) = \mathrm{tr}(FA) + \mathrm{tr}(GB) + \mathrm{tr}((C - F \otimes \mathrm{Id}_m - \mathrm{Id}_n \otimes G)T)\,,$$

where (13.57) was used in the last equality. Minimizing over $T \succeq 0$ gives a finite lower bound if and only if $C - F \otimes \mathrm{Id}_m - \mathrm{Id}_n \otimes G \succeq 0$, in which case the infimum in T is 0. This gives the dual program. When $A, B \succ 0$, the coupling $A \otimes B$ is strictly feasible, so Slater's theorem gives equality of primal and dual values and dual attainment.

For singular marginals, let P_A and P_B be their support projections. If T is feasible, then

$$\operatorname{tr}\big(((\mathrm{Id}_n - P_A) \otimes \mathrm{Id}_m)T\big) = \operatorname{tr}((\mathrm{Id}_n - P_A)A) = 0.$$

Positivity implies that T vanishes on $(\operatorname{supp} A)^\perp \otimes \mathbb{C}^m$; the analogous argument for B yields $T = (P_A \otimes P_B)T(P_A \otimes P_B)$. Thus the primal is exactly the problem obtained by compressing A, B and C to $\operatorname{supp} A \otimes \operatorname{supp} B$. The reduced marginals are positive definite, so Slater applies there. Equivalently, the unreduced dual values approach the reduced maximum after assigning sufficiently negative potentials on the orthogonal complements. This proves the stated supremum formula without a gap. □

The dual potentials have the usual scalar gauge freedom: replacing (F, G) by $(F + t\mathrm{Id}_n, G - t\mathrm{Id}_m)$ leaves both the constraint and the value unchanged because $\operatorname{tr}(A) = \operatorname{tr}(B) = 1$.

Entropic regularization and Bregman iterations. As in scalar OT, one obtains a smoother problem by adding the convex quantum entropy functional associated with the matrix logarithm.

Definition 13.51 (von Neumann quantum entropy)**.** For a density matrix or positive semidefinite matrix T, the shifted von Neumann entropy functional used here is

$$H(T) = \operatorname{tr}\big(T(\log T - \mathrm{Id})\big), \qquad \nabla H(T) = \log T,$$

with the convention $0 \log 0 = 0$ on eigenvalues. This is the convex negative quantum entropy; on trace-one states it differs from the physical entropy $-\operatorname{tr}(T \log T)$ by a sign and an additive constant.

For $\varepsilon > 0$ define $$\mathrm{QOT}_C^\varepsilon(A, B) = \min_{T \geq 0} \{\operatorname{tr}(CT) + \varepsilon H(T) : \operatorname{Tr}_B(T) = A, \operatorname{Tr}_A(T) = B\}. \tag{13.60}$$

This is the non-commutative analogue of entropic OT [429, 121]: the Shannon entropy of a coupling is replaced by the trace entropy of a density matrix.

Proposition 13.52 (Entropic quantum OT duality)**.** *Assume $A \succ 0$, $B \succ 0$ and $\varepsilon > 0$. Then (13.60) has a unique positive minimizer. Its dual is*

$$\mathrm{QOT}_C^\varepsilon(A, B) = \max_{F \in \mathbb{H}_n,\, G \in \mathbb{H}_m} \left\{ \operatorname{tr}(FA) + \operatorname{tr}(GB) - \varepsilon \operatorname{tr}\exp\left(\frac{F \otimes \mathrm{Id}_m + \mathrm{Id}_n \otimes G - C}{\varepsilon}\right)\right\}. \tag{13.61}$$

At optimality, primal and dual variables are linked by the Gibbs formula

$$T_e(F, G) = \exp\left(\frac{F \otimes \mathrm{Id}_m + \mathrm{Id}_n \otimes G - C}{\varepsilon}\right), \tag{13.62}$$

with $\operatorname{Tr}_B(T_e) = A$ and $\operatorname{Tr}_A(T_e) = B$.

Proof. The feasible set is compact and nonempty, and it contains the positive definite point $A \otimes B$. The trace entropy is strictly convex on positive semidefinite matrices, hence the regularized primal has a unique minimizer. This minimizer is positive definite: if it had a non-trivial kernel, moving toward $A \otimes B$ would have entropy directional derivative $-\infty$ at the boundary while changing the linear cost only at finite rate. Slater's condition justifies the Lagrange dual computation. The Fenchel identity

$$\sup_{T \geq 0} \operatorname{tr}(YT) - \varepsilon H(T) = \varepsilon \operatorname{tr}\exp(Y/\varepsilon)$$

is the matrix analogue of the scalar exponential conjugacy. Applying it to the Lagrangian of (13.60), with $Y = F \otimes \mathrm{Id}_m + \mathrm{Id}_n \otimes G - C$, gives (13.61). The stationarity condition of this Fenchel identity gives (13.62); differentiating the dual objective with respect to F and G yields the two marginal equations. □

Writing $K = \exp(-C/\varepsilon)$, the objective in (13.60) differs by a constant from ε times the quantum KL divergence

$$D_H(T|K) = \operatorname{tr}\big(T(\log T - \log K) - T + K\big).$$

The exact quantum analogue of Sinkhorn is an implicit alternating Bregman projection scheme onto the affine marginal sets

$$\mathcal{M}_A = \{T \succeq 0 : \operatorname{Tr}_B(T) = A\}, \qquad \mathcal{M}_B = \{T \succeq 0 : \operatorname{Tr}_A(T) = B\}.$$

Proposition 13.53 (Exact Bregman projections)**.** *Assume $A, B \succ 0$ and let $K = \exp(-C/\varepsilon)$. The minimizer of (13.60) is equivalently the minimizer of $D_H(T|K)$ over $\mathcal{M}_A \cap \mathcal{M}_B$. Moreover, if a current positive definite matrix has Gibbs form $T_e(F, G)$, then its Bregman projection onto $\mathcal{M}_A$ has the form $T_e(F^+, G)$, where F^+ is chosen so that $\operatorname{Tr}_B T_e(F^+, G) = A$. The projection onto $\mathcal{M}_B$ is analogous. Thus, when each one-block marginal equation is solved exactly, alternating Bregman projections are equivalent to alternating block maximization of the dual (13.61).*

Proof. Since $\log K = -C/\varepsilon$, the identity

$$\operatorname{tr}(CT) + \varepsilon H(T) = \varepsilon D_H(T|K) - \varepsilon \operatorname{tr}(K)$$

holds, so the primal minimizer is the constrained Bregman projection of K up to an additive constant. For the projection of a positive definite matrix S onto $\mathcal{M}_A$, the affine set contains the positive definite point $A \otimes \mathrm{Id}_m/m$; the entropy derivative $\log T - \log S$ is singular at the boundary, so the projection lies in the interior of the positive cone. Its Lagrangian first variation is $\log T - \log S - \Lambda \otimes \mathrm{Id}_m = 0$ for a Hermitian multiplier Λ. Hence $T = \exp(\log S + \Lambda \otimes \mathrm{Id}_m)$. If $S = T_e(F, G)$, this is again of the form $T_e(F + \varepsilon\Lambda, G)$. The multiplier is fixed by the marginal equation $\operatorname{Tr}_B(T) = A$. The same argument applies to $\mathcal{M}_B$. Finally, the first-order optimality condition for maximizing (13.61) over one block is exactly the corresponding marginal equation, so the Bregman and block-dual views coincide. □

In the diagonal case this proposition gives the usual multiplicative Sinkhorn updates. In the non-commutative case, however, the exact block equations

$$\operatorname{Tr}_B T_e(F, G) = A, \qquad \operatorname{Tr}_A T_e(F, G) = B$$

do not admit scalar division formulas, because the exponential of $F \otimes \mathrm{Id}_m + \mathrm{Id}_n \otimes G - C$ cannot be separated unless the local potential $F \otimes \mathrm{Id}_m + \mathrm{Id}_n \otimes G$ commutes with the cost C. When all matrices are diagonal in the same basis, commutativity restores the scalar form $T_e = \operatorname{diag}(u)K\operatorname{diag}(v)$ and the marginal equations reduce to the usual Sinkhorn divisions.

Algorithm 13.2 records this exact but implicit non-commutative analogue of Sinkhorn.

Algorithm 13.2: Exact quantum Bregman projections

Input: Positive definite density matrices A, B, cost C, regularization $\varepsilon > 0$, tolerance tol, iteration budget $L \geqslant 1$.
Output: Approximate quantum entropic coupling T and its partial-trace residual.
Initialize: Set Hermitian potentials $F^{(0)} = 0$ and $G^{(0)} = 0$.
For $\ell = 0, \dots, L-1$ **do**:
  $T^{(\ell)} = T_e(F^{(\ell)}, G^{(\ell)}) = \exp\left(\frac{F^{(\ell)} \otimes \mathrm{Id}_m + \mathrm{Id}_n \otimes G^{(\ell)} - C}{\varepsilon}\right)$.
  Solve A-projection equation: $\operatorname{Tr}_B T_e(F^+, G^{(\ell)}) = A$.
  Set $F^{(\ell+1)} = F^+$.
  Solve B-projection equation: $\operatorname{Tr}_A T_e(F^{(\ell+1)}, G^+) = B$.
  Set $G^{(\ell+1)} = G^+$.
  Set $T^{(\ell+1)} = T_e(F^{(\ell+1)}, G^{(\ell+1)})$ and $r^{(\ell+1)} = \max\{\| \operatorname{Tr}_B T^{(\ell+1)} - A\|_{\mathrm{F}}, \| \operatorname{Tr}_A T^{(\ell+1)} - B\|_{\mathrm{F}}\}$.
  If $r^{(\ell+1)} \leqslant$ tol **then**:
    Return $T^{(\ell+1)}$ and $r^{(\ell+1)}$.
Return $T^{(L)}$ and $r^{(L)}$.

Gurvits scaling and quantum Sinkhorn. The algorithm often called quantum Sinkhorn comes from the operator-scaling literature of Gurvits and subsequent developments [260, 261, 240, 234]. It replaces the true Gibbs coupling (13.62) by the symmetric factorization

$$T_s(F, G) = \exp\left(\frac{Z}{2\varepsilon}\right)\exp(-C/\varepsilon)\exp\left(\frac{Z}{2\varepsilon}\right) = (U \otimes V)K(U \otimes V), \qquad Z = F \otimes \mathrm{Id}_m + \mathrm{Id}_n \otimes G, \tag{13.63}$$

where $U = \exp(F/(2\varepsilon))$, $V = \exp(G/(2\varepsilon))$ and $K = \exp(-C/\varepsilon)$. If $[Z, C] = 0$, then $T_s(F, G) = T_e(F, G)$; otherwise this is a Strang-type symmetric surrogate.

Fix a Choi convention and let $\mathcal{K} : \mathbb{H}_m \to \mathbb{H}_n$ be the completely positive map represented by the positive Choi matrix K; let $\mathcal{K}^\star$ be its Hilbert–Schmidt adjoint. Up to the transpose dictated by the chosen Choi convention, the marginal equations for the symmetric coupling take the operator-scaling form

$$U\,\mathcal{K}(V^2)\,U = A, \qquad V\,\mathcal{K}^\star(U^2)\,V = B.$$

They can be enforced by the explicit congruence normalizations

$$\begin{aligned} R_V &= \mathcal{K}(V^2), \quad & U &\leftarrow R_V^{-1/2}\big(R_V^{1/2} A R_V^{1/2}\big)^{1/2} R_V^{-1/2}, \\ S_U &= \mathcal{K}^\star(U^2), \quad & V &\leftarrow S_U^{-1/2}\big(S_U^{1/2} B S_U^{1/2}\big)^{1/2} S_U^{-1/2}. \end{aligned} \tag{13.64}$$

These inverse square roots are well-defined when $K \succ 0$ and $U, V, A, B \succ 0$. Under the standard

Algorithm 13.3: Gurvits/operator scaling for quantum Sinkhorn

Input: Positive definite marginals A, B, positive definite kernel operator K, maps $\mathcal{K}, \mathcal{K}^\star$, tolerance tol, iteration budget $L \geqslant 1$.
Output: Symmetrically scaled coupling T_s and its marginal residual r.
Initialize: Set $U = \mathrm{Id}_n$ and $V = \mathrm{Id}_m$.
Set residual $r = +\infty$ and counter $\ell = 0$.
While $r >$ tol and $\ell < L$ **do**:
  $R_V = \mathcal{K}(V^2), \qquad U \leftarrow R_V^{-1/2}\big(R_V^{1/2} A R_V^{1/2}\big)^{1/2} R_V^{-1/2}.$
  $S_U = \mathcal{K}^\star(U^2), \qquad V \leftarrow S_U^{-1/2}\big(S_U^{1/2} B S_U^{1/2}\big)^{1/2} S_U^{-1/2}.$
  Set $T_s = (U \otimes V)K(U \otimes V)$ and r to the maximum of its two operator-marginal residuals against A and B.
  Set $\ell \leftarrow \ell + 1$.
Return T_s and r.

strict-positivity and scalability hypotheses, the alternating normalizations converge to the prescribed marginals [240, 234]; the finite algorithm returns once the residual is below tolerance or the iteration budget is exhausted. This is Gurvits/operator scaling with prescribed targets. When all matrices are diagonal it reduces to classical Sinkhorn scaling, and when the targets are proportional to identities it matches the usual bistochastic operator-scaling normalization, up to the conventional trace normalization.

Remark 13.54 (Gurvits scaling is not the exact Bregman scheme). It is important not to identify (13.64) with the exact Bregman scheme for (13.60). The exact Bregman step would enforce the marginals of $T_\varepsilon(F, G) = \exp((Z - C)/\varepsilon)$ and would be a block maximization of the true concave dual (13.61). Gurvits scaling instead enforces the marginals of the surrogate

$$T_s = \exp\left(\frac{Z}{2\varepsilon}\right)\exp(-C/\varepsilon)\exp\left(\frac{Z}{2\varepsilon}\right).$$

The two coincide in the commuting/diagonal regime, but in general the Baker–Campbell–Hausdorff commutator terms do not vanish. The Gurvits iteration should therefore be understood as a tractable symmetric operator-scaling approximation to entropic Q–OT, not as the literal alternating KL projection algorithm.

Remark 13.55 (Operator-valued couplings). The same definitions extend formally from matrices to separable Hilbert spaces by replacing density matrices with positive trace-class operators of trace one, observables with bounded self-adjoint operators and (13.57) with partial traces defined by duality against local bounded observables. If $\Pi(A, B)$ denotes positive trace-class operators with partial traces A and B, a bounded cost observable C gives the problem $\inf_{T \in \Pi(A,B)} \mathrm{Tr}(CT)$. For unbounded positive costs one must define the energy through the quadratic form or spectral truncations, and in the entropic case one must ensure that the Gibbs operator $\exp(-C/\varepsilon)$ is trace class and that the partial traces of the candidate coupling are well-defined. The matrix formulas above are therefore the clean finite-dimensional core; the operator version adds domain and compactness assumptions rather than a different algebraic structure.

13.5 Dynamic Time Warping

Dynamic time warping (DTW) compares ordered feature sequences when the same phenomenon may be observed under different clocks. It is historically rooted in speech recognition [541, 464], and is now a standard tool for time-series alignment, retrieval and classification [47, 393]. Like OT, it minimizes an aggregate feature mismatch over correspondences; unlike OT, those correspondences must respect chronology.

Ordered alignments versus transport couplings. For two empirical measures, Kantorovich OT minimizes a linear cost over the convex polytope of nonnegative matrices with prescribed row and column sums. DTW instead minimizes over the finite, non-convex set of connected monotone paths through the pairwise cost matrix. Every time index must be visited, but it may be visited repeatedly; the row and column sums therefore record endogenous visit counts rather than prescribed masses. Normalizing a path matrix produces a coupling only for these path-dependent marginals, not for fixed input histograms. Conversely, ordinary OT between the unordered empirical feature measures forgets chronology and may match indices in a crossing order. Temporal penalties, causal constraints, and joint OT–DTW models interpolate between the two viewpoints; for example, spatio-temporal alignment combines regularized OT for spatial comparison with soft-DTW for chronological alignment [290]. Here "dynamic" refers to Bellman's dynamic programming on the index grid, not to the transport PDEs of Chapter 14.

Discrete variational problem. Let $x = (x_i)_{i=1}^n$ and $y = (y_j)_{j=1}^m$ be two sequences in a feature space $\mathcal{Z}$, and set $\mathrm{C}_{i,j} = c(x_i, y_j)$ for a nonnegative cost $c : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}_+$. A warping path is a sequence

$$\omega = ((i_\ell, j_\ell))_{\ell=1}^L$$

that starts at $(1,1)$, ends at (n,m), and has increments

$$(i_{\ell+1} - i_\ell, j_{\ell+1} - j_\ell) \in \{(1,0), (0,1), (1,1)\}.$$

Denote the set of such paths by $\Omega_{n,m}$ and the corresponding incidence matrix by $(A_\omega)_{i,j} = \mathbf{1}_{\{(i,j)\in\omega\}}$. Its length satisfies $\max\{n,m\} \leqslant L \leqslant n+m-1$, its total mass is $\sum_{i,j}(A_\omega)_{i,j} = L$, and its two marginals are precisely the row and column visit counts.

Definition 13.56 (Dynamic time warping)**.** The DTW value associated with the feature cost c is

$$\mathrm{DTW}_c(x,y) \coloneqq \min_{\omega\in\Omega_{n,m}} \sum_{\ell=1}^{L} c(x_{i_\ell}, y_{j_\ell}) = \min_{\omega\in\Omega_{n,m}} \langle A_\omega, \mathrm{C}\rangle. \tag{13.65}$$

The definition is symmetric when c is symmetric, but it is generally not a metric: repetitions can give zero cost to distinct sequences, and the triangle inequality can fail. Step weights, slope constraints and a Sakoe–Chiba band are common variants that penalize excessive repetition or restrict the admissible temporal distortion [464, 393].

Dynamic programming. The monotone path structure converts the exponentially large variational problem (13.65) into a shortest-path computation on an acyclic grid.

Proposition 13.57 (DTW dynamic-programming recurrence)**.** *Set $D_{0,0} = 0$, $D_{i,0} = D_{0,j} = +\infty$ for $i, j > 0$, and, for $1 \leqslant i \leqslant n$, $1 \leqslant j \leqslant m$, define*

$$D_{i,j} = \mathrm{C}_{i,j} + \min\{D_{i-1,j}, D_{i,j-1}, D_{i-1,j-1}\}. \tag{13.66}$$

Then $D_{n,m} = \mathrm{DTW}_c(x,y)$. Storing one minimizing predecessor at every cell and backtracking from (n,m) recovers an optimal path. The computation takes $O(nm)$ time and $O(nm)$ memory; if only the value is required, two rows give $O(\min\{n,m\})$ memory.

Proof. Every admissible path reaching (i,j) enters it from exactly one of $(i-1,j)$, $(i,j-1)$ or $(i-1,j-1)$. Removing the last cell therefore leaves an admissible path to one of these predecessors, while appending (i,j) to any predecessor path gives an admissible path to (i,j). Minimizing over the three mutually exhaustive last steps proves (13.66) by induction on $i+j$. The complexity follows because each of the nm cells performs constant work. □

Algorithm 13.4 makes the value computation and path recovery explicit.

Algorithm 13.4: Dynamic time warping by dynamic programming

Input: Sequences $x = (x_i)_{i=1}^n$, $y = (y_j)_{j=1}^m$, local cost c.
Output: DTW value $D_{n,m}$ and an optimal warping path $\omega^\star$.
Initialize: Set $D \in (\mathbb{R} \cup \{+\infty\})^{(n+1)\times(m+1)}$ to $+\infty$, $D_{0,0} = 0$, and allocate predecessors B.
For $i = 1, \dots, n$ **do**:
  For $j = 1, \dots, m$ **do**:
    Set $\mathrm{C}_{i,j} = c(x_i, y_j)$.
    Choose the lexicographically first $(r^\star, s^\star) \in \operatorname{argmin}_{(r,s)\in\{(i-1,j),(i,j-1),(i-1,j-1)\}} D_{r,s}$.
    Set $B_{i,j} = (r^\star, s^\star)$ and $D_{i,j} = \mathrm{C}_{i,j} + D_{r^\star,s^\star}$.

Initialize: Set $(i, j) = (n, m)$ and $\omega^\star = [\,]$.
While $(i, j) \neq (0, 0)$ **do**:
  Prepend (i, j) to $\omega^\star$.
  Set $(i, j) \leftarrow B_{i,j}$.

Return $D_{n,m}$ and $\omega^\star$.

Continuous time warping. Discrete DTW depends on the sampling density because every visited cell contributes once. A direct continuous registration minimizes $\int_0^1 c(x(t), y(\gamma(t)))\mathrm{d}t$ over nondecreasing endpoint-fixing maps γ, but this one-clock formulation is asymmetric and privileges the parameterization of x. Continuous DTW instead traverses both clocks and measures mismatch per unit length in the parameter square [85].

For simplicity, let $x, y : [0, 1] \to \mathcal{Z}$ use normalized clocks. The symmetric continuous registration problem uses two monotone clocks.

Definition 13.58 (Continuous dynamic time warping)**.** Let $\Gamma_\uparrow$ contain pairs (φ, ψ) of absolutely continuous, nondecreasing surjections of $[0, 1]$ onto itself. Define

$$\mathrm{CDTW}_c(x, y) := \inf_{(\varphi,\psi)\in\Gamma_\uparrow} \int_0^1 c\big(x(\varphi(s)), y(\psi(s))\big)\big(\dot\varphi(s) + \dot\psi(s)\big)\mathrm{d}s. \tag{13.67}$$

The surjections satisfy $\varphi(0) = \psi(0) = 0$ and $\varphi(1) = \psi(1) = 1$. Because $\dot\varphi, \dot\psi \geqslant 0$ almost everywhere, the last factor is the ℓ^1 line element of the monotone path $s \mapsto (\varphi(s), \psi(s))$. Formula (13.67) is invariant under increasing reparameterizations of the auxiliary variable s and does not privilege either clock; when c is symmetric, the resulting functional is also symmetric in x and y. After parameterization by ℓ^1 arc length, it is simply the line integral of the feature mismatch along a monotone path. For physical clock intervals $[0, p]$ and $[0, q]$, the same formula uses $\varphi : [0, 1] \to [0, p]$ and $\psi : [0, 1] \to [0, q]$. Exact computation is substantially harder than the discrete recurrence: for arc-length-parametrized one-dimensional polygonal curves and the standard cost $c(u, v) = |u - v|$, Buchin, Nusser and Wong propagate piecewise-quadratic boundary costs in $O((n + m)^5)$ time [85]. This complexity statement does not apply to an arbitrary feature cost c.

Soft-DTW and the Sinkhorn analogy. The hard minimum in (13.66) is nonsmooth when several paths tie. Soft-DTW replaces it with the log-sum-exp soft minimum [168],

$$\mathrm{softmin}_\varepsilon(r_1, \dots, r_q) = -\varepsilon \log\left(\sum_{k=1}^q e^{-r_k/\varepsilon}\right), \tag{13.68}$$

and defines $D^\varepsilon_{0,0} = 0$, $D^\varepsilon_{i,0} = D^\varepsilon_{0,j} = +\infty$ for $i, j > 0$, together with

$$D^\varepsilon_{i,j} = \mathrm{C}_{i,j} + \mathrm{softmin}_\varepsilon\big(D^\varepsilon_{i-1,j}, D^\varepsilon_{i,j-1}, D^\varepsilon_{i-1,j-1}\big), \qquad \mathrm{sDTW}_{c,\varepsilon}(x, y) = D^\varepsilon_{n,m}. \tag{13.69}$$

Equivalently, it is the free energy of all monotone paths,

$$\mathrm{sDTW}_{c,\varepsilon}(x, y) = -\varepsilon \log \sum_{\omega\in\Omega_{n,m}} \exp\left(-\frac{\langle A_\omega, \mathrm{C}\rangle}{\varepsilon}\right). \tag{13.70}$$

To make the regularization explicit, let $\Delta(\Omega_{n,m})$ be the simplex of probability laws $q = (q_\omega)_\omega$ over paths and let $H(q) = -\sum_\omega q_\omega \log q_\omega$ be their Shannon entropy, with $0 \log 0 = 0$. The Gibbs variational identity gives

$$\mathrm{sDTW}_{c,\varepsilon}(x, y) = \min_{q \in \Delta(\Omega_{n,m})} \left\{ \sum_\omega q_\omega \langle A_\omega, C \rangle - \varepsilon H(q) \right\}. \tag{13.71}$$

Its unique minimizer is the Gibbs law

$$\mathbb{P}_\varepsilon(\omega) = \frac{\exp(-\langle A_\omega, C \rangle / \varepsilon)}{\sum_{\omega'} \exp(-\langle A_{\omega'}, C \rangle / \varepsilon)}, \qquad E_\varepsilon := \nabla_C \mathrm{sDTW}_{c,\varepsilon}(x, y). \tag{13.72}$$

Indeed, subtracting the value in (13.70) from the objective in (13.71) gives $\varepsilon \, \mathrm{KL}(q | \mathbb{P}_\varepsilon) \geqslant 0$. Moreover,

$$\mathrm{DTW}_c(x, y) - \varepsilon \log |\Omega_{n,m}| \leqslant \mathrm{sDTW}_{c,\varepsilon}(x, y) \leqslant \mathrm{DTW}_c(x, y), \tag{13.73}$$

because the partition sum is bounded below by its largest term and above by $|\Omega_{n,m}|$ times that term. Hence $\mathrm{sDTW}_{c,\varepsilon} \to \mathrm{DTW}_c$ as $\varepsilon \to 0$. Forward and backward dynamic programs compute its value and gradient in $O(nm)$ time and $O(nm)$ memory [168]. Differentiating the finite log-partition function gives

$$E_\varepsilon = \mathbb{E}_{\omega \sim \mathbb{P}_\varepsilon}[A_\omega]. \tag{13.74}$$

Thus $(E_\varepsilon)_{i,j}$ is the probability that a Gibbs path visits cell (i, j). The matrix E_ε is a diffuse expected alignment and converges, when the hard optimum is unique, to its path-incidence matrix.

The analogy with entropic OT is now exact at the level of free energies, but not at the level of feasible variables. Sinkhorn regularizes a coupling with prescribed marginals, whereas soft-DTW regularizes the path law q in (13.71). Its mean E_ε generally has neither prescribed row sums nor prescribed column sums, and the path entropy $H(q)$ cannot in general be recovered from E_ε alone. Algorithmically, Sinkhorn uses alternating matrix scaling, while soft-DTW uses forward–backward dynamic programming on an acyclic grid. Global-alignment kernels sum the same Gibbs weights over all paths [172, 165].

Raw soft-DTW also has an entropic self-bias and can be negative. In direct analogy with the Sinkhorn divergence of Section 8.10, the soft-DTW divergence is

$$\overline{\mathrm{sDTW}}_{c,\varepsilon}(x, y) = \mathrm{sDTW}_{c,\varepsilon}(x, y) - \frac{1}{2}\mathrm{sDTW}_{c,\varepsilon}(x, x) - \frac{1}{2}\mathrm{sDTW}_{c,\varepsilon}(y, y). \tag{13.75}$$

The correction always vanishes on the diagonal, but positivity requires care.

> **Remark 13.59 (Validity of the soft-DTW divergence).** Blondel, Mensch and Vert prove that (13.75) is nonnegative and vanishes only for identical sequences when either $c(u, v) = |u - v|$ in one dimension or
>
> $$c(u, v) = \delta(u, v) + \log\bigl(2 - e^{-\delta(u,v)}\bigr), \qquad \delta(u, v) = \tfrac{1}{2}\|u - v\|^2.$$
>
> For the ordinary squared Euclidean cost used in Figure 13.8, they prove for equal-length sequences that the diagonal is stationary and provide numerical evidence for nonnegativity, but do not prove it in general [64]. Thus debiasing alone should not be taken as a universal positivity theorem. Even for the costs covered by the result, no triangle inequality is asserted.

Figure 13.8 contrasts the hard minimizing path with the expected alignment (13.74) on an exactly time-warped oscillatory signal.

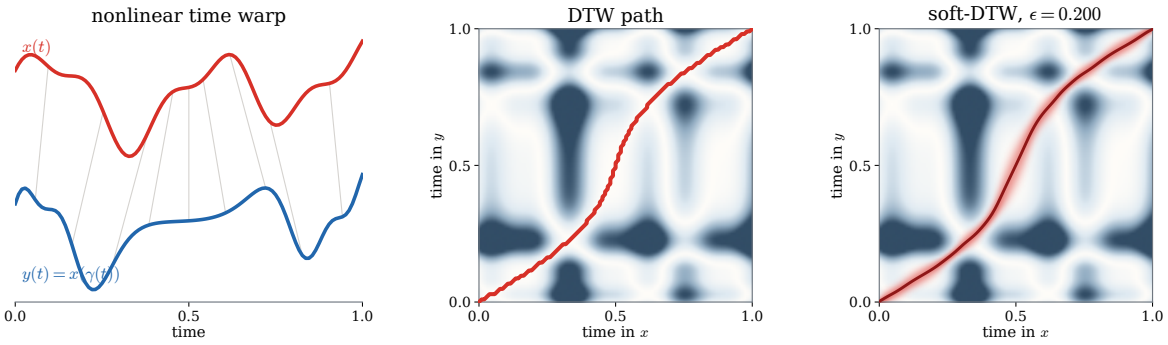


Figure 13.8: Hard and soft monotone alignments recover a nonlinear time warp. Left: an oscillatory signal x and the warped observation $y(t) = x(\gamma(t))$ for a smooth increasing map γ; thin gray segments mark exact corresponding times. Middle: the pairwise squared feature-cost matrix $C_{i,j} = |x_i - y_j|^2$, with the optimal DTW path shown in red. Right: the same matrix overlaid with the soft-DTW expected alignment E_ε from (13.74) at $\varepsilon = .200$; red intensity gives cell-visit probability and the dark red curve is its row-wise barycentric summary.

CHAPTER 14

Dynamic Optimal Transport

Optimal transport becomes especially powerful once distances between measures are seen as actions of moving mass. This chapter first develops the dynamic language: continuity equations describe admissible measure evolutions, while the Benamou–Brenier formula identifies $\mathcal{W}_2$ with a least-action principle. After extending this construction to other transport geometries, the chapter closes with variational mean field games, where congestion and a terminal cost turn the Benamou–Brenier action into a population-planning problem. These ideas prepare the gradient-flow and generative-model chapters that follow.

14.1 Evolutions over the Space of Measures

We start with the continuity equation because it is the common language for particles, densities and weak measure evolutions. It also makes precise which velocity fields actually move mass.

Lagrangian and Eulerian descriptions. We consider the evolution $t \mapsto \alpha_t \in \mathcal{P}(\mathbb{R}^d)$. Such an evolution can be described in a "Lagrangian" way as the advection of particles along a (time-dependent) vector field $v_t(x)$ in $\mathbb{R}^d$. At the particle level, this advection is governed by

$$\frac{\mathrm{d}x(t)}{\mathrm{d}t} = v_t(x(t)), \tag{14.1}$$

and we write T_t for the associated flow map, so that $T_t(x(0)) = x(t)$. The advected measure is then $\alpha_t = (T_t)_\sharp \alpha_0$. For discrete measures, $\alpha_t = \frac{1}{n}\sum_{i=1}^n \delta_{x_i(t)}$, meaning each $x_i(t)$ solves (14.1).

In the Eulerian description, the same motion is written directly on the evolving measure: the particle ODE becomes the PDE

$$\frac{\partial \alpha_t}{\partial t} + \operatorname{div}(v_t \alpha_t) = 0. \tag{14.2}$$

This PDE is often referred to as the advection equation, the continuity equation, or Liouville's equation when operating over a phase space. It is only a classical PDE when α_t has a smooth density. For general measures, and in particular for empirical measures, it is understood in the distributional sense: for every $\varphi \in C_c^1((0,1)\times\mathbb{R}^d)$,

$$\int_0^1 \int_{\mathbb{R}^d} (\partial_t \varphi(t,x) + \langle v_t(x), \nabla_x \varphi(t,x)\rangle)\, \mathrm{d}\alpha_t(x)\mathrm{d}t = 0. \tag{14.3}$$

This equation is obtained from (14.2) by integration by parts. Hence, for smooth positive densities, the classical and weak formulations are equivalent; the weak formulation is useful because it still makes sense for discrete measures whose particles evolve according to (14.1).

This weak identity provides the measure-theoretic notion of an admissible evolution.

Definition 14.1 (Admissible continuity-equation evolution)**.** A narrowly continuous curve $(\alpha_t)_{t\in[0,1]}$ and a Borel velocity field v are admissible between α_0 and α_1 if

$$\int_0^1 \int \|v_t(x)\|^2 \mathrm{d}\alpha_t(x)\mathrm{d}t < +\infty$$

and

$$\partial_t \alpha_t + \operatorname{div}(\alpha_t v_t) = 0$$

in distributions, with the prescribed endpoints and no flux through the boundary when the domain is bounded.

Proposition 14.2 (Lagrangian flows solve the continuity equation)**.** *Let $(T_t)_{t\in[0,1]}$ be a C^1 family of diffeomorphisms of $\mathbb{R}^d$ and define $\alpha_t = (T_t)_\sharp \alpha_0$. Assume that the derivatives below are integrable with respect to*

α_0, and define the Eulerian velocity field by

$$v_t(T_t(y)) = \partial_t T_t(y).$$

Then (α_t, v_t) solves the continuity equation in the weak sense (14.3). In particular, if $\alpha_0 = \frac{1}{n}\sum_i \delta_{x_i(0)}$ is empirical, then $\alpha_t = \frac{1}{n}\sum_i \delta_{x_i(t)}$ is empirical as well, with particle velocities $\dot{x}_i(t) = v_t(x_i(t))$.

Proof. Let $\varphi \in C_c^1((0,1)\times\mathbb{R}^d)$. Since $\alpha_t = (T_t)_\sharp \alpha_0$,

$$\frac{\mathrm{d}}{\mathrm{d}t}\int \varphi(t,x)\mathrm{d}\alpha_t(x) = \frac{\mathrm{d}}{\mathrm{d}t}\int \varphi(t,T_t(y))\mathrm{d}\alpha_0(y).$$

The chain rule gives

$$\frac{\mathrm{d}}{\mathrm{d}t}\int \varphi(t,T_t(y))\mathrm{d}\alpha_0(y) = \int \left(\partial_t\varphi(t,T_t(y)) + \langle \nabla_x\varphi(t,T_t(y)),\, \partial_t T_t(y)\rangle\right)\mathrm{d}\alpha_0(y).$$

Using the definition of v_t and the push-forward relation, this equals

$$\int \left(\partial_t\varphi(t,x) + \langle \nabla_x\varphi(t,x),\, v_t(x)\rangle\right)\mathrm{d}\alpha_t(x).$$

Integrating in time and using the boundary values of φ gives (14.3). □

From measure evolutions to vector fields. For a given evolution $(\alpha_t)_t$, there are typically infinitely many velocity fields v_t satisfying

$$\partial_t \alpha_t + \operatorname{div}(\alpha_t v_t) = 0. \tag{14.4}$$

This non-uniqueness comes from the kernel of the weighted divergence. The linear space of vector fields that leave a measure α invariant is

$$\mathcal{H}_\alpha = \left\{ v \in L^2(\alpha;\mathbb{R}^d)\ ;\ \operatorname{div}(\alpha v) = 0 \text{ in distributions} \right\}.$$

It is usually non-trivial: for instance, if α is an isotropic Gaussian, $\mathcal{H}_\alpha$ contains rotational vector fields generated by anti-symmetric matrices.

Dacorogna–Moser inversion. Reconstructing particles from an observed density evolution is therefore ill-posed. For a smooth positive density $\alpha_t = \rho_t\,\mathrm{d}x$, a simple choice, introduced by Dacorogna and Moser [173], is to impose that the flux $\rho_t v_t$ is a gradient field. With a fixed convention for the inverse Laplacian, this gives formally

$$v_t = -\frac{1}{\rho_t}\nabla\Delta^{-1}(\partial_t\rho_t), \tag{14.5}$$

with suitable boundary conditions, for instance vanishing at infinity. This formula is useful conceptually but delicate when ρ_t vanishes, and it does not generally produce a gradient velocity field.

The classical Dacorogna–Moser construction uses the linear density path. If $\alpha_i = \rho_i\,\mathrm{d}x$ are smooth positive densities with the same total mass on a bounded domain Ω, set

$$\alpha_t = (1-t)\alpha_0 + t\alpha_1 = \rho_t\,\mathrm{d}x, \qquad \rho_t = (1-t)\rho_0 + t\rho_1.$$

Choose a time-independent flux w satisfying

$$\operatorname{div} w = \rho_0 - \rho_1, \qquad w\cdot n = 0 \quad \text{on } \partial\Omega,$$

for instance $w = -\nabla\varphi$ with $\Delta\varphi = \rho_1 - \rho_0$ and Neumann boundary condition. Then

$$v_t = \frac{w}{\rho_t}$$

satisfies $\partial_t\rho_t + \operatorname{div}(\rho_t v_t) = 0$. The flow $\partial_t T_t = v_t\circ T_t$, $T_0 = \mathrm{Id}$, therefore transports $\rho_0\mathrm{d}x$ onto $\rho_t\mathrm{d}x$, and T_1 solves the prescribed-Jacobian problem $\rho_1(T_1(x))\det(\nabla T_1(x)) = \rho_0(x)$.

Least-square inversion and gradient structure. A more robust choice, used implicitly in flow matching, optimal transport and Wasserstein gradient flows, is to select among all admissible velocities the one with smallest kinetic energy:

$$\min_{v} \frac{1}{2}\int_0^1 \int_{\mathbb{R}^d} \|v_t(x)\|^2 \,\mathrm{d}\alpha_t(x)\mathrm{d}t \quad \text{subject to} \quad \partial_t \alpha_t + \operatorname{div}(\alpha_t v_t) = 0. \tag{14.6}$$

Proposition 14.3 (Least-square velocities are gradients)**.** *Assume that $\alpha_t = \rho_t \,\mathrm{d}x$ is a smooth positive density curve, that $\partial_t \rho_t$ has zero integral, and that boundary terms vanish. The minimizer of (14.6), if it exists, is a gradient field*

$$v_t = \nabla \varphi_t,$$

where φ_t, unique up to an additive constant on each connected component, solves the weighted Poisson equation

$$-\operatorname{div}(\rho_t \nabla \varphi_t) = \partial_t \rho_t, \qquad v_t = -\nabla \Delta_{\alpha_t}^{-1}(\partial_t \alpha_t), \quad \Delta_{\alpha_t}\varphi = \operatorname{div}(\alpha_t \nabla \varphi). \tag{14.7}$$

Proof. Introduce a Lagrange multiplier φ_t for the continuity equation. The constrained problem has the formal saddle formulation

$$\min_{v} \max_{\varphi} \int_0^1 \left[\frac{1}{2}\int_{\mathbb{R}^d} \|v_t(x)\|^2 \,\mathrm{d}\alpha_t(x) + \int_{\mathbb{R}^d} \varphi_t(x)\big(\operatorname{div}(\alpha_t v_t)(x) + \partial_t \alpha_t(x)\big)\mathrm{d}x\right] \mathrm{d}t.$$

Integrating by parts in the divergence term gives, for each t,

$$\int \left(\frac{1}{2}\|v_t\|^2 - \langle \nabla \varphi_t,\, v_t\rangle\right) \mathrm{d}\alpha_t + \int \varphi_t \,\partial_t \alpha_t.$$

The pointwise minimizer in v_t is therefore $v_t = \nabla \varphi_t$. Substituting this into the constraint $\partial_t \rho_t + \operatorname{div}(\rho_t v_t) = 0$ gives the weighted Poisson equation in (14.7). The inverse notation is just a shorthand for solving this equation on zero-mean right-hand sides, modulo additive constants. □

In general this inversion is still computationally demanding, but special choices of $(\alpha_t)_t$ lead to simpler formulas; this is the mechanism exploited later by flow matching in Section 16.1.

14.2 Benamou–Brenier dynamic formulation of OT

The dynamic formulation identifies $\mathcal{W}_2$ with the kinetic energy of the cheapest continuity-equation path. It is the point where OT becomes a least-action principle.

Benamou–Brenier formulation. Instead of assuming that a whole curve $(\alpha_t)_{t\in[0,1]}$ is prescribed, one only fixes its endpoints α_0 and α_1 and minimizes the least-square energy (14.6). The theorem of Benamou and Brenier states that this geodesic energy is exactly the squared Wasserstein distance [35].

The kinetic energy of an admissible evolution is the basic dynamic action.

Definition 14.4 (Benamou–Brenier action)**.** For an admissible pair (α_t, v_t), define

$$\mathrm{Act}_2(\alpha, v) := \int_0^1 \int_{\mathbb{R}^d} \|v_t(x)\|^2 \mathrm{d}\alpha_t(x)\mathrm{d}t.$$

Theorem 14.5 (Benamou–Brenier)**.** *For probability measures $\alpha_0, \alpha_1 \in \mathcal{P}_2(\mathbb{R}^d)$,*

$$\mathcal{W}_2^2(\alpha_0, \alpha_1) = \inf_{(\alpha_t, v_t)} \int_0^1 \int_{\mathbb{R}^d} \|v_t(x)\|^2 \,\mathrm{d}\alpha_t(x)\mathrm{d}t, \tag{14.8}$$

where the infimum is over (α_t, v_t) solving $\partial_t \alpha_t + \nabla\cdot(\alpha_t v_t) = 0$ with $\alpha_{t=0} = \alpha_0$ and $\alpha_{t=1} = \alpha_1$. If α_0 has a density and T is the optimal Monge map $T_\sharp \alpha_0 = \alpha_1$, a minimizing curve is

$$\alpha_t = ((1-t)\mathrm{Id} + tT)_\sharp \alpha_0, \qquad v_t\big((1-t)x + tT(x)\big) = T(x) - x \quad \text{for } \alpha_0\text{-a.e. } x \text{ and a.e. } t \in (0,1). \tag{14.9}$$

Proof. For the inequality "dynamic $\leqslant$ static", assume first that a Monge map T exists and define (α_t, v_t) by (14.9). Since the Lagrangian velocity $T(x) - x$ is independent of t,

$$\int_0^1 \int \|v_t\|^2 \, \mathrm{d}\alpha_t \mathrm{d}t = \int \|T(x) - x\|^2 \, \mathrm{d}\alpha_0(x),$$

so the dynamic cost is no larger than the static Monge cost. Without a Monge map, the same construction is made with an optimal coupling π: sample $(X, Y) \sim \pi$ and move along the straight path $\gamma_{X,Y}(t) = (1-t)X + tY$. This path measure has action $\int \|x - y\|^2 \mathrm{d}\pi(x, y)$; projecting path velocities onto their conditional mean at time t gives an admissible Eulerian velocity with no larger action, so the dynamic value is no larger than the Kantorovich value.

Conversely, for a smooth deterministic path, take the flow T_t defined by $\dot{T}_t = v_t \circ T_t$ and $T_0 = \mathrm{Id}$. Then $\alpha_t = (T_t)_\sharp \alpha_0$ and $(T_1)_\sharp \alpha_0 = \alpha_1$. Jensen's inequality gives

$$\|T_1(x) - x\|^2 \leqslant \int_0^1 \|v_t(T_t(x))\|^2 \mathrm{d}t.$$

After integration with respect to α_0, the Monge cost is bounded above by the dynamic action. For general finite-energy solutions of the continuity equation, the superposition principle lifts the curve to a probability measure on absolutely continuous paths; applying Jensen's inequality pathwise gives a coupling of the endpoints whose quadratic cost is no larger than the action. Thus the Kantorovich value is bounded above by the dynamic value. □

Convex moment-based reformulation. Although (14.8) is not jointly convex in (α_t, v_t), it becomes convex after replacing velocities by momenta. Given a velocity $v \in L^2(\alpha; \mathbb{R}^d)$, define the momentum

$$\omega \coloneqq \alpha v, \qquad \omega(A) = \int_A v(x) \, \mathrm{d}\alpha(x),$$

which is a finite $\mathbb{R}^d$-valued measure. The nonlinear relation $\omega = \alpha v$ is eliminated by the following measure perspective.

Definition 14.6 (Quadratic perspective action)**.**

$$J(a, m) \coloneqq \begin{cases} \|m\|^2/a, & a > 0, \\ 0, & a = 0 \text{ and } m = 0, \\ +\infty, & a = 0 \text{ and } m \neq 0, \end{cases} \qquad (a, m) \in [0, +\infty) \times \mathbb{R}^d. \tag{14.10}$$

If λ dominates α and $|\omega|$, define

$$\mathbb{J}(\alpha, \omega) \coloneqq \int J\left(\frac{\mathrm{d}\alpha}{\mathrm{d}\lambda}(x), \frac{\mathrm{d}\omega}{\mathrm{d}\lambda}(x)\right) \mathrm{d}\lambda(x). \tag{14.11}$$

The pointwise function is lower-semicontinuous, convex and positively 1-homogeneous: $J(\eta a, \eta m) = \eta J(a, m)$ for $\eta \geqslant 0$. The definition is independent of the dominating measure λ: after replacing λ by another dominating measure, both Radon–Nikodym densities are rescaled by the same factor, which cancels with the change of reference measure by the 1-homogeneity of J. Equivalently, this is the standard integral functional associated with a convex normal integrand, used in the measure-valued relaxation of dynamic OT [12]; see also Rockafellar's perspective construction [452]. In particular,

$$\mathbb{J}(\alpha, \omega) < +\infty \quad \Longleftrightarrow \quad \omega = v\alpha \text{ with } v \in L^2(\alpha; \mathbb{R}^d), \qquad \mathbb{J}(\alpha, \omega) = \int \|v\|^2 \, \mathrm{d}\alpha.$$

The Benamou–Brenier problem can now be written with the linear continuity equation in the pair (α_t, ω_t):

$$\mathcal{W}_2^2(\alpha_0, \alpha_1) = \inf_{\substack{\partial_t \alpha_t + \operatorname{div} \omega_t = 0 \\ \alpha_{t=0} = \alpha_0, \ \alpha_{t=1} = \alpha_1}} \int_0^1 \mathbb{J}(\alpha_t, \omega_t) \, \mathrm{d}t. \tag{14.12}$$

In the absolutely continuous case $\alpha_t = \rho_t\,\mathrm{d}x$ and $\omega_t = m_t\,\mathrm{d}x$, this becomes

$$\inf_{\substack{\partial_t\rho_t+\operatorname{div} m_t=0\\ \rho_{t=0}\mathrm{d}x=\alpha_0,\ \rho_{t=1}\mathrm{d}x=\alpha_1}} \int_0^1\int_{\mathbb{R}^d} J(\rho_t(x),m_t(x))\,\mathrm{d}x\,\mathrm{d}t = \inf_{\substack{\partial_t\rho_t+\operatorname{div} m_t=0\\ \rho_{t=0}\mathrm{d}x=\alpha_0,\ \rho_{t=1}\mathrm{d}x=\alpha_1}} \int_0^1\int_{\mathbb{R}^d} \frac{\|m_t(x)\|^2}{\rho_t(x)}\,\mathrm{d}x\,\mathrm{d}t,$$

with the conventions encoded in (14.10). This convex reformulation enables geodesic interpolation by convex optimization once the domain is discretized.

Dual Hamilton–Jacobi formulation. The momentum formulation also has a useful dual. It turns the least-action problem into a Hamilton–Jacobi subsolution inequality for a scalar potential, with equality on the part of space-time actually visited by the optimal curve. This is the dual dynamic formulation already present in the original Benamou–Brenier perspective [35]; see also the standard accounts [538, 469]. The constants below follow the convention of (14.12), where the kinetic energy has no factor $1/2$.

Proposition 14.7 (Dual Benamou–Brenier problem)**.** *Assume, for simplicity, that the densities are smooth, compactly supported, and that boundary terms vanish. Then the convex dynamic value* (14.12) *has the dual formulation*

$$\mathcal{W}_2^2(\alpha_0,\alpha_1) = \sup_{\varphi}\left\{\int_{\mathbb{R}^d}\varphi_1\,\mathrm{d}\alpha_1 - \int_{\mathbb{R}^d}\varphi_0\,\mathrm{d}\alpha_0\ ;\ \partial_t\varphi_t + \frac{1}{4}\|\nabla\varphi_t\|^2 \leqslant 0\right\}, \tag{14.13}$$

where the supremum is over smooth test potentials $\varphi : [0,1]\times\mathbb{R}^d \to \mathbb{R}$. In the general metric-measure statement, the same formula is understood after relaxation: φ is a Hamilton–Jacobi subsolution, for instance in the viscosity sense, and finite-dimensional discretizations follow directly from Fenchel–Rockafellar duality.

If (ρ, m) and φ are smooth primal and dual optimizers, then

$$m_t = \frac{\rho_t}{2}\nabla\varphi_t, \qquad \partial_t\varphi_t + \frac{1}{4}\|\nabla\varphi_t\|^2 = 0 \quad \text{on } \{\rho_t > 0\}. \tag{14.14}$$

Equivalently, the optimal Eulerian velocity is $v_t = m_t/\rho_t = \nabla\varphi_t/2$.

Proof. Let (ρ, m) satisfy the continuity equation and let φ be smooth. Multiplying the constraint by φ and integrating by parts gives

$$\int_{\mathbb{R}^d}\varphi_1\,\mathrm{d}\alpha_1 - \int_{\mathbb{R}^d}\varphi_0\,\mathrm{d}\alpha_0 = \int_0^1\int_{\mathbb{R}^d}(\rho_t\,\partial_t\varphi_t + \langle m_t,\,\nabla\varphi_t\rangle)\,\mathrm{d}x\mathrm{d}t.$$

If $\partial_t\varphi_t + \|\nabla\varphi_t\|^2/4 \leqslant 0$, Young's inequality yields, pointwise,

$$\rho\,\partial_t\varphi + \langle m,\,\nabla\varphi\rangle \leqslant -\frac{\rho}{4}\|\nabla\varphi\|^2 + \langle m,\,\nabla\varphi\rangle \leqslant J(\rho,m),$$

with the perspective convention (14.10). Hence the dual objective of every feasible φ is no larger than the action of every feasible primal pair.

Conversely, introduce φ as a Lagrange multiplier for $\partial_t\rho + \operatorname{div} m = 0$. After integration by parts, and up to the fixed endpoint contribution in (14.13), the pointwise minimization over (ρ, m) contains

$$J(\rho,m) - \rho\,\partial_t\varphi - \langle m,\,\nabla\varphi\rangle.$$

For fixed $\rho > 0$, its minimum over m is attained at $m = \rho\nabla\varphi/2$ and equals $-\rho(\partial_t\varphi + \|\nabla\varphi\|^2/4)$. The recession cases in (14.10) give the same conclusion when $\rho = 0$. Thus the infimum over $\rho \geqslant 0$ is finite exactly when the Hamilton–Jacobi inequality holds, in which case the pointwise infimum is 0. Fenchel–Rockafellar duality then gives no duality gap. Equality in the two pointwise inequalities gives (14.14). □

This also recovers the static Kantorovich inequality from a dynamic principle. If γ is any smooth curve with $\gamma(0) = x$ and $\gamma(1) = y$, then

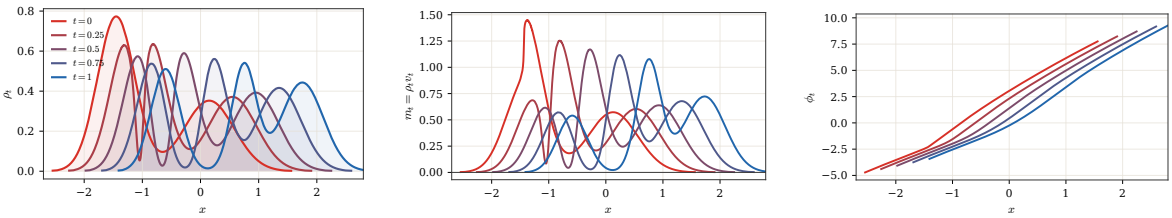
$$\frac{\mathrm{d}}{\mathrm{d}t}\varphi_t(\gamma(t)) = \partial_t\varphi_t(\gamma(t)) + \langle\nabla\varphi_t(\gamma(t)),\,\dot\gamma(t)\rangle \leqslant \|\dot\gamma(t)\|^2,$$

where the last inequality uses $\partial_t \varphi + \|\nabla \varphi\|^2/4 \leqslant 0$ and Young's inequality. Integrating and minimizing over curves gives

$$\varphi_1(y) - \varphi_0(x) \leqslant \|x - y\|^2.$$

Thus $(-\varphi_0, \varphi_1)$ is a feasible static Kantorovich dual pair for the quadratic cost. At optimality the inequality is saturated on the endpoint pairs connected by the primal characteristics, which is the Eulerian version of complementary slackness.

Figure 14.1 displays these primal–dual relations for a one-dimensional mixture transport, including the Hamilton–Jacobi contact identity along the active mass.



density ρ_t momentum density m_t dual potential φ_t

Figure 14.1: One-dimensional Benamou–Brenier primal and dual solutions. The endpoints are Gaussian mixtures and the solution is computed from monotone quantile interpolation. The left and middle panels show the primal density and the momentum density $m_t = \rho_t v_t$. The right panel shows the dual Hamilton–Jacobi potential, normalized up to an additive constant. Along the active transported mass, the notebook checks the primal–dual identities $m_t = \rho_t \partial_x \varphi_t / 2$ and $\partial_t \varphi_t + |\partial_x \varphi_t|^2/4 = 0$.

Proximal splitting. The convex momentum formulation also explains the original Benamou–Brenier solver. Papadakis, Peyré and Oudet [414] showed that, after discretization, the ALG2 scheme is a Douglas–Rachford splitting, equivalently ADMM on the Fenchel–Rockafellar dual. Suppressing discretization indices, write $U = (\rho, m)$, where m is the density of the momentum measure. Using the perspective (14.10), set

$$\mathcal{F}(U) = \int_0^1 \int_{\mathbb{R}^d} J(\rho_t(x), m_t(x))\, \mathrm{d}x\, \mathrm{d}t, \qquad C = \{(\rho, m)\ ;\ \partial_t \rho + \operatorname{div} m = 0,\ \rho_0 \mathrm{d}x = \alpha_0,\ \rho_1 \mathrm{d}x = \alpha_1\},$$

and let $\mathcal{G} = \iota_C$ be the indicator of this affine continuity constraint. The convex Benamou–Brenier problem is therefore

$$\min_U \ \mathcal{F}(U) + \mathcal{G}(U).$$

On the resulting Hilbert space of unknowns, or formally for an L^2-type product structure, the two proximal operators are

$$\operatorname{prox}_{\tau\mathcal{F}}(\bar U) = \operatorname{argmin}_U \frac{1}{2}\|U - \bar U\|_{\mathsf{H}}^2 + \tau \mathcal{F}(U), \qquad \operatorname{prox}_{\tau\mathcal{G}}(\bar U) = \operatorname{argmin}_{U \in C} \frac{1}{2}\|U - \bar U\|_{\mathsf{H}}^2.$$

For the standard product metric, the first map is local in (t, x): it is the proximal operator of the convex perspective J. The second map is the orthogonal projection onto the affine set defined by the divergence equation and endpoint constraints. Douglas–Rachford then alternates these two simple operations:

$$\begin{aligned} U^{(\ell+1)} &= \operatorname{prox}_{\tau\mathcal{F}}(Z^{(\ell)}), \\ \widetilde U^{(\ell+1)} &= \operatorname{prox}_{\tau\mathcal{G}}(2U^{(\ell+1)} - Z^{(\ell)}), \\ Z^{(\ell+1)} &= Z^{(\ell)} + \widetilde U^{(\ell+1)} - U^{(\ell+1)}. \end{aligned}$$

Equivalently one may swap the roles of $\mathcal{F}$ and $\mathcal{G}$. At convergence, the two shadow sequences $U^{(\ell)}$ and $\widetilde U^{(\ell)}$ agree and give a minimizer of the convex dynamic problem. This viewpoint is useful because it separates the nonlinear but pointwise perspective proximal step from the global but linear projection onto the continuity equation [414, 201].

Figure 14.2 complements the Eulerian optimization viewpoint with the Lagrangian picture: matched particles travel along the straight characteristics of the minimizing curve.

Algorithm 14.1: Douglas–Rachford for dynamic Benamou–Brenier

Input: Functionals $\mathcal{F}, \mathcal{G} = \iota_C$, proximal parameter $\tau > 0$, initial field $Z^{(0)}$, tolerance tol > 0, iteration budget $L \geqslant 1$.
Output: Discrete density-momentum field $U^\star$.
For $\ell = 0, \dots, L-1$ **do**:
  $U^{(\ell+1)} = \mathrm{prox}_{\tau\mathcal{F}}(Z^{(\ell)})$.
  Project reflected point: $\widetilde{U}^{(\ell+1)} = \mathrm{prox}_{\tau\mathcal{G}}(2U^{(\ell+1)} - Z^{(\ell)}) = \mathrm{Proj}_C(2U^{(\ell+1)} - Z^{(\ell)})$.
  Update $Z^{(\ell+1)} = Z^{(\ell)} + \widetilde{U}^{(\ell+1)} - U^{(\ell+1)}$.
  If $\|U^{(\ell+1)} - \widetilde{U}^{(\ell+1)}\| \leqslant$ tol **then**:
    Return $U^{(\ell+1)}$.
Return $U^{(L)}$.

density sequence midpoint velocities

Figure 14.2: Benamou–Brenier geodesic between two sampled silhouettes. A discrete quadratic OT plan between finely subsampled cat and two-disks point clouds induces the McCann interpolation $Z_t = (1-t)X + tY$, which is the Lagrangian realization of the least-action solution. The left panel renders local color images of the smaller-bandwidth kernel-smoothed densities with enough padding to include the full silhouettes. The right panel overlays shortened velocity arrows centered at evenly subsampled midpoint particles $Z_{1/2}$; each displayed arrow runs in data coordinates from a source-side tail to a target-side head along the matched characteristic direction $Y - X$, but is not drawn as the full endpoint segment from X to Y.

Path-space formulation. Let $\mathcal{S} = C([0,1]; \mathbb{R}^d)$ be the space of continuous paths endowed with the uniform topology. For $t \in [0,1]$ define the evaluation map

$$e_t : \mathcal{S} \to \mathbb{R}^d, \qquad e_t(\gamma) = \gamma(t).$$

The Benamou–Brenier cost admits the equivalent formulation

$$\mathcal{W}_2^2(\alpha_0, \alpha_1) = \inf_{M \in \mathcal{P}(\mathcal{S})} \left\{ \int_{\mathcal{S}} \int_0^1 \|\dot\gamma(t)\|^2 \mathrm{d}t \, \mathrm{d}M(\gamma) \; ; \; (e_0)_\sharp M = \alpha_0, \; (e_1)_\sharp M = \alpha_1 \right\}.$$

The inner kinetic energy is understood as $+\infty$ when γ is not absolutely continuous. If α_0 has a density, the minimizer M^* is unique. Its time marginals reproduce the optimal curve: $\alpha_t = (e_t)_\sharp M^*$ for all t. Furthermore, for a.e. t, denoting $\dot e_t(\gamma) = \dot\gamma(t)$ on absolutely continuous paths, the conditional law of the velocity is deterministic:

$$(e_t, \dot e_t)_\sharp M^*(\mathrm{d}x, \mathrm{d}q) = \alpha_t(\mathrm{d}x)\delta_{v_t^*(x)}(\mathrm{d}q),$$

where v_t^* is the optimal velocity field in the Benamou–Brenier formulation. Hence M^* concentrates on straight-line geodesics and, for a.e. t, assigns exactly one direction at α_t-a.e. spatial point.

14.3 Generalized Dynamic Wasserstein Distances

The quadratic Benamou–Brenier formula is only one instance of a broader fixed-mass dynamic language. The goal of this section is to define a large family of geodesic-like distances on spaces of probability measures by modifying the action minimized in the Benamou–Brenier formula. The objects introduced here are metric: they specify admissible curves, tangent variables and path energies. All descent constructions are postponed to Section 15.5, where these distances are used to generate gradient-flow PDE models.

Path actions. The common starting point is an instantaneous cost for moving a measure with a prescribed Eulerian velocity. Integrating this cost along continuity-equation curves produces the associated endpoint geometry.

In the mass-preserving Euclidean setting, the basic input is an instantaneous action $\mathbb{A}(\alpha, w)$, where α is the current measure and w is an admissible velocity representative. Its endpoint geometry is defined dynamically.

Definition 14.8 (Generalized dynamic action distance)**.** Let $\mathbb{A}$ be a nonnegative extended-valued action on measure–velocity pairs. Define

$$\mathsf{D}^2_{\mathbb{A}}(\alpha_0, \alpha_1) := \inf_{\alpha_t, v_t} \left\{ \int_0^1 \mathbb{A}(\alpha_t, v_t)\, \mathrm{d}t : \partial_t \alpha_t + \operatorname{div}(\alpha_t v_t) = 0,\ \alpha_{t=0} = \alpha_0,\ \alpha_{t=1} = \alpha_1 \right\}. \tag{14.15}$$

Equivalently, one may quotient by velocity fields that induce the same first-order variation of the measure. The formula (14.15) should be read as a dynamic definition of the distance, not as a property automatically satisfied by an arbitrary discrepancy. Some standard distances, such as $\mathcal{W}_p$, are first written with a p-homogeneous action and then squared by taking a constant-speed parametrization; this normalization is made explicit below. Different choices of $\mathbb{A}$ change the resulting geometry; Section 15.5 later reuses these choices when dynamics are introduced.

Quadratic, or Riemannian, tangent actions. A particularly transparent case occurs when $w \mapsto \mathbb{A}(\alpha, w)$ is quadratic. For simplicity, take admissible velocities in $L^2(\alpha; \mathbb{R}^d)$; in some applications this Hilbert space is replaced by a closed subspace encoding additional constraints. If the polarization of $\mathbb{A}$ is represented by a positive self-adjoint operator $Q_\alpha : L^2(\alpha; \mathbb{R}^d) \to L^2(\alpha; \mathbb{R}^d)$,

$$\mathbb{A}(\alpha; w, z) = \langle Q_\alpha w, z \rangle_{L^2(\alpha)}, \qquad \mathbb{A}(\alpha, w) = \langle Q_\alpha w, w \rangle_{L^2(\alpha)},$$

and if the quadratic form is nondegenerate after quotienting velocities that induce the same measure variation, then the least-action distance generated by this tensor is

$$\mathsf{D}^2_Q(\alpha_0, \alpha_1) = \mathsf{D}^2_{\mathbb{A}}(\alpha_0, \alpha_1) = \inf_{\substack{\partial_t \alpha_t + \operatorname{div}(\alpha_t v_t) = 0 \\ \alpha_{t=0} = \alpha_0,\ \alpha_{t=1} = \alpha_1}} \int_0^1 \langle Q_{\alpha_t} v_t, v_t \rangle_{L^2(\alpha_t)}\, \mathrm{d}t. \tag{14.16}$$

The usual $\mathcal{W}_2$ geometry corresponds to $Q_\alpha = \mathrm{Id}$ in this simplified notation. Thus Q_α records how the chosen geometry deforms the Euclidean $L^2(\alpha)$ tangent norm: no deformation for $\mathcal{W}_2$, and a nontrivial tensor for generalized Riemannian geometries. In Section 15.5, the same tensor is reused as a preconditioner for metric descent.

Local velocity actions. Many dynamic distances are local with respect to a reference measure λ. Write $\alpha = a\lambda$. A velocity action is specified by a pointwise integrand

$$A : [0, +\infty) \times \mathbb{R}^d \to [0, +\infty], \qquad (a, w) \mapsto A(a, w),$$

where $a \in \mathbb{R}_+$ is a density value and $w \in \mathbb{R}^d$ is a velocity value, and defines

$$\mathbb{A}(\alpha, w) = \int A\left(\frac{\mathrm{d}\alpha}{\mathrm{d}\lambda}(x), w(x)\right) \mathrm{d}\lambda(x). \tag{14.17}$$

For a fixed reference λ, this covers density-dependent mobilities and congestion constraints. If, in addition, A is positively 1-homogeneous in its first variable, $A(\eta a, w) = \eta A(a, w)$ for $\eta \geqslant 0$, then the same formula is intrinsic: replacing λ by another dominating measure gives the same value. The usual Benamou–Brenier action is the model case

$$A_2(a, w) = a\|w\|^2, \qquad \mathbb{A}(\alpha, w) = \int \|w\|^2 \mathrm{d}\alpha. \tag{14.18}$$

Here $a = \mathrm{d}\alpha/\mathrm{d}\lambda$ is the density of transported mass and w is the Eulerian velocity. The factor a says that moving twice as much mass with the same velocity costs twice as much.

Homogeneous momentum actions. The same action can be written in momentum variables, and this is the form in which convexity and metric properties are easiest to read. Set $\omega = \alpha w$, so that ω is a vector-valued measure. When the local description is written with the same reference λ, so that $\alpha = a\lambda$ and $\omega = m\lambda$, the pointwise momentum perspective is

$$J_A(a,m) := \begin{cases} A(a,m/a), & a > 0, \\ 0, & a = 0 \text{ and } m = 0, \\ +\infty, & a = 0 \text{ and } m \neq 0, \end{cases} \tag{14.19}$$

and the measure action relative to λ is

$$\mathbb{J}_{A,\lambda}(\alpha,\omega) := \int J_A\left(\frac{\mathrm{d}\alpha}{\mathrm{d}\lambda}, \frac{\mathrm{d}\omega}{\mathrm{d}\lambda}\right) \mathrm{d}\lambda, \tag{14.20}$$

with value $+\infty$ if α or the total variation $|\omega|$ is not absolutely continuous with respect to λ. This zero-density convention is the lower-semicontinuous one for the superlinear actions used below; other growths use the corresponding recession extension. If A is positively 1-homogeneous in a, then J_A is jointly 1-homogeneous: $J_A(\eta a, \eta m) = \eta J_A(a,m)$. In that intrinsic case the value of $\mathbb{J}_{A,\lambda}$ is independent of the dominating reference measure, and we write simply $\mathbb{J}_A$. For $A_2(a,w) = a\|w\|^2$, one recovers the quadratic perspective

$$J_2(a,m) = \frac{\|m\|^2}{a},$$

which is the integrand used in the convex Benamou–Brenier formulation (14.12). In this case $\omega = \alpha w$; if $\alpha = \rho \mathrm{d}x$ and $\omega = m\mathrm{d}x$, then $m = \rho w$, and $J_2(\rho, m) = \rho\|w\|^2$.

Proposition 14.9 (Concave mobilities give convex momentum actions)**.** *Let $I \subset [0,+\infty)$ be a convex interval, let $\theta : I \to [0,+\infty)$ be concave, and let $L : \mathbb{R}^d \to [0,+\infty]$ be convex with $L(0) = 0$. Define, on the set where $\theta(a) > 0$,*

$$J_{\theta,L}(a,m) := \theta(a) L\left(\frac{m}{\theta(a)}\right), \qquad A_{\theta,L}(a,w) := J_{\theta,L}(a,aw) = \theta(a) L\left(\frac{aw}{\theta(a)}\right), \tag{14.21}$$

and extend $J_{\theta,L}$ to the boundary by lower semicontinuity. Then $J_{\theta,L}$ is convex in (a,m). In particular, this single construction contains

$$A(a,w) = aL(w) \quad \text{by taking } \theta(a) = a,$$

and the concave-mobility quadratic action

$$A(a,w) = \frac{a^2\|w\|^2}{\theta(a)} \quad \text{by taking } L(u) = \|u\|^2.$$

Proof. The perspective $P(s,m) = sL(m/s)$ of a convex function is convex on $s > 0$. Since $L(0) = 0$, it is also nonincreasing in the scalar s for fixed m: if $s_2 \geqslant s_1 > 0$, then

$$L(m/s_2) = L\big((s_1/s_2)(m/s_1)\big) \leqslant (s_1/s_2)L(m/s_1),$$

hence $P(s_2,m) \leqslant P(s_1,m)$. Let $a_\zeta = (1-\zeta)a_0 + \zeta a_1$ and $m_\zeta = (1-\zeta)m_0 + \zeta m_1$. By concavity of θ,

$$\theta(a_\zeta) \geqslant (1-\zeta)\theta(a_0) + \zeta\theta(a_1).$$

Using the monotonicity in s, and then the convexity of P, gives

$$J_{\theta,L}(a_\zeta, m_\zeta) = P(\theta(a_\zeta), m_\zeta) \leqslant P((1-\zeta)\theta(a_0) + \zeta\theta(a_1), m_\zeta) \leqslant (1-\zeta)J_{\theta,L}(a_0,m_0) + \zeta J_{\theta,L}(a_1,m_1).$$

The boundary cases follow by the chosen lower-semicontinuous extension. □

> **Remark 14.10 (Perspective transform gives convexity).** The convexity mechanism in Proposition 14.9 is the same one used earlier for φ-divergences. In both cases one starts from a convex function of a ratio and multiplies by the denominator: the density-ratio perspective $v\varphi(u/v)$ proves the joint convexity and lower-semicontinuity of $\mathcal{D}_\varphi$ in Proposition 7.17, see (7.6) and (7.8); the momentum perspective $J_A(a,m) = A(a,m/a)$ in (14.19) plays the same role for dynamic transport actions. The classical

> Benamou–Brenier relaxation uses the quadratic perspective $J(a,m) = \|m\|^2/a$ from (14.10)–(14.11); concave mobilities use the modified perspective (14.21); and unbalanced dynamic transport uses the three-variable analogue $J_\kappa(a,m,r)$ in (14.39)–(14.40). Thus the same perspective transform is the hidden convexification behind density-ratio divergences, balanced dynamic OT, concave-mobility distances and Wasserstein–Fisher–Rao dynamics.

The next proposition isolates the additional assumptions under which the momentum formulation generated by A defines a genuine path metric rather than only a variational principle.

Proposition 14.11 (Homogeneous dynamic actions define distances)**.** *Fix a reference measure λ, omitted from the notation only in the intrinsic case where $\mathbb{J}_{A,\lambda}$ does not depend on λ. Assume that the momentum perspective J_A defined in (14.19) is lower semicontinuous, convex in (a,m), even in m, and satisfies $J_A(a,0) = 0$. Assume moreover that, for some $r > 1$,*

$$J_A(a,\xi m) = |\xi|^r J_A(a,m) \qquad \text{for all admissible } a,\ \text{all } m,\ \text{and all } \xi \in \mathbb{R},$$

and that $J_A(a,m) = 0$ if and only if $m = 0$. Equivalently, the evenness, homogeneity and nondegeneracy assumptions translate, away from zero density, into

$$A(a,-w) = A(a,w), \qquad A(a,\xi w) = |\xi|^r A(a,w), \qquad A(a,w) = 0 \Longleftrightarrow w = 0 \qquad (a > 0).$$

Define, on every fixed-mass class,

$$\mathsf{D}_{A,\lambda}(\alpha_0,\alpha_1) := \inf_{\substack{\partial_t \alpha_t + \operatorname{div} \omega_t = 0 \\ \alpha_{t=0}=\alpha_0,\ \alpha_{t=1}=\alpha_1}} \left(\int_0^1 \mathbb{J}_{A,\lambda}(\alpha_t,\omega_t)\,\mathrm{d}t \right)^{1/r}.$$

Assume finally that the relaxed dynamic problem is sequentially closed and attains its infimum whenever the value is finite. Then $\mathsf{D}_{A,\lambda}$ is an extended distance on each finite-action component: it is symmetric, satisfies the triangle inequality, and $\mathsf{D}_{A,\lambda}(\alpha_0,\alpha_1) = 0$ only when $\alpha_0 = \alpha_1$. In the intrinsic case, we write D_A.

Proof. The constant curve $\alpha_t = \alpha_0$, $\omega_t = 0$, gives $\mathsf{D}_{A,\lambda}(\alpha_0,\alpha_0) = 0$. Conversely, if $\mathsf{D}_{A,\lambda}(\alpha_0,\alpha_1) = 0$, attainment provides a relaxed zero-action minimizer, which satisfies

$$\int_0^1 \mathbb{J}_{A,\lambda}(\alpha_t,\omega_t)\,\mathrm{d}t = 0.$$

Since $\mathbb{J}_{A,\lambda} \geqslant 0$, this implies $\omega_t = 0$ for a.e. t. The continuity equation then gives $\partial_t \alpha_t = 0$ in the sense of distributions, so $\alpha_0 = \alpha_1$.

Symmetry follows by time reversal. If $(\alpha_t,\omega_t)_{t\in[0,1]}$ connects α_0 to α_1, then

$$\widetilde{\alpha}_t = \alpha_{1-t}, \qquad \widetilde{\omega}_t = -\omega_{1-t}$$

connects α_1 to α_0, satisfies the same continuity equation, and has the same action because J_A is even in m.

It remains to prove the triangle inequality. Let (α^1,ω^1) connect α_0 to α_1 with action E_1, and let (α^2,ω^2) connect α_1 to α_2 with action E_2. For $\tau \in (0,1)$, concatenate the two curves after the time changes $s = t/\tau$ and $s = (t-\tau)/(1-\tau)$:

$$(\alpha_t,\omega_t) = \begin{cases} \big(\alpha^1_{t/\tau},\ \tau^{-1}\omega^1_{t/\tau}\big), & 0 \leqslant t \leqslant \tau, \\ \big(\alpha^2_{(t-\tau)/(1-\tau)},\ (1-\tau)^{-1}\omega^2_{(t-\tau)/(1-\tau)}\big), & \tau \leqslant t \leqslant 1. \end{cases}$$

The rescaled curve is admissible from α_0 to α_2. By r-homogeneity in the momentum variable, its action is $\tau^{1-r}E_1 + (1-\tau)^{1-r}E_2$. Writing $X = E_1^{1/r}$ and $Y = E_2^{1/r}$, the optimal allocation is $\tau = X/(X+Y)$, with the evident limiting convention if $X + Y = 0$, and gives

$$\inf_{\tau\in(0,1)} \left\{ \tau^{1-r}E_1 + (1-\tau)^{1-r}E_2 \right\} = (X+Y)^r.$$

Hence $\mathsf{D}_{A,\lambda}(\alpha_0,\alpha_2) \leqslant E_1^{1/r} + E_2^{1/r}$. Taking E_1 and E_2 arbitrarily close to the two optimal action values yields

$$\mathsf{D}_{A,\lambda}(\alpha_0,\alpha_2) \leqslant \mathsf{D}_{A,\lambda}(\alpha_0,\alpha_1) + \mathsf{D}_{A,\lambda}(\alpha_1,\alpha_2),$$

with the usual extended-real convention when one of the distances is infinite. □

Without homogeneity or nondegeneracy, the same momentum action remains useful as a variational principle, but its r-th root need not define a distance.

Example 14.12 (Wasserstein-p action). The usual $\mathcal{W}_p$ distances correspond to changing only the homogeneity of the Benamou–Brenier action. The case $p = 2$ is the original formula [35]; for $1 < p < +\infty$, the same dynamic viewpoint gives the standard description of absolutely continuous curves in Wasserstein spaces [12, 538, 469]. The specific objects to insert in the general framework above are

$$A_p(a, w) = a\|w\|^p, \qquad J_p(a, m) = \begin{cases} \|m\|^p / a^{p-1}, & a > 0, \\ 0, & (a, m) = (0, 0), \\ +\infty, & a = 0,\ m \neq 0, \end{cases}$$

With $A = A_p$, Proposition 14.11 gives the usual identity $\mathsf{D}_{A_p} = \mathcal{W}_p$. The corresponding squared length-space normalization is

$$\mathbb{A}_p(\alpha, w) = \left(\int \|w\|^p \, \mathrm{d}\alpha \right)^{2/p}.$$

This is the squared version of the p-homogeneous action: minimizing $\int_0^1 \mathbb{A}_p(\alpha_t, v_t)\mathrm{d}t$ gives $\mathcal{W}_p^2$ after constant-speed reparametrization. Thus A_p, J_p denote the local p-homogeneous velocity and momentum densities, whereas $\mathbb{A}_p$ denotes the squared tangent action used in the length formulation. The endpoint $p = 1$ can be treated separately: the perspective becomes $J_1(a, m) = \|m\|$, and the dynamic problem collapses to Beckmann's formulation of $\mathcal{W}_1$ [31].

Concave-mobility actions. One can instead keep a quadratic momentum action and change the mobility. Dolbeault, Nazaret and Savaré introduced this construction as a class of generalized transport distances adapted to nonlinear diffusion [189]. The defining objects are collected below.

Definition 14.13 (Concave-mobility dynamic distance)**.** Let $I \subset [0, +\infty)$ be a convex interval and let $\theta : I \to [0, +\infty)$ be concave. Define

$$J_\theta(a, m) := \begin{cases} \|m\|^2 / \theta(a), & \theta(a) > 0, \\ 0, & \theta(a) = 0 \text{ and } m = 0, \\ +\infty, & \theta(a) = 0 \text{ and } m \neq 0. \end{cases}$$

The corresponding velocity action is

$$A_\theta(a, w) = J_\theta(a, aw) = \frac{a^2 \|w\|^2}{\theta(a)}.$$

For a fixed reference measure λ, set

$$\mathsf{W}_{\theta,\lambda}(\alpha_0, \alpha_1) := \mathsf{D}_{A_\theta,\lambda}(\alpha_0, \alpha_1).$$

The convexity of J_θ is the special case $L(u) = \|u\|^2$ of Proposition 14.9. This is why concavity of the mobility, rather than convexity, is the structural condition that makes the continuity-equation formulation convex.

Fix now a reference measure λ. If $\alpha = a\lambda$, the induced squared tangent action is

$$\mathbb{A}_{\theta,\lambda}(\alpha, w) = \int A_\theta(a(x), w(x)) \, \mathrm{d}\lambda(x) = \int \frac{a(x)^2}{\theta(a(x))} \|w(x)\|^2 \, \mathrm{d}\lambda(x),$$

and it is set to $+\infty$ when $\alpha \not\ll \lambda$. Equivalently, on the set where $\theta(a(x)) > 0$,

$$\mathbb{A}_{\theta,\lambda}(\alpha, w) = \int \frac{a(x)}{\theta(a(x))} \|w(x)\|^2 \, \mathrm{d}\alpha(x).$$

Hence this is a local Riemannian case, in the sense of (14.16), whenever the multiplier $a(x)/\theta(a(x))$ is finite and positive. The associated tensor is the multiplication operator

$$\big(Q_{\theta,\lambda,\alpha} w\big)(x) = \frac{a(x)}{\theta(a(x))} w(x), \qquad \alpha = a\lambda, \tag{14.22}$$

defined α-a.e. on the set where $\theta(a) > 0$. Except for linear mobilities $\theta(a) = ca$, and in particular the normalized case $\theta(a) = a$ which recovers $\mathcal{W}_2$, the pointwise velocity action $A_\theta(a, w)$ is not positively

1-homogeneous in the density variable a. Consequently the construction is not intrinsic under a change of λ: the resulting distance depends on the chosen reference measure and is finite only between endpoints that can be joined by a finite-action curve with $\alpha_t \ll \lambda$.

The subscript λ recalls that the action is measured through the density $a = \mathrm{d}\alpha/\mathrm{d}\lambda$. Equivalently, because this action is quadratic in the momentum, $\mathsf{W}^2_{\theta,\lambda}$ is the path value (14.15) with $\mathbb{A} = \mathbb{A}_{\theta,\lambda}$.

Proposition 14.14 (Concave-mobility dynamic distances)**.** *For a fixed reference measure λ, and under the standard compactness and lower-semicontinuity hypotheses ensuring existence of relaxed minimizers [189], $\mathsf{W}_{\theta,\lambda} = \mathsf{D}_{A_\theta,\lambda}$ is an extended distance on each finite-action component contained in $\{\alpha : \alpha \ll \lambda\}$. In particular, whenever the relevant component has finite pairwise distances, $\mathsf{W}_{\theta,\lambda}$ is a genuine metric there.*

Proof. Proposition 14.9, applied with $L(u) = \|u\|^2$, gives the lower-semicontinuous convex momentum density J_θ. It is even in m, satisfies $J_\theta(a, 0) = 0$, and obeys $J_\theta(a, \xi m) = |\xi|^2 J_\theta(a, m)$. Moreover $J_\theta(a, m) = 0$ if and only if $m = 0$, with the boundary convention used in its definition. For the fixed reference λ, the hypotheses of Proposition 14.11 therefore hold with $r = 2$; the existence assumption is supplied by the compactness hypothesis in the statement. That proposition gives symmetry, separation and the triangle inequality for $\mathsf{D}_{A_\theta,\lambda}$, hence for $\mathsf{W}_{\theta,\lambda}$. □

The choice $\theta(a) = a$ recovers $\mathcal{W}_2$. Other choices encode different geometry: $\theta(a) = a^\gamma$ with $0 < \gamma \leqslant 1$ changes the cost of moving dilute mass, while $\theta(a) = a(1 - a/M)$ on $[0, M]$ models a volume-filling or exclusion effect, since mobility vanishes both at vacuum and at the maximal density. The distance is comparable with $\mathcal{W}_2$ on classes where $\theta(a)$ is bounded above and below by positive multiples of a; otherwise zero-mobility barriers can make some pairs infinitely far apart. Its use as a geometry for nonlinear diffusion is discussed in Section 15.5.

Dynamic spectral Wasserstein distances. The static spectral distances of Section 11.5 penalize a coupling through the covariance of its displacement. A dynamic version keeps the continuity equation but replaces the pointwise kinetic energy by a gauge of the whole velocity covariance. The resulting action is nonlocal in space: velocity directions are charged globally through their covariance, rather than independently at each point.

The spectral geometry charges the covariance of the velocity field through a monotone gauge.

Definition 14.15 (Dynamic spectral Wasserstein distance)**.** Let γ be a monotone spectral gauge on $\mathbb{S}^d_+$. Define

$$\mathbb{A}_\gamma(\alpha, v) := \gamma\left(\int v(x)v(x)^\top \mathrm{d}\alpha(x)\right). \tag{14.23}$$

The associated distance is

$$\mathsf{W}^2_{\gamma,\mathrm{dyn}}(\alpha_0, \alpha_1) := \mathsf{D}^2_{\mathbb{A}_\gamma}(\alpha_0, \alpha_1). \tag{14.24}$$

The trace gauge gives the usual Wasserstein tangent action, while the operator gauge $\gamma(M) = \lambda_{\max}(M)$ charges only the largest directional velocity variance. In density–momentum variables, this corresponds to the measure action

$$\mathbb{J}_\gamma(\alpha, \omega) = \gamma\left(\int \left(\frac{\mathrm{d}\omega}{\mathrm{d}\alpha}\right)\left(\frac{\mathrm{d}\omega}{\mathrm{d}\alpha}\right)^\top \mathrm{d}\alpha\right),$$

or, when $\alpha = \rho\,\mathrm{d}x$ and $\omega = m\,\mathrm{d}x$,

$$\mathbb{J}_\gamma(\rho, m) = \gamma\left(\int \frac{m(x)m(x)^\top}{\rho(x)}\,\mathrm{d}x\right).$$

This functional is convex in the density–momentum fields (ρ, m) by the matrix perspective, together with the monotonicity and convexity of γ. It is nevertheless not, in general, obtained by integrating a pointwise action density, because the covariance is computed globally before applying γ. It becomes local only for linear spectral gauges. For instance, if $\gamma(M) = \mathrm{tr}(GM)$ with $G \succeq 0$, then the velocity and momentum densities are

$$A_{\mathrm{lin}}(a, w) = a\, w^\top G w, \qquad J_{\mathrm{lin}}(a, m) = \frac{m^\top G m}{a},$$

and the trace gauge, $G = \mathrm{Id}$ in $\gamma(M) = \mathrm{tr}(GM)$, recovers the usual Benamou–Brenier action.

The following result, in the form used for normalized spectral flows in [427], shows that this dynamic construction is not merely infinitesimal: it exactly recovers the static displacement-covariance formulation.

Proposition 14.16 (Static/dynamic equality for spectral Wasserstein)**.** *Let γ be a monotone spectral gauge and let $\alpha_0, \alpha_1 \in \mathcal{P}_2(\mathbb{R}^d)$. Then*

$$\mathsf{W}^2_{\gamma,\mathrm{dyn}}(\alpha_0, \alpha_1) = \mathcal{W}^2_\gamma(\alpha_0, \alpha_1).$$

The common value defines a distance on $\mathcal{P}_2(\mathbb{R}^d)$; the triangle inequality is established by the robust representation in Proposition 11.43.

Proof. We first prove $\mathsf{W}^2_{\gamma,\mathrm{dyn}} \leqslant \mathcal{W}^2_\gamma$. Let $\pi \in \mathcal{U}(\alpha_0, \alpha_1)$ and let $(X, Y) \sim \pi$. Set $Z_t = (1-t)X + tY$ and $\alpha_t = (Z_t)_\sharp \pi$. Define the Eulerian velocity as the conditional mean $v_t(z) := \mathbb{E}[Y - X \mid Z_t = z]$. Then (α_t, v_t) solves the continuity equation. If

$$M_\pi := \int (x-y)(x-y)^\top \mathrm{d}\pi(x,y), \qquad C_t := \int v_t(z) v_t(z)^\top \mathrm{d}\alpha_t(z),$$

conditional Jensen gives $C_t \preceq M_\pi$. Indeed, for every direction $u \in \mathbb{R}^d$,

$$u^\top C_t u = \mathbb{E}\big[\mathbb{E}[\langle u, Y - X\rangle \mid Z_t]^2\big] \leqslant \mathbb{E}[(\langle u, Y - X\rangle)^2] = u^\top M_\pi u.$$

By the Loewner monotonicity of γ, $\int_0^1 \gamma(C_t)\mathrm{d}t \leqslant \gamma(M_\pi)$. Taking the infimum over π gives the first inequality.

Conversely, let (α_t, v_t) be an admissible finite-action competitor. Since γ is a finite positive gauge on the finite-dimensional cone $\mathbb{S}^d_+$, it is equivalent to the trace on this cone; hence finite spectral action implies the usual finite kinetic energy needed for the superposition principle. Thus there exists a probability law η on absolutely continuous paths $\omega_\cdot$ such that $(e_t)_\sharp \eta = \alpha_t$ and $\dot\omega_t = v_t(\omega_t)$ for a.e. t. Let $\pi = (e_0, e_1)_\sharp \eta$ be the induced endpoint coupling. With $C_t = \int v_t v_t^\top \mathrm{d}\alpha_t$, Jensen's inequality along each path gives

$$(\omega_1 - \omega_0)(\omega_1 - \omega_0)^\top = \left(\int_0^1 \dot\omega_t \mathrm{d}t\right)\left(\int_0^1 \dot\omega_t \mathrm{d}t\right)^\top \preceq \int_0^1 \dot\omega_t \dot\omega_t^\top \mathrm{d}t.$$

After integration over η, $M_\pi \preceq \int_0^1 C_t \mathrm{d}t$. Monotonicity of γ, followed by convexity, yields

$$\gamma(M_\pi) \leqslant \gamma\left(\int_0^1 C_t \mathrm{d}t\right) \leqslant \int_0^1 \gamma(C_t)\mathrm{d}t.$$

Since π is admissible for the static problem, $\mathcal{W}^2_\gamma(\alpha_0, \alpha_1)$ is bounded by the action of the dynamic competitor. Taking the infimum over all competitors gives the reverse inequality.

The crucial assumption is the monotonicity of γ: both parts of the proof produce Loewner-order comparisons between covariance matrices, and these comparisons imply inequalities between actions only for monotone gauges. □

The use of this geometry for normalized flows, including the operator-gauge connection with Muon-type normalization, is developed in Section 15.5.

Kernelized Benamou–Brenier distances. A different way to deform the Benamou–Brenier geometry is to keep the local continuity equation but to measure velocities in a reproducing-kernel Hilbert space rather than in $L^2(\alpha)$. This construction is motivated by Stein variational gradient descent, studied later in Section 16.2: the kernel makes the velocity field smooth and computable from particles, at the price of defining a much more restrictive transport geometry.

The kernelized distance uses a vector-valued RKHS as its restricted velocity class.

Definition 14.17 (Kernelized Benamou–Brenier distance)**.** Let k be a positive definite kernel on $\mathbb{R}^d$ with scalar RKHS $\mathcal{H}_k$. Set

$$\mathcal{H}_k^d := \mathcal{H}_k \times \cdots \times \mathcal{H}_k, \qquad \|v\|^2_{\mathcal{H}_k^d} := \sum_{\ell=1}^d \|v_\ell\|^2_{\mathcal{H}_k}.$$

$$\mathbb{A}_k(\alpha, v) := \|v\|^2_{\mathcal{H}_k^d}, \qquad \mathcal{W}_k^2(\alpha_0, \alpha_1) := \mathsf{D}^2_{\mathbb{A}_k}(\alpha_0, \alpha_1). \tag{14.25}$$

This is the vector-valued analogue of the scalar RKHS norm used for MMDs in Section 7.2. Formula (14.15) is restricted to velocities $v_t \in \mathcal{H}_k^d$. The action is independent of α; the measure enters only through the continuity equation. This type of Stein geometry was introduced in the analysis of SVGD by Liu and Wang [346, 345] and later developed geometrically in [194, 405]. The important caveat is that the admissible tangent space is the smooth RKHS class, not the whole Wasserstein tangent space.

Proposition 14.18 (Kernelized dynamic distance)**.** *Assume that the kernel has uniformly bounded diagonal,*

$$\kappa_k^2 := \sup_{x \in \mathbb{R}^d} k(x, x) < +\infty.$$

Then the quantity $\mathcal{W}_k$ defined by (14.25) is an extended distance on probability measures. More precisely, it is symmetric, satisfies the triangle inequality, and $\mathcal{W}_k(\alpha_0, \alpha_1) = 0$ implies $\alpha_0 = \alpha_1$. It can take the value $+\infty$ between different finite-action components.

Proof. The constant curve gives zero self-distance. To prove separation without assuming existence of a minimizing curve, use the RKHS evaluation bound $\|v(x)\| \leqslant \kappa_k \|v\|_{\mathcal{H}_k^d}$. For every $\varphi \in C_c^1(\mathbb{R}^d)$ and every admissible curve of action E, the weak continuity equation and Cauchy–Schwarz give

$$\begin{aligned}\left| \int \varphi \, \mathrm{d}(\alpha_1 - \alpha_0) \right| &\leqslant \int_0^1 \int \|\nabla \varphi(x)\| \, \|v_t(x)\| \, \mathrm{d}\alpha_t(x) \mathrm{d}t \\ &\leqslant \kappa_k \|\nabla \varphi\|_\infty \left(\int_0^1 \|v_t\|^2_{\mathcal{H}_k^d} \mathrm{d}t \right)^{1/2} = \kappa_k \|\nabla \varphi\|_\infty \sqrt{E}.\end{aligned}$$

Taking the infimum over admissible curves shows that $\mathcal{W}_k(\alpha_0, \alpha_1) = 0$ forces equality of the integrals of every compactly supported smooth test function, hence $\alpha_0 = \alpha_1$.

Symmetry follows by time reversal and by replacing v_t with $-v_{1-t}$. For the triangle inequality, concatenate an almost optimal curve from α_0 to α_1 of action E_1 with one from α_1 to α_2 of action E_2. Compressing the first curve into a time interval of length τ multiplies its action by $1/\tau$, and compressing the second into length $1-\tau$ multiplies its action by $1/(1-\tau)$. Optimizing over $\tau \in (0,1)$ gives $(\sqrt{E_1} + \sqrt{E_2})^2$, and taking infima yields the triangle inequality. □

One should read $\mathcal{W}_k$ as an extended distance on finite-action components, not as a replacement for $\mathcal{W}_2$ on all of $\mathcal{P}_2(\mathbb{R}^d)$. A useful sufficient condition for finiteness is that the endpoints lie on the same RKHS-flow orbit: if there exists $v \in L^2([0,1]; \mathcal{H}_k^d)$ whose flow map Φ_t solves $\dot{\Phi}_t = v_t \circ \Phi_t$ and satisfies $\alpha_1 = (\Phi_1)_\sharp \alpha_0$, then $\mathcal{W}_k^2(\alpha_0, \alpha_1) \leqslant \int_0^1 \|v_t\|^2_{\mathcal{H}_k^d} \mathrm{d}t$. In particular, for strictly positive definite kernels, two discrete measures with the same weights and distinct moving support points are at finite distance whenever their atoms can be connected by noncolliding smooth paths, because RKHS interpolation constructs vector fields realizing the prescribed atom velocities along the paths.

The same condition also explains the limitation. If k is smooth enough that $\mathcal{H}_k^d$ embeds into Lipschitz vector fields, finite-action curves are induced by flows of regular maps. Atomic measures therefore remain atomic with the same number of atoms along such curves, so a Dirac mass cannot be transported at finite kernelized action to a measure with a density. This lack of splitting is precisely what makes the geometry useful for deterministic particle methods, and also what makes it a nontrivial extended object rather than a full probability metric.

14.4 Nonlocal Wasserstein Distances

The local dynamic distances above transport mass through a vector field on the base space and a classical continuity equation. Nonlocal geometries use a different tangent model: the elementary motion is an exchange across an edge or a jump from x to y. The common data are a reversible kernel K, a symmetric edge or jump measure J, a pairwise increment $\bar{\nabla}$, and a pair-space action $\mathbb{A}_K$. The tangent variable is therefore attached to pairs of points, not to a single point x, so these constructions are not simply obtained by choosing another pointwise local cost $A(a, w)$.

There are two complementary versions. On a finite state space, the goal is to put a Wasserstein-like geometry on the probability simplex so that the entropy gradient flow is exactly a prescribed reversible

Markov chain [361, 384, 145]. On a continuum space, the same edge calculus becomes a jump calculus over $\mathcal{X} \times \mathcal{X}$, which models nonlocal motion, heavy-tailed jumps, and fractional-type diffusion; this is the construction of Erbar [205], building on nonlinear mobilities [189], with subsequent metric and asymptotic refinements [493].

In both settings the canonical mobility for entropy is the logarithmic mean.

Definition 14.19 (Logarithmic mean)**.**

$$\theta(a,b) := \begin{cases} \dfrac{a-b}{\log a - \log b}, & a,b>0 \text{ and } a \neq b, \\ a, & a = b, \\ 0, & ab = 0, \end{cases} \tag{14.26}$$

for $a, b \geqslant 0$.

It appears because $\theta(a,b)(\log a - \log b) = a - b$, which is the edge-wise chain rule identifying entropy-driven flows with the underlying reversible Markov or jump dynamics.

Continuum jump kernels. The continuum version replaces graph edges by a symmetric measure on pairs. Its action is still quadratic, but the tangent variable is an antisymmetric jump velocity $v(x,y)$, and the mobility depends simultaneously on the two endpoint density values. It is best viewed as a convex action on the pair space $\mathcal{X} \times \mathcal{X}$, rather than as an integral of independent costs attached to single base points x.

Let $(\mathcal{X}, \mathfrak{m})$ be a reference measure space, and let $K(x,\cdot)$ be a nonnegative measure on $\mathcal{X}$ for each $x \in \mathcal{X}$, possibly of infinite total mass. We write this kernel as $K(x, \mathrm{d}y)$ to emphasize that the integration variable is y. The pair measure J on $\mathcal{X} \times \mathcal{X}$ is defined by testing against nonnegative measurable functions Φ:

$$\int_{\mathcal{X}\times\mathcal{X}} \Phi(x,y)\, \mathsf{J}(\mathrm{d}x, \mathrm{d}y) := \int_{\mathcal{X}} \left(\int_{\mathcal{X}} \Phi(x,y)\, K(x, \mathrm{d}y) \right) \mathfrak{m}(\mathrm{d}x). \tag{14.27}$$

The reversibility assumption is precisely that this measure J is symmetric, i.e. invariant under $(x,y) \mapsto (y,x)$. For a density $\rho = \mathrm{d}\alpha/\mathrm{d}\mathfrak{m}$, write

$$\bar{\nabla}\varphi(x,y) := \varphi(y) - \varphi(x)$$

for the nonlocal gradient, and use the logarithmic mean θ defined in (14.26). These objects define the nonlocal continuity equation and its least-action distance.

Definition 14.20 (Continuum nonlocal Wasserstein distance)**.** For $\alpha_t = \rho_t \mathfrak{m}$ and antisymmetric $v_t(x,y)$, require, for every test function φ,

$$\frac{\mathrm{d}}{\mathrm{d}t} \int \varphi \, \mathrm{d}\alpha_t = \frac{1}{2} \iint \bar{\nabla}\varphi(x,y)\, v_t(x,y)\, \theta(\rho_t(x), \rho_t(y))\, \mathsf{J}(\mathrm{d}x, \mathrm{d}y). \tag{14.28}$$

Define

$$\mathbb{A}_K(\alpha, v) := \frac{1}{2} \iint |v(x,y)|^2 \theta(\rho(x), \rho(y))\, \mathsf{J}(\mathrm{d}x, \mathrm{d}y),$$

and

$$\mathcal{W}_K^2(\alpha_0, \alpha_1) := \inf_{\rho_t, v_t} \int_0^1 \mathbb{A}_K(\alpha_t, v_t)\, \mathrm{d}t, \tag{14.29}$$

where the infimum is over all curves solving (14.28) with endpoints α_0, α_1.

Here v is not a vector field on $\mathcal{X}$ but an antisymmetric velocity on pairs (x,y); the action is therefore genuinely nonlocal.

Proposition 14.21 (Metric properties of nonlocal transport)**.** *Under the regularity and irreducibility assumptions of [205], $\mathcal{W}_K$ is an extended distance on the set of probability measures absolutely continuous with respect to $\mathfrak{m}$, and finite-distance pairs are connected by constant-speed geodesics.*

Proof. We use the analytic compactness and lower-semicontinuity theorem of [205] for the logarithmic-mean action. In particular, action-bounded sequences of admissible curves are compact for the narrow

topology, the weak nonlocal continuity equation is closed under this convergence, and the action is lower semicontinuous. These facts are the nonlocal analogues of the compactness theorem used in the Benamou–Brenier formulation.

Nonnegativity is immediate from the definition of $\mathbb{A}_K(\alpha, v)$. If $\alpha_0 = \alpha_1$, the constant curve $\rho_t = \rho_0$, $v_t = 0$, is admissible and has zero action, hence $\mathcal{W}_K(\alpha_0, \alpha_0) = 0$.

Symmetry follows by time reversal. If (ρ_t, v_t) transports α_0 to α_1, set $\tilde{\rho}_t = \rho_{1-t}$, $\tilde{v}_t = -v_{1-t}$. For every test function φ, differentiating $\int \varphi \tilde{\rho}_t \, \mathrm{d}\mathfrak{m}$ and using (14.28) shows that $(\tilde{\rho}_t, \tilde{v}_t)$ solves the same continuity equation from α_1 to α_0. Since the action is quadratic in v, $\mathbb{A}_K(\tilde{\alpha}_t, \tilde{v}_t) = \mathbb{A}_K(\alpha_{1-t}, v_{1-t})$, so the action is unchanged.

For the triangle inequality, let (ρ^0_t, v^0_t) connect α_0 to α_1 with action E_0, and let (ρ^1_t, v^1_t) connect α_1 to α_2 with action E_1. For $0 < \zeta < 1$, concatenate the two curves after the rescaling

$$(\rho_t, v_t) = \begin{cases} \big(\rho^0_{t/\zeta},\ \zeta^{-1} v^0_{t/\zeta}\big), & 0 \leqslant t \leqslant \zeta, \\ \big(\rho^1_{(t-\zeta)/(1-\zeta)},\ (1-\zeta)^{-1} v^1_{(t-\zeta)/(1-\zeta)}\big), & \zeta < t \leqslant 1. \end{cases}$$

The factors ζ^{-1} and $(1-\zeta)^{-1}$ are exactly those needed for the weak continuity equation after changing time. Since $v \mapsto \mathbb{A}_K(\alpha, v)$ is quadratic,

$$\int_0^1 \mathbb{A}_K(\alpha_t, v_t)\, \mathrm{d}t = \frac{E_0}{\zeta} + \frac{E_1}{1-\zeta}.$$

Optimizing in ζ, or taking $\zeta = \sqrt{E_0}/(\sqrt{E_0} + \sqrt{E_1})$ when both actions are positive, gives a concatenated action $(\sqrt{E_0} + \sqrt{E_1})^2$. Taking infima over the two curves yields

$$\mathcal{W}_K(\alpha_0, \alpha_2) \leqslant \mathcal{W}_K(\alpha_0, \alpha_1) + \mathcal{W}_K(\alpha_1, \alpha_2).$$

It remains to justify definiteness. If $\mathcal{W}_K(\alpha_0, \alpha_1) = 0$, choose admissible curves with actions tending to zero. By the compactness and lower-semicontinuity theorem quoted above, a subsequence converges to an admissible curve (ρ_t, v_t) of zero action. Hence $v_t = 0$ for $\theta(\rho_t(x), \rho_t(y))\mathsf{J}(\mathrm{d}x, \mathrm{d}y)\mathrm{d}t$-a.e. (t, x, y). Substituting in the weak continuity equation gives $\frac{\mathrm{d}}{\mathrm{d}t} \int \varphi \, \mathrm{d}\alpha_t = 0$ for every admissible test function φ. The irreducibility/separation assumption in [205] ensures that these test functions determine the measure, so α_t is constant and $\alpha_0 = \alpha_1$.

Finally, if $\mathcal{W}_K(\alpha_0, \alpha_1) < +\infty$, the same direct-method compactness applied to a minimizing sequence gives a minimizer of (14.29). Reparametrizing this minimizing curve by metric arclength gives a constant-speed curve. Equivalently, for every $0 \leqslant s < t \leqslant 1$, the restriction of the minimizer to $[s, t]$, rescaled to unit time, is admissible between α_s and α_t; the triangle inequality and equality of the total minimal action force

$$\mathcal{W}_K(\alpha_s, \alpha_t) = (t-s)\mathcal{W}_K(\alpha_0, \alpha_1)$$

after this constant-speed parametrization. Thus finite-distance pairs are joined by constant-speed geodesics. The consequences for entropy dynamics and fractional PDE examples are developed in Section 15.6. □

For a fixed jump kernel, this geometry is genuinely nonlocal and does not coincide with ordinary $\mathcal{W}_2$. The local metric is nevertheless recovered in a small-jump limit. More explicitly, on $\mathcal{X} = \mathbb{R}^d$, let $\eta(z) = \bar{\eta}(\|z\|)$ be a nonnegative radial profile with

$$M_2(\eta) := \int_{\mathbb{R}^d} \|z\|^2 \eta(z)\, \mathrm{d}z \in (0, +\infty),$$

and define

$$K_\varepsilon(x, \mathrm{d}y) := \eta_\varepsilon(y - x)\, \mathrm{d}y, \qquad \eta_\varepsilon(z) := \varepsilon^{-d} \eta(z/\varepsilon). \tag{14.30}$$

Radiality implies symmetry and isotropy, and a change of variables gives

$$\int \|y - x\|^2 K_\varepsilon(x, \mathrm{d}y) = \varepsilon^2 M_2(\eta), \qquad \int (y-x)(y-x)^\top K_\varepsilon(x, \mathrm{d}y) = \varepsilon^2 \frac{M_2(\eta)}{d} \mathrm{Id}. \tag{14.31}$$

Equation (14.31) is the precise sense in which the second-moment jump scale is ε. To obtain a nontrivial local limit, one simultaneously accelerates the jump rate by ε^{-2} and sets

$$\widehat{K}_\varepsilon := \frac{2d}{\varepsilon^2 M_2(\eta)} K_\varepsilon, \qquad \int (y-x)(y-x)^\top \widehat{K}_\varepsilon(x, \mathrm{d}y) = 2\mathrm{Id}.$$

Multiplying a jump kernel by a constant $c > 0$ divides the associated distance by $\sqrt{c}$. Hence, under the regularity and irreducibility hypotheses of [493], for endpoints supported in a fixed compact set,

$$\mathcal{W}_{\widehat{K}_\varepsilon} = \varepsilon \sqrt{\frac{M_2(\eta)}{2d}}\, \mathcal{W}_{K_\varepsilon} \underset{\varepsilon \to 0}{\longrightarrow} \mathcal{W}_2 .$$

This makes precise how sharply concentrated isotropic jumps recover the local Benamou–Brenier geometry. Without isotropy, the covariance matrix in (14.31) need not be proportional to Id, and the limit is instead an anisotropic Wasserstein geometry.

Discrete Wasserstein distances on Markov chains. The finite-state version keeps the same pair-space philosophy, but with a finite graph of admissible exchanges. It is not the naive Euclidean metric on the simplex. The key idea, introduced by Maas and independently developed in related forms by Mielke and by Chow–Huang–Li–Zhou, is to use the transition graph of a reversible Markov chain to define both the admissible directions and the mobility of the mass [361, 384, 145]. The resulting distance is the finite-dimensional counterpart of the Wasserstein geometry behind the heat equation, and it is also the geometric structure used in numerical work on discrete metric-measure spaces [206, 207]. Its entropy interpretation is stated later in Proposition 15.61.

Let $\mathcal{X} = \{1, \dots, n\}$ and let $K = (K_{i,j})$ denote the off-diagonal transition rates of an irreducible continuous-time Markov chain which is reversible with respect to a probability vector π, so that $\pi_i K_{i,j} = \pi_j K_{j,i}$ for $i \neq j$. We write

$$\mathsf{J}_{i,j} := \pi_i K_{i,j} = \pi_j K_{j,i}$$

for the symmetric edge measure, the finite counterpart of the jump measure used above. The transported object is a mass histogram

$$\Sigma_n := \left\{ a \in \mathbb{R}_+^n : \sum_i a_i = 1 \right\}.$$

Relative densities only enter as auxiliary variables with respect to the invariant law:

$$\rho_i(a) := \frac{a_i}{\pi_i}, \qquad a_i = \pi_i \rho_i(a).$$

The logarithmic mean θ defined in (14.26) is the mobility selected so that the entropy calculus later recovers exactly the Markov evolution; see Proposition 15.61. The identity $\theta(a,b)(\log a - \log b) = a - b$ converts entropy gradients into density differences along graph edges. For a potential $\psi \in \mathbb{R}^n$, set

$$\bar{\nabla}\psi(i,j) := \psi_j - \psi_i.$$

The finite nonlocal divergence, tangent action, and distance are the discrete counterparts of Definition 14.20.

Definition 14.22 (Discrete Markov-chain Wasserstein distance)**.**

$$(\mathcal{K}_\rho \psi)_i := \sum_j K_{i,j} \theta(\rho_i, \rho_j)(\psi_i - \psi_j), \qquad (\mathcal{L}_a \psi)_i := \pi_i (\mathcal{K}_{\rho(a)} \psi)_i. \tag{14.32}$$

Its associated tangent action is

$$\mathbb{A}_K(a, \psi) := \frac{1}{2} \sum_{i,j} \mathsf{J}_{i,j} \theta(\rho_i(a), \rho_j(a)) (\bar{\nabla}\psi(i,j))^2. \tag{14.33}$$

$$\mathcal{W}_K^2(a_0, a_1) := \inf_{a_t, \psi_t} \int_0^1 \mathbb{A}_K(a_t, \psi_t)\, \mathrm{d}t, \qquad \dot{a}_t + \mathcal{L}_{a_t} \psi_t = 0, \tag{14.34}$$

with endpoint conditions $a_{t=0} = a_0$ and $a_{t=1} = a_1$.

The tangent variable is the potential ψ, or equivalently the induced edge flux, rather than an ambient Euclidean vector field. In edge-flux variables, flux is allowed only where $K_{i,j} > 0$, and its kinetic denominator is the logarithmic mean of the two relative endpoint densities.

The first nontrivial finite Markov geometries already show how the logarithmic mean bends the simplex. In both examples below, take the uniform random walk on the complete neighbor graph, i.e. every pair of distinct points is declared neighboring. Thus $\pi_i = 1/n$, and $K_{i,j} = 1/(n-1)$ for $i \neq j$.

Example 14.23 (Two-point complete graph). On Σ_2, write $a_0 = (r, 1-r)$ and $a_1 = (s, 1-s)$. Since there is only one edge, the discrete dynamic problem reduces to the scalar Riemannian length

$$\mathcal{W}_K(a_0, a_1) = \left| \int_s^r \frac{\mathrm{d}u}{\sqrt{\theta(u, 1-u)}} \right|, \qquad 0 < r, s < 1. \tag{14.35}$$

This formula is closed but not Euclidean: the logarithmic mean changes the cost of moving mass depending on the current split between the two points.

Example 14.24 (Three-point complete graph). On Σ_3, the complete neighbor graph is a triangle. For $a \in \operatorname{int}(\Sigma_3)$, set

$$\Theta_{i,j}(a) := \frac{1}{2}\theta(a_i, a_j), \qquad 1 \leqslant i < j \leqslant 3.$$

For a fixed a, write $\Theta_{i,j} = \Theta_{i,j}(a)$. For a tangent vector $h \in \mathbb{R}^3$ with $h_1 + h_2 + h_3 = 0$, orient the edges as $1 \to 2$, $1 \to 3$, $2 \to 3$. The squared norm induced by the discrete Wasserstein metric is

$$\|h\|_a^2 = \min_{m_{1,2}, m_{1,3}, m_{2,3}} \left\{ \frac{m_{1,2}^2}{\Theta_{1,2}} + \frac{m_{1,3}^2}{\Theta_{1,3}} + \frac{m_{2,3}^2}{\Theta_{2,3}} \right\}, \tag{14.36}$$

subject to

$$h_1 + m_{1,2} + m_{1,3} = 0, \qquad h_2 - m_{1,2} + m_{2,3} = 0, \qquad h_3 - m_{1,3} - m_{2,3} = 0.$$

Eliminating the three edge fluxes gives an explicit formula. With $D = \Theta_{1,2}^{-1} + \Theta_{1,3}^{-1} + \Theta_{2,3}^{-1}$,

$$m_{1,2}^* = \frac{h_2/\Theta_{2,3} - h_1/\Theta_{1,3}}{D}, \qquad m_{1,3}^* = -h_1 - m_{1,2}^*, \qquad m_{2,3}^* = m_{1,2}^* - h_2,$$

and $\|h\|_a^2$ is obtained by inserting these values in (14.36). Therefore

$$\mathcal{W}_K^2(a_0, a_1) = \inf_{a_t \in \operatorname{int}(\Sigma_3)} \int_0^1 \|\dot{a}_t\|_{a_t}^2 \, \mathrm{d}t, \qquad a_{t=0} = a_0, \quad a_{t=1} = a_1. \tag{14.37}$$

Thus the three-state distance is an explicit two-dimensional Riemannian geodesic problem on the open triangle. The formula is simple enough to compute directly, but it already shows the main difference with Euclidean geometry on the simplex: the local metric depends nonlinearly on the current density through logarithmic edge mobilities.

Figure 14.3 visualizes these small-dimensional geometries and compares them with the ordinary Wasserstein distance associated with the 0/1 ground metric, for which $\mathcal{W}_2^2$ is exactly the total variation distance.

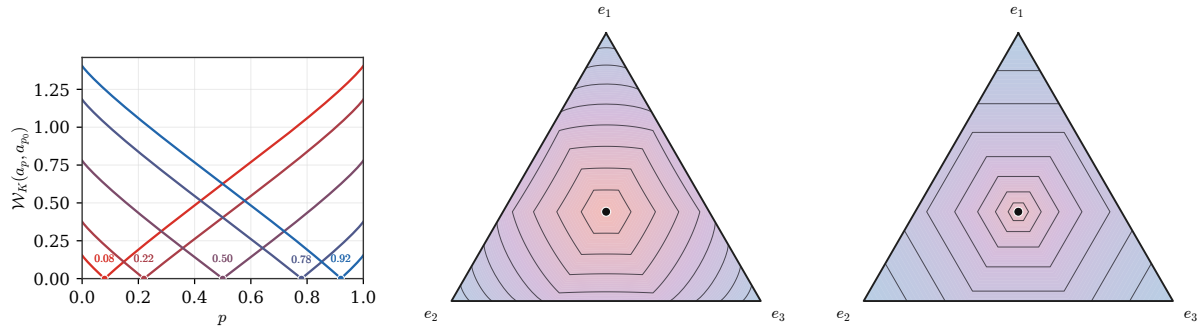


Figure 14.3: Discrete Wasserstein distances on small Markov-chain simplices. Left: the closed-form profiles $r \mapsto \mathcal{W}_K(a_r, a_{r_0})$, with $a_r = (r, 1-r)$, for several anchors r_0 on Σ_2. Middle: numerical level sets of $\mathcal{W}_K(a, \bar{a})$ on Σ_3, where $\bar{a} = (1/3, 1/3, 1/3)$, computed from the local Riemannian norm induced by the complete-neighbor Markov chain. Right: the corresponding level sets for the ordinary $\mathcal{W}_2$ distance with $d(i,j) = 1$ for $i \neq j$, i.e. $\mathcal{W}_2^2(a, \bar{a}) = \|a - \bar{a}\|_{\mathrm{TV}}$.

14.5 Dynamic Unbalanced Wasserstein Distances

Balanced dynamic distances keep the total mass fixed: their tangent vectors are transport velocities or fluxes satisfying a continuity equation. Unbalanced distances use a different tangent model. A tangent vector now has both a spatial component and a reaction component, so mass can move, disappear, and reappear. This section isolates this reaction–transport geometry before its use for gradient flows in Section 15.7.

Balance equation and tangent variables. Unbalanced dynamic transport is obtained by allowing mass to be created and destroyed along the path. The continuity equation is replaced by a balance equation, and, at the density level, an admissible tangent direction is a pair (m, s): the flux density m transports mass in space, while the source density s changes the amount of mass locally. This dynamic formulation underlies the Hellinger–Kantorovich and Wasserstein–Fisher–Rao metrics [339, 143]; its equivalence with static entropy-transport and cone formulations is developed in [340, 141]. A representative quadratic balance equation is

$$\partial_t \rho_t + \nabla \cdot m_t = s_t, \qquad \int_0^1 \int \left(\frac{\|m_t\|^2}{\rho_t} + \kappa^2 \frac{s_t^2}{\rho_t} \right) \mathrm{d}x \, \mathrm{d}t,$$

with the usual perspective convention: zero flux and zero source through zero density cost nothing, whereas nonzero flux or source through zero density has infinite cost. At the measure level these densities become the flux measure $\omega_t = m_t \, \mathrm{d}x$ and the signed source measure $\sigma_t = s_t \, \mathrm{d}x$.

Reaction–transport action. The action attaches one price to displacement and another to local growth. Its velocity and momentum forms are collected in one definition.

Definition 14.25 (Dynamic Wasserstein–Fisher–Rao distance)**.**

For $\kappa > 0$, define

$$A_\kappa(a, w, g) := a\left(\|w\|^2 + \kappa^2 g^2\right), \tag{14.38}$$

and

$$J_\kappa(a, m, r) := \begin{cases} (\|m\|^2 + \kappa^2 r^2)/a, & a > 0, \\ 0, & (a, m, r) = (0, 0, 0), \\ +\infty, & a = 0,\ (m, r) \neq (0, 0). \end{cases} \tag{14.39}$$

If λ dominates α, $|\omega|$, and $|\sigma|$, set

$$\mathbb{J}_\kappa(\alpha, \omega, \sigma) := \int J_\kappa\left(\frac{\mathrm{d}\alpha}{\mathrm{d}\lambda}, \frac{\mathrm{d}\omega}{\mathrm{d}\lambda}, \frac{\mathrm{d}\sigma}{\mathrm{d}\lambda} \right) \mathrm{d}\lambda. \tag{14.40}$$

The induced dynamic Wasserstein–Fisher–Rao distance is

$$\mathrm{WFR}_\kappa^2(\alpha_0,\alpha_1) := \inf_{\substack{\partial_t \alpha_t + \nabla\cdot\omega_t = \sigma_t \\ \alpha_{t=0}=\alpha_0,\ \alpha_{t=1}=\alpha_1}} \int_0^1 \mathbb{J}_\kappa(\alpha_t,\omega_t,\sigma_t)\mathrm{d}t. \tag{14.41}$$

Here w is velocity, g is relative growth, $m = aw$ is flux, and $r = ag$ is source density. The parameter κ fixes the relative cost of reaction and transport. Since J_κ is convex and positively one-homogeneous, the measure action is independent of λ; finite action forces $\omega, \sigma \ll \alpha$.

Static and dynamic viewpoints. The dynamic balance-equation formula is not a separate object from static unbalanced OT: it is the least-action representation of the same cone distance. To make the normalization explicit, define on the cone $\mathfrak{C}[\mathbb{R}^d]$ the squared cost

$$\Delta_\kappa\big((x,r),(y,s)\big)^2 := 4\kappa^2\left[r^2+s^2-2rs\cos\left(\frac{\|x-y\|}{2\kappa}\wedge\frac{\pi}{2}\right)\right]. \tag{14.42}$$

The radii encode masses through the weighted projection P_2 defined in Section 11.1. Accordingly, set

$$\mathrm{CW}_\kappa(\alpha_0,\alpha_1) := \inf_{\substack{\gamma\in\mathcal{M}_+(\mathfrak{C}[\mathbb{R}^d]^2) \\ \mathsf{P}_2\gamma_1=\alpha_0,\ \mathsf{P}_2\gamma_2=\alpha_1}} \int \Delta_\kappa\big((x,r),(y,s)\big)^2 \mathrm{d}\gamma(x,r,y,s). \tag{14.43}$$

For $\kappa = 1/2$, this is exactly the normalization of the Hellinger–Kantorovich cone cost used in Section 11.1. The next proposition identifies its general rescaling with the dynamic action above.

Thus Definition 14.25 gives both the reaction–transport action and the induced dynamic WFR metric.

Proposition 14.26 (Static/dynamic equivalence for unbalanced OT)**.** *For nonnegative finite measures α_0, α_1 on $\mathbb{R}^d$, the dynamic value (14.41) equals the static cone value (14.43). Hence $\mathrm{WFR}_\kappa = \mathrm{CW}_\kappa^{1/2}$ is the geodesic distance generated by the balance-equation least-action problem.*

Proof. The cone construction turns variation of mass into radial motion and spatial transport into angular motion on $\mathfrak{C}[\mathbb{R}^d]$. The normalization can be checked on a smooth cone path. If its base position is x_t, its cone radius is r_t, and the projected mass is $a_t = r_t^2$, then its relative growth rate is $g_t = \dot a_t/a_t = 2\dot r_t/r_t$. The infinitesimal energy induced by (14.42) is therefore

$$4\kappa^2\dot r_t^2 + r_t^2\|\dot x_t\|^2 = a_t\left(\|\dot x_t\|^2 + \kappa^2 g_t^2\right),$$

which is precisely the velocity action (14.38).

Applying the Benamou–Brenier theorem on the cone to the lifted endpoint measures now gives a dynamic least-action problem whose static value is CW_κ. Projecting a cone curve back to the base space with the squared cone radius as weight produces a measure curve α_t, a spatial flux measure ω_t, and a source measure σ_t satisfying the balance equation. The cone kinetic energy decomposes exactly into the three-variable perspective action $\mathbb{J}_\kappa$ in (14.41). Conversely, any finite-action triple $(\alpha_t,\omega_t,\sigma_t)$ can be lifted, after relaxation, to a cone curve whose radial and spatial velocities realize the source and flux with the same action. This is the standard static/dynamic identification for the Hellinger–Kantorovich and Wasserstein–Fisher–Rao metrics [339, 340, 143, 141].

The two infima are therefore equal; lower semicontinuity gives the general finite-measure statement from the smooth positive case. □

Balanced versus unbalanced interpolations. The distinction is visible already for one-dimensional mixtures with mismatched modal masses. Balanced transport must physically move the excess mass, whereas unbalanced transport can trade transport against reaction. Figure 14.4 uses entropic balanced and KL-relaxed barycenters as a qualitative numerical surrogate; its unbalanced row illustrates the reaction–transport mechanism but is not asserted to be an exact WFR_κ geodesic.

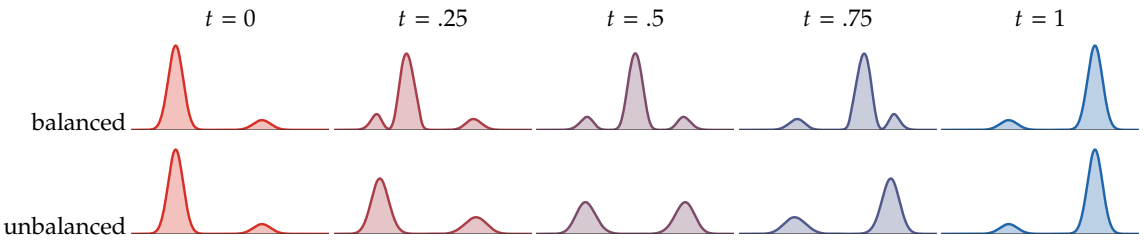


Figure 14.4: Balanced and unbalanced Sinkhorn-barycenter interpolations between two one-dimensional Gaussian mixtures with swapped modal masses. The balanced row conserves total mass, so excess mass from the dominant left mode must move along the line toward the dominant right target mode, producing transient mass in the middle. The unbalanced row uses KL-relaxed marginal constraints; mass can be attenuated near overrepresented modes and recreated near underrepresented modes, giving a reaction–transport interpolation closer to the Wasserstein–Fisher–Rao intuition.

14.6 Variational Mean Field Games

Mean field games use a population law to summarize strategic interactions among a very large number of individually negligible agents. This section only considers the special *potential* subclass in which the equilibrium system is the optimality condition of one convex planning problem. In that setting, the relation with dynamic OT is particularly transparent: one augments the Benamou–Brenier kinetic action by congestion and may replace a prescribed final measure by a terminal penalty.

From individual control to a population equilibrium. Mean field games were introduced by Lasry and Lions [323] and, in a parallel large-population stochastic-control framework, by Huang, Malhamé and Caines [281]. Given a candidate population density $(\rho_t)_{t\in[0,1]}$, a representative deterministic agent starting from x chooses a path γ by minimizing

$$\inf_{\gamma(0)=x} \left\{ \int_0^1 \left(\frac{1}{2}\|\dot\gamma(t)\|^2 + g\big(\rho_t(\gamma(t))\big) \right) \mathrm{d}t + \Psi(\gamma(1)) \right\}. \tag{14.44}$$

Here the kinetic term prices individual effort, $g(\rho)$ is the running cost created by the local population density, and Ψ assigns a cost to the final state. A mean field Nash equilibrium requires consistency: when all agents use an optimal feedback for (14.44), the resulting law of their positions must be precisely $\alpha_t = \rho_t\,\mathrm{d}x$.

A Benamou–Brenier planning problem. The consistency condition is usually coupled to an individual Hamilton–Jacobi–Bellman equation. A global variational reduction is available when the local coupling is a first variation: assume

$$g(r) = G'(r) \tag{14.45}$$

for a proper convex congestion potential $G : [0,+\infty) \to \mathbb{R}\cup\{+\infty\}$. Starting from $\alpha_0 = \rho_0\,\mathrm{d}x$, the associated potential MFG is the following planning problem.

Definition 14.27 (Variational mean-field-game planning problem)**.**

$$\inf_{(\rho_t,v_t)} \left\{ \int_0^1 \int_{\mathbb{R}^d} \left(\frac{1}{2}\rho_t(x)\|v_t(x)\|^2 + G(\rho_t(x)) \right) \mathrm{d}x\,\mathrm{d}t + \int_{\mathbb{R}^d} \Psi(x)\rho_1(x)\mathrm{d}x \right\}, \tag{14.46}$$

subject to

$$\partial_t\rho_t + \operatorname{div}(\rho_t v_t) = 0, \qquad \rho_{t=0}\,\mathrm{d}x = \alpha_0. \tag{14.47}$$

Unlike in (14.8), the final measure is not prescribed. The terminal potential Ψ selects desirable final states softly, while the integral of $G(\rho_t)$ penalizes crowded intermediate configurations. It is important that the planner pays $G(\rho)$ rather than $\rho g(\rho)$: the marginal cost perceived by one infinitesimal agent is the first variation $G'(\rho) = g(\rho)$. For the usual congestion models, g is nondecreasing, equivalently G is convex; superlinear choices of G increasingly penalize concentration.

This potential formulation and its deterministic and diffusive variants are developed by Benamou, Carlier and Santambrogio [40]; first-order MFG systems with local couplings are analyzed, in particular, by Cardaliaguet and Graber [102].

Convex momentum formulation. As in the Benamou–Brenier problem, the density–velocity integrand and the constraint in (14.46)–(14.47) are not jointly convex. Set $\omega_t = \alpha_t v_t$, or $m_t = \rho_t v_t$ when densities exist, and define the congestion functional

$$C_G(\alpha) := \begin{cases} \displaystyle\int_{\mathbb{R}^d} G(\rho(x))\mathrm{d}x, & \alpha = \rho\,\mathrm{d}x, \\ +\infty, & \text{otherwise.} \end{cases} \tag{14.48}$$

Using the measure perspective action (14.11), the same problem becomes the generalized Benamou–Brenier formula

$$\inf_{(\alpha_t,\omega_t)} \left\{ \int_0^1 \left(\frac{1}{2}\mathbb{J}(\alpha_t,\omega_t) + C_G(\alpha_t) \right) \mathrm{d}t + \int_{\mathbb{R}^d} \Psi\,\mathrm{d}\alpha_1 \right\}, \qquad \partial_t \alpha_t + \operatorname{div}\omega_t = 0, \quad \alpha_{t=0} = \alpha_0. \tag{14.49}$$

Equivalently, in density–momentum variables, the running integrand is

$$\frac{1}{2}J(\rho_t, m_t) + G(\rho_t) = \frac{\|m_t\|^2}{2\rho_t} + G(\rho_t), \tag{14.50}$$

with the vacuum convention of (14.10), under the affine equation $\partial_t \rho_t + \operatorname{div} m_t = 0$.

The perspective term $(\alpha, \omega) \mapsto \mathbb{J}(\alpha, \omega)$ is convex by (14.10)–(14.11), while C_G is convex whenever G is proper and convex. Since the terminal term is linear and the constraints are affine, problem (14.49) is then convex in (α, ω). If G is nonconvex, the momentum substitution still convexifies the kinetic term and linearizes the constraint, but it does not convexify the congestion term.

Optimality system and game interpretation. The potential structure matters because the optimality conditions of the planner reproduce the best-response and consistency equations of the game. The following smooth calculation makes this identification explicit.

Proposition 14.28 (Potential MFG system)**.** *Assume that an optimizer of* (14.49) *has a smooth positive density ρ, that G is differentiable, and set $g = G'$. Then there is a value function u such that*

$$\begin{cases} -\partial_t u + \dfrac{1}{2}\|\nabla u\|^2 = g(\rho), & u_{t=1} = \Psi, \\ \partial_t \rho - \operatorname{div}(\rho \nabla u) = 0, & \rho_{t=0}\,\mathrm{d}x = \alpha_0, \end{cases} \qquad m = -\rho\nabla u. \tag{14.51}$$

Proof. Use $-u$ as the multiplier for $\partial_t \rho + \operatorname{div} m = 0$. After integration by parts, the space–time Lagrangian density is

$$\frac{\|m\|^2}{2\rho} + G(\rho) + \rho\,\partial_t u + m \cdot \nabla u.$$

Stationarity in m gives $m = -\rho\nabla u$. Stationarity in ρ, followed by this substitution, gives the first equation in (14.51). The free terminal density contributes $(\Psi - u_1)\rho_1$, hence $u_1 = \Psi$ under the positivity assumption. Primal feasibility gives the second equation. □

The backward Hamilton–Jacobi–Bellman equation describes the representative agent's optimal response to the density-dependent cost $g(\rho)$; the forward continuity equation transports the population under the resulting feedback $v = -\nabla u$. Their coupling enforces the Nash consistency condition. For nonsmooth G, $g(\rho)$ is replaced by an element of $\partial G(\rho)$; hard density caps then produce a pressure supported on the saturated region.

Hard congestion through a bottleneck. A global hard crowd-capacity constraint is obtained from

$$G_\kappa(r) = \iota_{[0,\kappa]}(r) = \begin{cases} 0, & 0 \leqslant r \leqslant \kappa, \\ +\infty, & \text{otherwise.} \end{cases} \tag{14.52}$$

In order to make the cap active while retaining a prescribed terminal preference, one may replace the linear terminal term in (14.49) by the convex penalty

$$\Gamma_\eta(\rho_1) = \frac{\eta}{2}\int_\Omega |\rho_1(x) - \rho_\star(x)|^2\,\mathrm{d}x, \tag{14.53}$$

where $\rho_\star$ is a desired terminal density. Such L^1 and L^2 endpoint penalties are part of the augmented-Lagrangian experiments of Benamou and Carlier [36]; the two-room hard-congestion geometry follows Benamou, Carlier and Santambrogio [40].

The role of endpoint relaxation depends on the geometry. If two fixed endpoint densities on $\mathbb{R}^d$ satisfy a spatially constant cap, their ordinary $\mathcal{W}_2$ geodesic satisfies the same cap by Proposition 15.22, so the cap does not alter that geodesic. This argument no longer applies inside a nonconvex domain, where straight displacement segments may leave the domain and a narrow passage creates a genuine capacity bottleneck. The terminal penalty (14.53) additionally lets the planner trade terminal mismatch against kinetic cost while enforcing the running cap everywhere.

Figure 14.5 illustrates this mechanism in two rooms connected by a one-cell doorway. Here we specialize the terminal term to the hard constraint $\rho_{t=1} = \rho_\star$, equivalently the indicator terminal cost $\Psi = \iota_{\{\rho_\star\}}$. Both endpoints are uniform on mirrored disks. Writing $M_0 = \|\rho_0\|_\infty$, the constrained experiment uses the tight feasible capacity $\kappa = M_0$: the crowd starts at saturation and cannot compress above its initial packing density. The uncapped hard-endpoint path forms a thin, dense jet through the doorway, whereas the capped path builds a broad saturated queue upstream. The convex density–momentum problem is discretized on a 48×31 spatial grid with eight time slabs. Douglas–Rachford splitting computes the variational solution, and a final Dykstra projection onto the intersection of the affine continuity/end-point constraints and the global density box removes the small feasibility error of the finite splitting iterate.

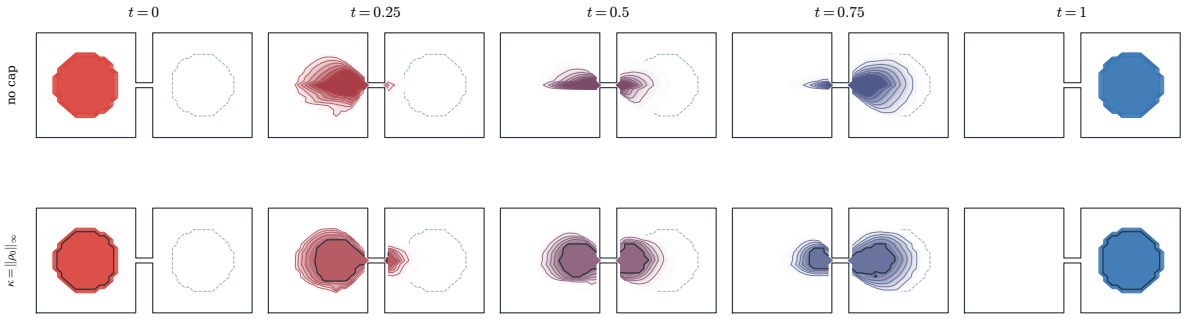


Figure 14.5: Hard-congestion variational MFG through two communicating rooms. The initial and hard terminal profiles are uniform on mirrored disks; the dashed outlines show the latter, and the five columns advance from red to blue. The top row has no cap and concentrates into a narrow doorway jet. The bottom row imposes $\kappa = \|\rho_0\|_\infty$ everywhere, forcing a broader saturated queue. Dark contours mark regions within 1% of the cap, and both rows use the same density levels.

> **Remark 14.29 (Scope of the variational reduction).** A general MFG need not be the optimality system of a global convex functional. The reduction above applies because the coupling is the first variation of the potential C_G and the terminal cost is linear, or more generally convex, in the final law. More general nonlocal, non-potential or non-monotone couplings retain the coupled best-response/consistency interpretation but generally fall outside (14.49).

Computational consequence. The convexification is not merely formal. After conservative space–time discretization, the objective is a sum of local perspective and congestion terms subject to one affine divergence constraint. Augmented-Lagrangian and primal–dual methods can therefore alternate local proximal updates with the solution of a linear space–time equation, closely paralleling the Benamou–Brenier solver. This route was developed for transport and variational MFGs by Benamou and Carlier [36] and is reviewed with numerical examples in [40].

The chapter has separated three dynamic transport mechanisms—local advection, nonlocal jumps, and coupled displacement–reaction—and then shown how the local Benamou–Brenier action becomes a convex population game after adding a potential congestion cost. The next chapter uses dynamic transport geometries to turn energies on measures into evolution equations.

CHAPTER 15

Wasserstein Gradient Flows

Once $\mathcal{W}_2$ is understood as a dynamic metric, one can run gradient descent directly on the space of measures. This chapter derives the Wasserstein gradient from the JKO minimizing-movement scheme, develops convergence through geodesic convexity and functional inequalities, and applies the calculus to interacting particles and mean-field neural networks. It then changes the underlying geometry to obtain mobility, spectral, nonlocal, unbalanced, conditional, and inertial flows.

15.1 Minimizing Movements and Wasserstein Gradients

This first section explains how a variational implicit-Euler step on measures gives rise, in the small-step limit, to a continuity equation driven by the Wasserstein gradient of the energy.

We now consider a function $f(\alpha)$ and seek a minimizing evolution $(\alpha_t)_t$. The general strategy is the following implicit Euler construction.

Definition 15.1 (JKO minimizing movement)**.** Given $\tau > 0$ and $\alpha^{(0)}$, define recursively

$$\alpha^{(\ell+1)} \in \operatorname*{argmin}_{\alpha \in \mathcal{P}_2(\mathbb{R}^d)} \left\{ \frac{1}{2\tau} \mathcal{W}_2(\alpha^{(\ell)}, \alpha)^2 + f(\alpha) \right\}. \tag{15.1}$$

The piecewise-constant interpolation $\alpha_\tau(t) = \alpha^{(\ell)}$ for $t \in [\ell\tau, (\ell+1)\tau)$ is called a minimizing movement.

Euclidean gradient flows. If we restrict (15.1) to finite dimensions and assume $\alpha_t = \delta_{x(t)}$ and $\alpha = \delta_x$ (single Dirac measures), this matches the implicit Euler scheme:

$$x(t+\tau) \in \operatorname*{argmin}_x \left\{ \frac{1}{2\tau} \|x - x(t)\|^2 + h(x) \right\},$$

where $h(x) = f(\delta_x)$. When h is differentiable and the minimizer is unique, the solution is characterized by the implicit Euler formula

$$x(t+\tau) = (\mathrm{Id} + \tau \nabla h)^{-1}(x(t)).$$

In contrast, the explicit Euler scheme is:

$$x(t+\tau) = (\mathrm{Id} - \tau \nabla h)(x(t)) = x(t) - \tau \nabla h(x(t)).$$

Both schemes converge as $\tau \to 0$ to:

$$\dot{x}(t) = -\nabla h(x(t)). \tag{15.2}$$

Vertical derivatives (first- and higher-order variations). Deriving a PDE from the JKO scheme requires distinguishing affine from transport perturbations. The mixture $\alpha_\vartheta = (1-\vartheta)\alpha + \vartheta\beta$ changes mass without prescribing how particles move and defines the *linear functional*, *flat* or *vertical* derivative. The horizontal derivative introduced below instead follows a transport displacement.

Definition 15.2 (First variation)**.** Assume that $f(\alpha) < +\infty$. A measurable function $\delta f(\alpha)$ is a first variation of f at α if, for every $\beta \in \mathcal{P}(\mathbb{R}^d)$ such that $\delta f(\alpha)$ is integrable with respect to $|\beta - \alpha|$ and the mixture α_ϑ remains in the domain of f for small $\vartheta > 0$,

$$f(\alpha_\vartheta) = f(\alpha) + \vartheta \int_{\mathbb{R}^d} [\delta f(\alpha)](x) \mathrm{d}(\beta - \alpha)(x) + o(\vartheta) \qquad (\vartheta \downarrow 0).$$

Because $\beta - \alpha$ has zero mass, $\delta f(\alpha)$ is determined only up to an additive constant. It represents the directional linear form

$$Df(\alpha)[\eta] := \int_{\mathbb{R}^d} [\delta f(\alpha)](x)\mathrm{d}\eta(x)$$

whenever the zero-mass signed perturbation η is admissible. For example, if $f(\alpha) = \int V \,\mathrm{d}\alpha$, then $\delta f(\alpha) = V$: the vertical derivative records the value $V(x)$, whereas the horizontal derivative will record its spatial gradient $\nabla V(x)$.

Higher derivatives are mostly unnecessary for the first-order flows studied in this chapter, but they provide the appropriate notion of smoothness for functionals of measures. For this brief account, fix a compact set $\mathcal{X} \subset \mathbb{R}^d$ and denote by $\mathcal{M}_0(\mathcal{X})$ the finite signed Borel measures on $\mathcal{X}$ with zero total mass.

Definition 15.3 (Higher-order vertical derivatives)**.** Set $D^0 f = f$. The functional f admits vertical derivatives through order r if there are symmetric multilinear forms

$$D^j f(\alpha) : \mathcal{M}_0(\mathcal{X})^j \longrightarrow \mathbb{R}, \qquad 1 \leqslant j \leqslant r,$$

such that, along every mixture segment in $\mathcal{P}(\mathcal{X})$ and for fixed $\eta_1, \dots, \eta_{j-1} \in \mathcal{M}_0(\mathcal{X})$,

$$\frac{\mathrm{d}}{\mathrm{d}\vartheta} D^{j-1} f(\alpha_\vartheta)[\eta_1, \dots, \eta_{j-1}] = D^j f(\alpha_\vartheta)[\beta - \alpha, \eta_1, \dots, \eta_{j-1}].$$

At the endpoints, derivatives are understood one-sided.

When these multilinear forms have integral kernels, the notation

$$D^j f(\alpha)[\eta_1, \dots, \eta_j] = \int_{\mathcal{X}^j} \frac{\delta^j f}{\delta \alpha^j}(\alpha, x_1, \dots, x_j)\mathrm{d}\eta_1(x_1) \cdots \mathrm{d}\eta_j(x_j)$$

is customary. Because the η_i have zero mass, these kernels have additive gauge freedoms; one can fix them by requiring zero mean in each spatial variable. This calculus and its relation to differential calculus on Wasserstein space are developed in [115, 124].

Remark 15.4 (Higher smoothness and faster particle-polynomial approximation). Section 4.4 approximates f by the particle polynomial $\mathsf{B}_n f(\alpha) = \mathbb{E} f(\widehat{\alpha}_n)$ from (4.8). Proposition 4.34 uses only metric Hölder regularity of order $0 < \eta \leqslant 1$. On a geodesic Wasserstein space, a global metric Hölder exponent larger than one would force f to be constant, so higher smoothness must instead be measured after subtracting a Taylor polynomial.
Let $s > 1$, set $r = \lceil s \rceil - 1$, and define

$$\mathrm{Tay}^r_\alpha f(\beta) := f(\alpha) + \sum_{j=1}^{r} \frac{1}{j!} D^j f(\alpha)[(\beta - \alpha)^{\otimes j}].$$

We say that f has Wasserstein–Taylor regularity of order s when its vertical derivatives exist through order r and

$$\left| f(\beta) - \mathrm{Tay}^r_\alpha f(\beta) \right| \leqslant L \mathcal{W}_p(\alpha, \beta)^s. \tag{15.3}$$

The convention $r = \lceil s \rceil - 1$ treats an integer s as the Lipschitz endpoint $C^{s-1,1}$.
Combining the weak empirical expansion of [124] with (15.3) gives, when the centered derivative kernels are uniformly bounded and with $r_{n,p,d}$ from (10.3),

$$\|\mathsf{B}_n f - f\|_\infty \lesssim \begin{cases} r^{\,s}_{n,p,d}, & 1 < s \leqslant 2, \\ n^{-1} + r^{\,s}_{n,p,d}, & s > 2. \end{cases}$$

The reason is simple: $\mathbb{E}(\widehat{\alpha}_n - \alpha) = 0$, so the linear Taylor term vanishes in expectation, while the quadratic term generally produces a bias of order n^{-1}.
A signed Richardson–Romberg combination

$$\mathsf{B}^{[J]}_n f = \sum_{j=1}^{J} a_j \mathsf{B}_{m_j n} f, \qquad \sum_{j=1}^{J} a_j = 1, \qquad \sum_{j=1}^{J} a_j m_j^{-k} = 0 \quad (1 \leqslant k < J),$$

for distinct positive integers m_j, cancels the first $J-1$ bias coefficients. If $s > 2J-2$ and the derivative kernels through order $2J-2$ are uniformly bounded, then

$$\|\mathsf{B}_n^{[J]} f - f\|_\infty \lesssim n^{-J} + r_{n,p,d}^s.$$

Thus higher vertical smoothness can be converted into faster polynomial approximation, although none of this machinery is needed for the first-order PDEs below.

Horizontal derivative (Wasserstein gradient). The vertical derivative asks how f reacts when mass is reweighted. Wasserstein geometry asks how it reacts when mass is moved through a Lagrangian displacement $(\mathrm{Id} + \vartheta v)_\sharp \alpha$. When the first variation is differentiable in x, its spatial gradient represents this horizontal differential.

As $\tau \to 0$, under suitable conditions on f, (15.1) defines a continuous evolution $t \mapsto \alpha_t$. As discussed earlier, this evolution can be described by the Lagrangian advection formula (14.1). The key infinitesimal object is the vector field representing the differential in the Wasserstein metric.

Definition 15.5 (Wasserstein gradient)**.** Assume that f admits a smooth first variation $\delta f(\alpha)$. In the smooth formal calculus on $\mathcal{P}_2(\mathbb{R}^d)$, the Wasserstein gradient of f at α is the gradient vector field

$$\mathbb{W} f(\alpha) = \nabla_x \delta f(\alpha).$$

Remark 15.6 ($\mathcal{W}_2$-Lipschitz functionals and bounded gradients). Section 4.4 introduced Wasserstein–Lipschitz regularity as the endpoint of Wasserstein–Hölder regularity; see Definition 4.30. An L-Lipschitz functional has descending metric slope at most L. Conversely, on a geodesic metric space, if the slope is a strong upper gradient uniformly bounded by L, integrating it along geodesics shows that the functional is L-Lipschitz. In the smooth Otto calculus, under the chain-rule assumptions used in Proposition 15.7, this slope is

$$|\partial f|_{\mathcal{W}_2}(\alpha) = \|\mathbb{W} f(\alpha)\|_{L^2(\alpha)}.$$

Thus

$$\sup_\alpha \|\mathbb{W} f(\alpha)\|_{L^2(\alpha)} \leqslant L$$

is the differential counterpart of $\mathcal{W}_2$-Lipschitz regularity.

The associated formal gradient flow is the continuity equation

$$\frac{\partial \alpha_t}{\partial t} + \operatorname{div}(-\mathbb{W} f(\alpha_t)\alpha_t) = 0. \tag{15.4}$$

The following proposition explains why this vector field is the Riemannian gradient for the $L^2(\alpha)$ metric on velocities.

Proposition 15.7 (Formal Wasserstein gradient)**.** *Assume that f admits a smooth first variation $\delta f(\alpha)$ and that α has a smooth positive density. Let v be a sufficiently regular velocity field, compactly supported or satisfying a no-flux boundary condition. For perturbations generated by $(\mathrm{Id} + \tau v)_\sharp \alpha$, the differential of f is*

$$\frac{\mathrm{d}}{\mathrm{d}\tau}_{|\tau=0} f\big((\mathrm{Id} + \tau v)_\sharp \alpha\big) = \int \langle \nabla \delta f(\alpha)(x),\, v(x)\rangle \mathrm{d}\alpha(x).$$

Hence, for the Riemannian metric $\|v\|^2_{L^2(\alpha)} = \int \|v\|^2 \mathrm{d}\alpha$, the Wasserstein gradient is the vector field

$$\mathbb{W} f(\alpha) = \nabla \delta f(\alpha).$$

Proof. The push-forward expansion gives, in the sense of distributions,

$$(\mathrm{Id} + \tau v)_\sharp \alpha = \alpha - \tau \operatorname{div}(\alpha v) + o(\tau).$$

Using the definition of the first variation,

$$f((\mathrm{Id} + \tau v)_\sharp \alpha) = f(\alpha) - \tau \int \delta f(\alpha) \operatorname{div}(\alpha v) \mathrm{d}x + o(\tau).$$

An integration by parts, with either compact support or vanishing boundary flux, gives

$$-\int \delta f(\alpha)\,\mathrm{div}(\alpha v)\mathrm{d}x = \int \langle \nabla \delta f(\alpha),\, v\rangle \mathrm{d}\alpha.$$

By definition of the Riesz representative for the $L^2(\alpha)$ metric, this representative is $\nabla \delta f(\alpha)$. □

The Wasserstein gradient-flow viewpoint already appears in John D. Lafferty's PhD work, published as "The Density Manifold and Configuration Space Quantization", under the name "density manifold". It was then systematically developed by Otto, who exposed the formal Riemannian structure of this space [411]. Rigorous metric-space treatments and numerical JKO schemes can be found in [12, 38, 425, 230].

From the JKO step to the velocity field. A first-order expansion of the JKO step explains why (15.4) uses the vector field $\mathbb{W} f(\alpha)$. This is a formal local argument: assume that α_t is absolutely continuous and that, for the nearby competitors selected by the JKO step, the optimal map from α_t can be written as $\mathrm{Id} + \tau v$. For these optimal displacement representatives,

$$\mathcal{W}_2^2\big(\alpha_t, (\mathrm{Id} + \tau v)_\sharp \alpha_t\big) = \tau^2 \|v\|^2_{L^2(\alpha_t)}.$$

The local JKO objective therefore reads

$$\min_v \frac{\tau}{2}\|v\|^2_{L^2(\alpha_t)} + f\big((\mathrm{Id} + \tau v)_\sharp \alpha_t\big),$$

where the minimization is restricted to these representatives. The two expansions

$$(\mathrm{Id} + \tau v)_\sharp \alpha_t = \alpha_t - \tau\,\mathrm{div}(v\alpha_t) + o(\tau),$$
$$f\big((\mathrm{Id} + \tau v)_\sharp \alpha_t\big) = f(\alpha_t) + \tau \int \langle \nabla_x \delta f(\alpha_t)(x),\, v(x)\rangle \mathrm{d}\alpha_t(x) + o(\tau)$$

give the first-order problem

$$\min_v f(\alpha_t) + \tau \int \left[\frac{1}{2}\|v(x)\|^2 + \langle \mathbb{W} f(\alpha_t)(x),\, v(x)\rangle\right] \mathrm{d}\alpha_t(x) + o(\tau).$$

Its pointwise minimizer is $v = -\mathbb{W} f(\alpha_t)$, which gives the velocity in the continuity equation. A rigorous passage from the discrete scheme to a curve of maximal slope uses compactness and lower-semicontinuity rather than this formal parametrization; see [38, 12].

Metric derivative and curves of maximal slope. The same descent principle admits a coordinate-free formulation, which is the right language for limits of JKO schemes.

Definition 15.8 (Metric derivative and absolutely continuous curve)**.** For $1 \leqslant p \leqslant +\infty$, a curve $x : [0,T] \to (\mathcal{X}, d)$ belongs to $AC^p([0,T];\mathcal{X})$ if some $g \in L^p(0,T)$ satisfies $d(x_s, x_t) \leqslant \int_s^t g(r)\mathrm{d}r$. Absolutely continuous means AC^1. Its metric derivative is

$$|\dot{x}_t| := \lim_{h\to 0} \frac{d(x_{t+h}, x_t)}{|h|}$$

for a.e. t.

Energy descent is encoded by the slope and its upper-gradient property.

Definition 15.9 (Metric slope and curve of maximal slope)**.** For lower-semicontinuous $f : \mathcal{X} \to (-\infty, +\infty]$, define

$$|\partial f|(x) := \begin{cases} \limsup_{\substack{y\to x \\ y\neq x}} \dfrac{(f(x) - f(y))_+}{d(x,y)}, & f(x) < +\infty, \\ +\infty, & f(x) = +\infty. \end{cases}$$

It is a strong upper gradient if $|f(y_t) - f(y_s)| \leqslant \int_s^t |\partial f|(y_r)|\dot{y}_r|\mathrm{d}r$ along every absolutely continuous

y. When this property holds, a curve x_t is a curve of maximal slope for f if

$$f(x_t) + \frac{1}{2}\int_s^t |\dot{x}_r|^2 \mathrm{d}r + \frac{1}{2}\int_s^t |\partial f|^2(x_r)\mathrm{d}r \leqslant f(x_s), \qquad 0 \leqslant s < t.$$

This is the metric-space formulation of gradient descent [12]; the inequality is stable under weak limits, while in smooth Hilbert or Wasserstein settings it is usually saturated. In $\mathcal{X} = \mathcal{P}_2(\mathbb{R}^d)$ with $d = \mathcal{W}_2$, an absolutely continuous curve admits a continuity-equation velocity v_t and satisfies $|\dot{\alpha}_t|_{\mathcal{W}_2} \leqslant \|v_t\|_{L^2(\alpha_t)}$, with equality for the minimal velocity. For smooth energies, $|\partial f|(\alpha_t) = \|\mathbb{W} f(\alpha_t)\|_{L^2(\alpha_t)}$, so the maximal-slope condition reduces to the velocity-field equation $v_t = -\mathbb{W} f(\alpha_t)$ derived above. The energy consequences of this viewpoint are used below in Section 15.2.

We now detail examples of such Wasserstein gradient flows.

Example 15.10 (Application to single-cell gradient-flow models). Instead of only coupling successive snapshots, one may posit that a latent population law evolves by

$$\partial_t \alpha_t + \operatorname{div}(\alpha_t v_t) = 0, \qquad v_t = -\mathbb{W} f(\alpha_t) = -\nabla \delta f(\alpha_t).$$

Here $\delta f(\alpha_t)$ is the first variation, with the convention introduced above. For example, if $\alpha = \rho\,\mathrm{d}x$, the energy $f(\alpha) = \int V \mathrm{d}\alpha + \sigma^2 \int \rho \log \rho \,\mathrm{d}x$ gives a drift toward a potential landscape V together with diffusion of strength σ^2. In single-cell modeling, V, interaction terms or a neural approximation of v_t can be fitted from unpaired population snapshots. This links Waddington-landscape intuition to the mathematical language of dynamic OT, Fokker–Planck equations and flow matching [520, 326, 308].

Figure 15.1 shows the same minimizing-movement construction both in density coordinates and through the motion of quantiles.

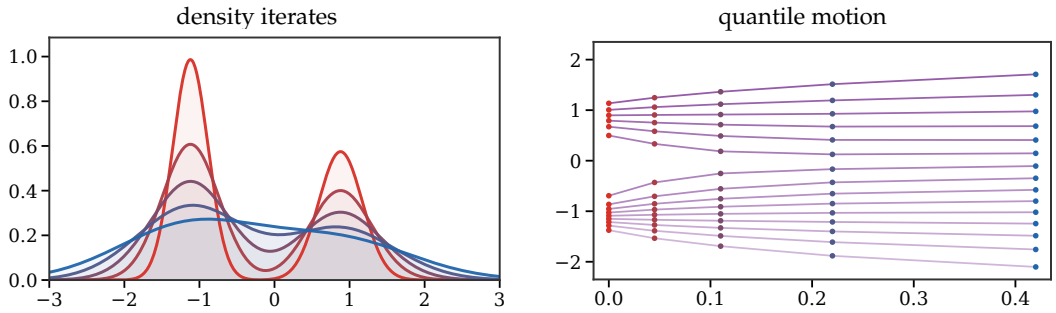


Figure 15.1: JKO minimizing movements for the entropy flow in one dimension. The left panel displays successive implicit-Euler minimizers for the heat equation, colored from red to blue. The right panel tracks inverse CDF values $Q_t(s) = F_t^{-1}(s)$ for selected probability levels s, giving a Lagrangian view of the proximal movement in Wasserstein space.

Discrete evolutions. If f admits a first variation whose spatial gradient is well defined at empirical measures, the characteristic form of (15.4) preserves the number and weights of the atoms. For equal weights, write $\alpha_t = \frac{1}{n}\sum_i \delta_{x_i(t)}$. The particles $X(t) := (x_i(t))_i$ evolve according to the coupled ODEs

$$\dot{x}_i(t) = -n \nabla_{x_i} F(X(t)), \tag{15.5}$$

where $F(X) := f\left(\frac{1}{n}\sum_i \delta_{x_i}\right)$ and the factor n comes from the empirical Wasserstein metric $\frac{1}{n}\sum_i \|\dot{x}_i\|^2$.

Linear functionals. The first benchmark is a functional whose first variation does not depend on the current measure.

Example 15.11 (Linear potentials generate independent particles).

Take a linear functional $$f(\alpha) = \int h(x)\mathrm{d}\alpha(x). \tag{15.6}$$

Then $\delta f(\alpha) = h$ is independent of α, and the flow (15.4) becomes

$$\frac{\partial \alpha_t}{\partial t} + \operatorname{div}(-\nabla h \alpha_t) = 0.$$

This implies particles move independently according to the usual gradient flow (15.2).

Shannon neg-entropy. Entropy gives the opposite benchmark: the flow is no longer a deterministic push-forward of particles, but a diffusion of density.

Definition 15.12 (Score function)**.** For $\alpha = \rho\,\mathrm{d}x$ with differentiable positive density ρ, its score is $\nabla \log \rho$. Relative to β, its relative score is $\nabla \log(\mathrm{d}\alpha/\mathrm{d}\beta)$ whenever this logarithmic density has a weak gradient.

Example 15.13 (Entropy generates heat and porous-medium flows). The canonical density-dependent functional is the Shannon neg-entropy

$$f(\alpha) = \int \log\left(\frac{\mathrm{d}\alpha}{\mathrm{d}x}(x)\right)\mathrm{d}\alpha(x). \tag{15.7}$$

If $\alpha = \rho \mathrm{d}x$, then $\delta f(\alpha) = \log\rho + 1$, up to an immaterial additive constant, so $\mathbb{W} f(\alpha) = \nabla \log\rho$ is the score. The flow (15.4) becomes the heat equation

$$\partial_t \rho_t = \Delta \rho_t.$$

Other entropy functionals lead to nonlinear diffusion equations; finite-volume and particle discretizations are discussed in [116, 242, 361, 204].

For a generalized entropy $$f(\rho \mathrm{d}x) = \int g(\rho(x))\mathrm{d}x, \tag{15.8}$$

with a scalar convex function g, one obtains nonlinear diffusions in the smooth-density regime:

$$\frac{\partial \rho_t}{\partial t} = \Delta(P(\rho_t)),$$

where the pressure P satisfies $P'(s) = sg''(s)$. For example, $g(s) = s\log(s)$ gives $P(s) = s$ and recovers (15.7), while $g(s) = s^m/(m-1)$, $m > 1$, gives $P(s) = s^m$ up to an additive constant and yields the porous-medium equation.

Remark 15.14 (Two gradient-flow interpretations of the heat equation). The heat equation already illustrates that a PDE does not determine a unique gradient-flow structure. Write $\alpha = \rho \mathrm{d}x$ on a flat domain, with periodic or no-flux boundary conditions. In the Hilbert geometry of densities, the squared distance and the Dirichlet energy are

$$d^2_{L^2}(\alpha,\beta) = \int |\rho_\alpha(x) - \rho_\beta(x)|^2 \mathrm{d}x, \qquad \mathcal{D}(\alpha) = \frac{1}{2}\int \|\nabla\rho(x)\|^2 \mathrm{d}x,$$

so $\delta\mathcal{D}/\delta\rho = -\Delta\rho$ and the L^2 gradient flow $\partial_t\rho_t = -\delta\mathcal{D}/\delta\rho(\rho_t)$ gives $\partial_t\rho_t = \Delta\rho_t$. This viewpoint says that heat flow decreases oscillations and regularizes the density. In Wasserstein geometry, the same equation is instead the $\mathcal{W}_2$ gradient flow of the Shannon entropy $\int \rho\log\rho \mathrm{d}x$, so the driving mechanism is entropic spreading. The example is a useful warning: explaining a dynamics as a gradient flow requires specifying both an energy and a metric, and different pairs (f, d) may produce the same PDE.

Figure 15.2 contrasts the spreading induced by the linear heat equation with the compactly supported, density-dependent propagation of two porous-medium flows.

Interaction energies. In a similar spirit, to obtain nonlinear evolutions without requiring the measure to have density, one can consider

$$f(\alpha) := \iint k(x,y)\mathrm{d}\alpha(x)\mathrm{d}\alpha(y). \tag{15.9}$$

For a symmetric kernel k:

$$\delta f(\alpha)(x) = 2\int k(x,y)\mathrm{d}\alpha(y), \quad \mathbb{W} f(\alpha)(x) = 2\int \nabla_x k(x,y)\mathrm{d}\alpha(y).$$

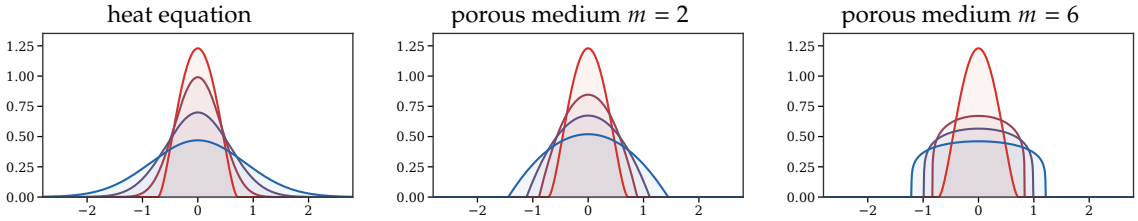


Figure 15.2: Entropy-driven Wasserstein gradient flows from the same compact initial density. The heat flow is generated by Shannon entropy $g(\rho) = \rho \log \rho$ and instantly develops Gaussian tails. The porous-medium flows use the power entropy $g(\rho) = \rho^m/(m-1)$, hence $\partial_t \rho = \Delta(\rho^m)$: the middle panel has $m = 2$, while the right panel has the stronger nonlinearity $m = 6$, i.e. $\partial_t \rho = \Delta(\rho^6)$. Larger powers diffuse mainly where the density is high, producing a flatter core and a sharper compact free boundary.

For $\alpha_0 = \frac{1}{n}\sum_i \delta_{x_i}$, the flow (15.4) implies particles $(x_i(t))_i$ obey:

$$\dot{x}_i(t) = -\frac{2}{n}\sum_j \nabla k(x_i(t), x_j(t)).$$

If k is positive definite, or more generally conditionally positive definite on signed measures of zero total mass as for the energy-distance kernel $k(x, y) = -\|x - y\|$, and one minimizes the squared kernel discrepancy to a teacher distribution β, then

$$\|\alpha - \beta\|_k^2 = \iint k \mathrm{d}\alpha \mathrm{d}\alpha - 2\int \left(\int k(x, y)\mathrm{d}\beta(y)\right) \mathrm{d}\alpha(x) + \text{constant}.$$

Thus MMD-type training energies are exactly an interaction energy plus a linear potential; the teacher distribution appears through the potential $x \mapsto -2\int k(x, y)\mathrm{d}\beta(y)$. The corresponding empirical Wasserstein gradient flow is

$$\dot{x}_i(t) = -\frac{2}{n}\sum_j \nabla_x k(x_i(t), x_j(t)) + 2\int \nabla_x k(x_i(t), y)\mathrm{d}\beta(y).$$

The first term is a kernelized self-interaction, while the second is the attraction induced by the continuous teacher kernel mean. At the continuum level, characteristic positive-definite kernels, and the Euclidean energy-distance kernel on probability measures, have β as the unique minimizer of $\|\alpha - \beta\|_k^2$. For a finite number of particles, however, the flow can only form a kernelized quadrature of β, and small particle systems may cover the target modes poorly. Figure 15.3 illustrates this finite-particle effect.

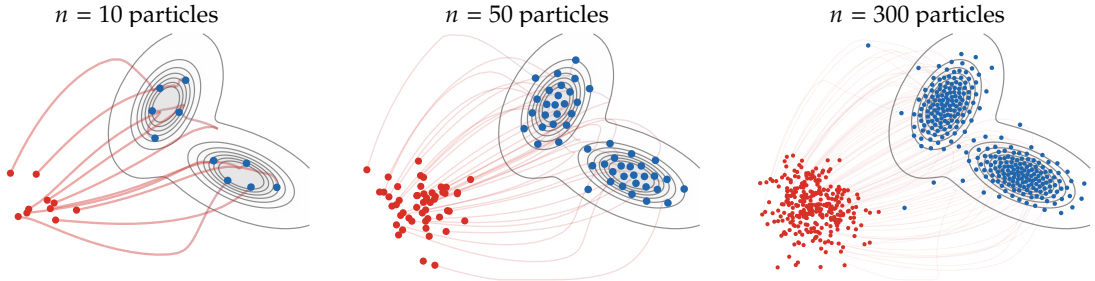


Figure 15.3: Particle count in an energy-distance Wasserstein particle flow. The squared MMD-type discrepancy uses the energy-distance kernel $k(x, y) = -\|x - y\|$ and targets a smooth two-Gaussian teacher distribution. The teacher itself is shown only through true density contours, while red dots are a compact shifted Gaussian initialization placed away from the target, red-to-blue curves show a thinned subset of particle trajectories, and blue dots show the stabilized long-time particles. With too few particles, the empirical measure forms a sparse kernelized quadrature and may under-cover the target modes; increasing n makes the particle cloud approximate the continuous target geometry more faithfully.

The preceding particle flow minimizes a squared kernel discrepancy. Figure 15.4 contrasts this with flows generated by three discrepancies toward the same target. For the energy distance and $\mathcal{W}_2$, the

energy is the unsquared distance, so the Wasserstein velocity is normalized by the current discrepancy value away from the target. The KL case is different: it is a relative-entropy gradient flow and therefore follows the Fokker–Planck diffusion toward the target. The comparison keeps the endpoint fixed and makes the geometry visible through the transient densities.

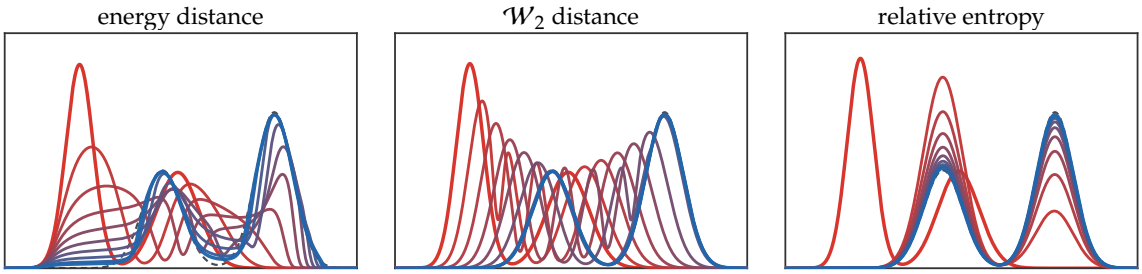


Figure 15.4: Wasserstein gradient flows of three discrepancies toward the same one-dimensional Gaussian-mixture target, shown as a dashed black density. The source density is also a Gaussian mixture. Red-to-blue curves are snapshots of α_t. The energy-distance and $\mathcal{W}_2$ panels descend the unsquared distances and use many equal-weight quantile particles: in one dimension the energy distance is obtained from $\mathrm{ED}^2 = 2\int |F_\alpha - F_\beta|^2 \mathrm{d}x$, while $\mathcal{W}_2$ uses monotone quantile matching. The relative-entropy panel solves the reversible Fokker–Planck equation $\partial_t \rho = \partial_x(\rho_\beta\, \partial_x(\rho/\rho_\beta))$ by an implicit finite-volume scheme, so that ρ_β is the discrete stationary density.

Figure 15.5 complements this target-matching example by isolating three self-interaction regimes.

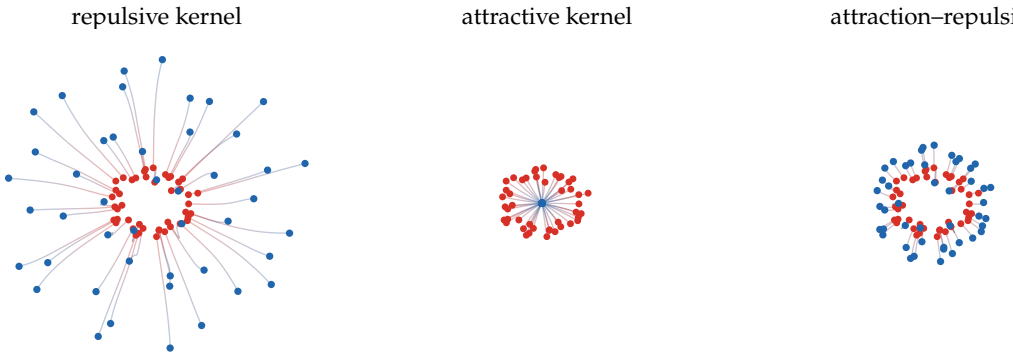


Figure 15.5: Interaction-energy particle flows for three choices of k. A positive Gaussian kernel $k(x,y) = \exp(-\|x-y\|^2/(2\sigma^2))$ produces short-range repulsion under Wasserstein descent; changing its sign produces attraction and collapse; adding a quadratic long-range attraction to the repulsive kernel yields a balanced attraction–repulsion dynamics. The curves use arclength-based red-to-blue coloring along a longer integration of the coupled particle ODE (15.5).

Figure 15.6 then compares the trajectory geometries produced by several discrepancies on common source and target clouds.

Stochastic particles and McKean–Vlasov limits. More generally, deterministic particle flows have stochastic counterparts, where Brownian noise at the particle level becomes an entropy term at the level of measures. If the drift does not depend on the empirical measure, each particle evolves independently according to

$$\mathrm{d}X_t = b(X_t)\mathrm{d}t + \sqrt{2}\sigma \mathrm{d}B_t,$$

and the one-particle law $\alpha_t = \rho_t \mathrm{d}x$ directly satisfies the linear Fokker–Planck equation

$$\partial_t \rho_t = -\operatorname{div}(b\rho_t) + \sigma^2 \Delta \rho_t.$$

Example 15.15 (Langevin drift as a free-energy flow). If $b = -\nabla V$, this linear Fokker–Planck equation is the $\mathcal{W}_2$ gradient flow of the free energy

$$\rho \mapsto \int V\rho \,\mathrm{d}x + \sigma^2 \int \rho \log \rho \,\mathrm{d}x.$$

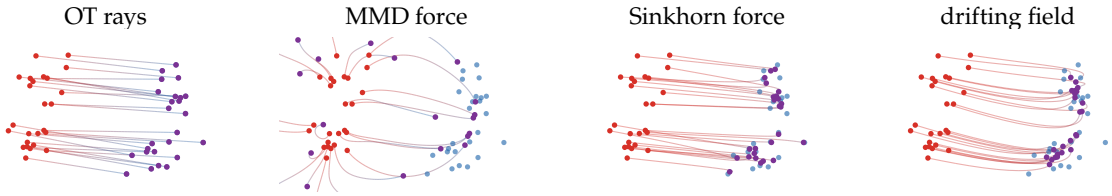


Figure 15.6: Particle trajectories induced by different discrepancy geometries. The red particles and blue target cloud are the same in all panels. Straight OT displacement produces rays from an optimal matching; an MMD-type witness field gives smoother nonlocal forces; the Sinkhorn-divergence force is an entropic, debiased transport attraction; and the normalized drifting field combines attraction to data with self-repulsion. The figure is qualitative: it compares geometric behavior, not solver performance.

The mean-field case is different: the drift is recomputed from the current empirical distribution of all particles,

$$\mathrm{d}X_i^n(t) = b(X_i^n(t), \alpha_t^n)\mathrm{d}t + \sqrt{2}\sigma \mathrm{d}B_i(t), \qquad \alpha_t^n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i^n(t)}.$$

For finite n, the empirical law α_t^n is itself random. Under suitable Lipschitz, growth and chaotic-initialization assumptions, propagation of chaos states that finitely many particles become asymptotically independent as $n \to \infty$, all with the same deterministic law $\rho_t \mathrm{d}x$; equivalently, the empirical measure α_t^n converges in probability to this law. The limiting density solves the nonlinear Fokker–Planck, or McKean–Vlasov, equation

$$\partial_t \rho_t = -\operatorname{div}\big(b(x, \rho_t)\rho_t\big) + \sigma^2 \Delta \rho_t.$$

When the interaction drift has variational form

$$b(x, \rho) = -\nabla \frac{\delta \mathcal{E}}{\delta \rho}(x),$$

this PDE is the Wasserstein gradient flow of the entropy-regularized energy

$$\mathcal{E}(\rho) + \sigma^2 \int \rho \log \rho \, \mathrm{d}x.$$

For a prescribed target $\beta = \rho_\beta \mathrm{d}x$, take $b = \sigma^2 \nabla \log \rho_\beta$. The SDE, the Fokker–Planck equation, and the deterministic continuity equation with score velocity are then three representations of the same gradient flow:

$$\partial_t \rho_t = \sigma^2 \operatorname{div}\left(\rho_t \nabla \log \frac{\rho_t}{\rho_\beta}\right), \qquad v_t = \sigma^2\big(\nabla \log \rho_\beta - \nabla \log \rho_t\big). \tag{15.10}$$

Figure 15.7 compares numerical realizations of precisely these three descriptions.

Gradient flows under density constraints. The same minimizing-movement formalism also handles nonsmooth energies. A prototypical example is a linear energy, generated by a potential, with a hard upper bound on the density,

$$f_{\rho_{\max}}(\rho \mathrm{d}x) = \int_\Omega h(x)\rho(x)\mathrm{d}x + \iota_{\mathcal{K}_{\rho_{\max}}}(\rho \mathrm{d}x), \qquad \mathcal{K}_{\rho_{\max}} = \{\rho \mathrm{d}x \in \mathcal{P}_2(\Omega)\ ;\ 0 \leqslant \rho \leqslant \rho_{\max} \text{ a.e.}\}, \tag{15.11}$$

where $\Omega \subset \mathbb{R}^d$ is a convex domain and $\mathcal{K}_{\rho_{\max}}$ is assumed nonempty; if Ω has finite volume, this requires $\rho_{\max}|\Omega| \geqslant 1$. The indicator term has no classical first variation; it contributes the normal cone to $\mathcal{K}_{\rho_{\max}}$ in the Wasserstein first-order condition. The JKO step becomes

$$\alpha^{(\ell+1)} \in \operatorname*{argmin}_{\alpha \in \mathcal{K}_{\rho_{\max}}} \frac{1}{2\tau} \mathcal{W}_2^2(\alpha^{(\ell)}, \alpha) + \int_\Omega h \, \mathrm{d}\alpha. \tag{15.12}$$

This is the mathematical mechanism behind macroscopic crowd-motion models with congestion, where the desired velocity $-\nabla h$ is projected onto velocities that do not increase density beyond the maximal value $\rho_{\max}$ [370, 471].

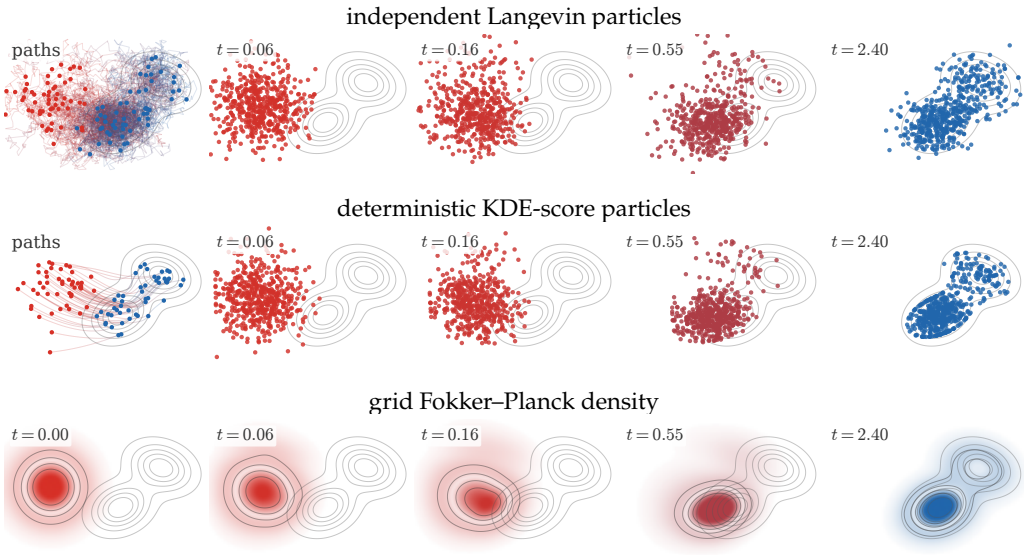


Figure 15.7: Three numerical representations of an entropy-regularized Wasserstein gradient flow. The target β is a two-Gaussian density shifted to the right of an initially isotropic Gaussian density. The first row simulates independent Langevin particles and displays a thinned set of trajectories in the left panel. The second row evolves many deterministic particles with the score velocity in (15.10), estimating $\nabla \log \rho_t$ by a kernel-density estimator; only representative trajectories and particle subsets are displayed. The third row solves the corresponding Fokker–Planck equation on a grid, starting from the initial density in the left panel. The remaining columns use front-loaded times, so that the onset of the flow and the later deformation toward a bimodal law are both visible.

The hard cap is not especially benign numerically: the projection in the JKO step and the pressure field below can be difficult to compute accurately. Its advantage is geometric. Proposition 15.22 in Section 15.2 shows that $\mathcal{K}_{\rho_{\max}}$ is geodesically convex, so the indicator of the constraint is harmless for the usual geodesic-convex convergence guarantees.

The formal PDE contains a pressure field p_t, the Lagrange multiplier of the constraint. With no-flux boundary condition $\rho_t \nabla(h + p_t) \cdot n = 0$ on $\partial\Omega$, the constrained gradient flow is the complementarity system

$$\partial_t \rho_t = \operatorname{div}\left(\rho_t \nabla(h + p_t)\right), \qquad 0 \leqslant \rho_t \leqslant \rho_{\max}, \qquad p_t \geqslant 0, \qquad p_t(\rho_{\max} - \rho_t) = 0. \tag{15.13}$$

The pressure vanishes in the unsaturated region and acts only where $\rho_t = \rho_{\max}$, pushing mass away from congested zones so that the density cap remains satisfied.

Figure 15.8 shows how decreasing the cap converts free concentration into a saturated region whose width is fixed by mass conservation.

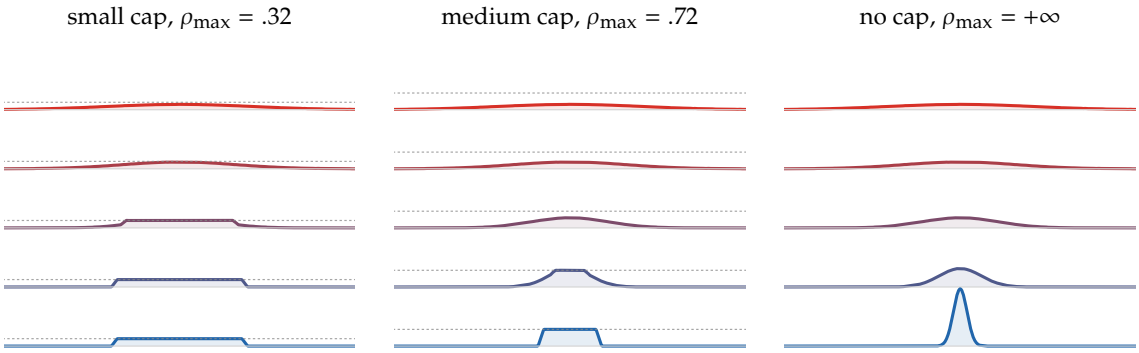


Figure 15.8: Density-constrained gradient flow under a quadratic attraction. The one-dimensional flow is computed in quantile variables by explicit Lagrangian descent followed by the isotonic projection enforcing $\partial_s q \geqslant 1/\rho_{\max}$. Each panel stacks five times vertically, colored from red to blue, using the same vertical density scale in all three panels. The dashed gray line marks the maximal density $\rho_{\max}$ for the constrained runs. A small cap forces the mass to form a wide saturated block, a medium cap allows a narrower congested region, and the unconstrained flow concentrates freely near the minimizer of h.

Multi-species gradient flows. Several applications involve several densities living on the same physical space: chemical species, color channels, cell populations or competing phases. Let $\Omega \subset \mathbb{R}^d$ be the physical domain and assume that the component masses $\mathbf{m} = (m_1, \dots, m_p)$ are positive and fixed. A convenient notation is

$$\mathcal{P}_{2,\mathbf{m}}(\Omega; \mathbb{R}_+^p) = \left\{ \alpha = (\alpha_1, \dots, \alpha_p) \,;\; \alpha_i \in \mathcal{M}_+(\Omega),\; \alpha_i(\Omega) = m_i,\; \int_\Omega \|x\|^2 \mathrm{d}\alpha_i(x) < +\infty \right\}. \tag{15.14}$$

Since the components are finite measures rather than necessarily probabilities, set

$$\mathcal{W}_{2,\oplus}^2(\alpha, \eta) = \sum_{i=1}^p m_i \, \mathcal{W}_2^2\left(\frac{\alpha_i}{m_i}, \frac{\eta_i}{m_i}\right). \tag{15.15}$$

This is the diagonal, independent-channel geometry; cross-effects may still enter through the energy or through constraints. In the dynamic language used later for vector-valued measures, this is the special case of a diagonal mobility; see Proposition 13.4. The corresponding JKO step is

$$\alpha^{(\ell+1)} \in \operatorname*{argmin}_{\alpha \in \mathcal{P}_{2,\mathbf{m}}(\Omega;\mathbb{R}_+^p)} \frac{1}{2\tau} \mathcal{W}_{2,\oplus}^2(\alpha^{(\ell)}, \alpha) + f(\alpha). \tag{15.16}$$

For smooth densities and first variations $\varphi_i = \delta f / \delta \rho_i$, this yields

$$\partial_t \rho_i = \operatorname{div}(\rho_i \nabla \varphi_i), \qquad \varphi_i = \frac{\delta f}{\delta \rho_i}, \qquad i = 1, \dots, p. \tag{15.17}$$

The equations are independent only if f is separable. Otherwise the coupling enters through the first variations φ_i, not through the metric tensor. Carlier, Chizat and Laborde [107] use displacement smoothness of entropic OT to prove well-posedness of Wasserstein gradient flows that include multi-species systems. Congestion and shared-density variants are closely related to crowd and traffic models [370, 113, 471].

A second useful model imposes a pointwise composition constraint. Given a fixed reference measure $\beta = \rho_\beta \mathrm{d}x$ with total mass $\sum_i m_i$, set

$$C_\beta = \left\{ \alpha \in \mathcal{P}_{2,\mathbf{m}}(\Omega; \mathbb{R}_+^p) \,;\; \sum_{i=1}^p \alpha_i = \beta \right\}. \tag{15.18}$$

The constrained JKO step is

$$\alpha^{(\ell+1)} \in \operatorname*{argmin}_{\alpha \in C_\beta} \frac{1}{2\tau} \mathcal{W}_{2,\oplus}^2(\alpha^{(\ell)}, \alpha) + f(\alpha). \tag{15.19}$$

In the smooth-density regime, the formal constrained flow has a common pressure λ_t:

$$\partial_t \rho_i = \operatorname{div}\left(\rho_i \nabla(\varphi_i - \lambda_t)\right), \qquad \operatorname{div}(\rho_\beta \nabla \lambda_t) = \operatorname{div}\left(\sum_{i=1}^{p} \rho_i \nabla \varphi_i\right). \tag{15.20}$$

For the separable Shannon entropy $f(\alpha) = \sum_i \int_\Omega \rho_i \log \rho_i \mathrm{d}x$, this becomes

$$\partial_t \rho_i = \Delta \rho_i - \operatorname{div}(\rho_i \nabla \lambda_t), \qquad \operatorname{div}(b \nabla \lambda_t) = \Delta b. \tag{15.21}$$

When b is constant, the pressure can be chosen constant and the system reduces to independent heat equations which nevertheless preserve the pointwise sum $\sum_i \rho_i = b$. This product geometry should be contrasted with the vector-valued Benamou–Brenier distances of Section 13.1, where the metric itself can couple the components through a non-diagonal mobility.

Figure 15.9 visualizes this constant-total-density case for two, three, and five species.

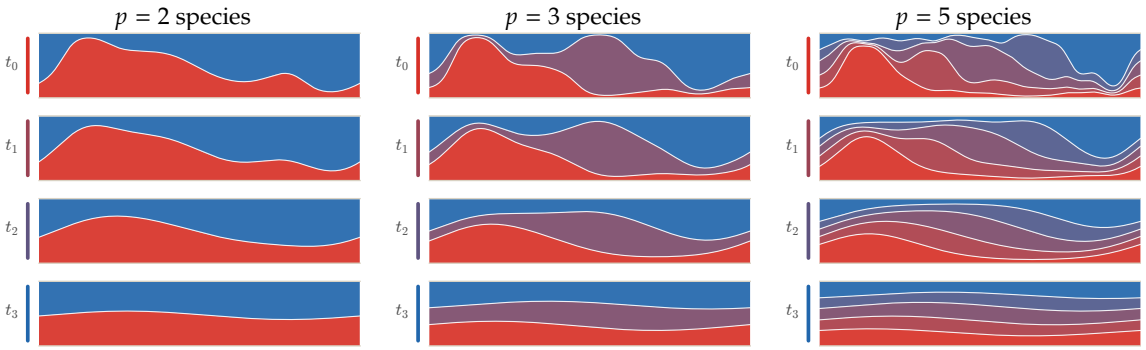


Figure 15.9: Multi-species entropy flow under the shared-density constraint $\sum_i \rho_i = 1$ on the periodic interval. Each row is a time snapshot of the stacked densities, with the species shown as colored bands and time increasing from top to bottom. For a constant total density, the pressure in (15.21) is constant, so each component follows the heat equation while the sum of the bands remains exactly flat.

15.2 Geodesic Convexity and Convergence

Geodesic convexity is the convexity notion adapted to Wasserstein geometry. It is the condition that turns the formal gradient-flow calculus into a convergence theory.

Geodesics and convexity. A constant-speed $\mathcal{W}_2$ geodesic between α_0 and α_1 is obtained, as in Definition 4.12, from any optimal coupling $\pi^\star \in \mathcal{U}(\alpha_0, \alpha_1)$ by the McCann interpolation

$$\alpha_t = ((1-t)P_0 + tP_1)_\sharp \pi^\star, \qquad t \in [0,1],$$

where $P_0(x,y) = x$ and $P_1(x,y) = y$. If the optimal plan is induced by a Brenier map T, this reduces to $((1-t)\mathrm{Id} + tT)_\sharp \alpha_0$. The coupling formula is important because geodesics exist even when no Monge map exists, for instance when a Dirac mass must split.

Definition 15.16 (Geodesic convexity)**.** A functional f on $\mathcal{P}_2(\mathbb{R}^d)$ is geodesically convex if for every $\mathcal{W}_2$ geodesic $(\alpha_t)_t$,

$$f(\alpha_t) \leqslant (1-t) f(\alpha_0) + t f(\alpha_1).$$

It is λ-geodesically convex if the right-hand side is improved by $-\frac{\lambda}{2} t(1-t)\mathcal{W}_2^2(\alpha_0, \alpha_1)$.

Proposition 15.17 (Geodesic convexity of linear and quadratic energies)**.** *The following formal statements hold on $\mathcal{P}_2(\mathbb{R}^d)$.*

1. *If h is convex, then the linear energy $\alpha \mapsto \int h \mathrm{d}\alpha$ is geodesically convex; if h is λ-strongly convex, it is λ-geodesically convex. Conversely, if this linear energy is geodesically convex for all Dirac endpoints, then h is convex.*

2. *More generally, let $k \geqslant 2$ and let $H_k : (\mathbb{R}^d)^k \to \mathbb{R} \cup \{+\infty\}$ be convex. Then the polynomial energy*

$$\alpha \mapsto \int_{(\mathbb{R}^d)^k} H_k(x_1, \dots, x_k)\, \mathrm{d}\alpha(x_1) \cdots \mathrm{d}\alpha(x_k)$$

is geodesically convex whenever the integral is well defined. If H_k is λ-strongly convex on the product space, this energy is $k\lambda$-geodesically convex. In particular, the quadratic interaction

$$\alpha \mapsto \frac{1}{2} \iint W(x, y) \mathrm{d}\alpha(x) \mathrm{d}\alpha(y)$$

is geodesically convex when W is convex on $\mathbb{R}^d \times \mathbb{R}^d$, and it is λ-geodesically convex when W is λ-strongly convex there.

Proof. Let π be an optimal coupling between α_0 and α_1, and set $X_t = (1-t)X_0 + tX_1$ under $(X_0, X_1) \sim \pi$. Convexity of h gives $h(X_t) \leqslant (1-t)h(X_0) + th(X_1)$, and strong convexity gives the additional term $-\frac{\lambda}{2}t(1-t)\|X_0 - X_1\|^2$. Integrating proves the first claim. Necessity follows by testing on Dirac masses: the geodesic between δ_x and δ_y is $\delta_{(1-t)x+ty}$, so geodesic convexity of $\int h \mathrm{d}\alpha$ gives the usual convexity inequality for h.

For the polynomial statement, take k independent copies $(X_0^\ell, X_1^\ell)_{\ell=1}^k$ of the same optimal coupling and define $X_t^\ell = (1-t)X_0^\ell + tX_1^\ell$. Convexity of H_k on the product space gives

$$H_k(X_t^1, \dots, X_t^k) \leqslant (1-t)H_k(X_0^1, \dots, X_0^k) + tH_k(X_1^1, \dots, X_1^k).$$

Integrating over the product coupling proves the claim. If H_k is λ-strongly convex, the additional term is

$$-\frac{\lambda}{2}t(1-t)\sum_{\ell=1}^k \mathbb{E}\|X_0^\ell - X_1^\ell\|^2 = -\frac{k\lambda}{2}t(1-t)\,\mathcal{W}_2^2(\alpha_0, \alpha_1),$$

which proves the stated constant. For the quadratic interaction, the prefactor $1/2$ cancels the factor $k = 2$, leaving the constant λ. □

The polynomial criterion is useful but restrictive for kernel losses. The RKHS/MMD kernels introduced in Section 7.2 are usually chosen for positive definiteness and statistical smoothing, not for convexity as functions of (x, y). Gaussian and Laplace kernels, for instance, are not convex on $\mathbb{R}^d \times \mathbb{R}^d$, so positive definiteness of k does not imply that $\alpha \mapsto \mathrm{MMD}_k^2(\alpha, \beta)$ is geodesically convex. In fact, such MMD losses are not geodesically convex in general. One-dimensional monotone transport gives a sharper exception for distance-type kernels, because ordered quantiles keep the sign of pairwise differences fixed along geodesics; this is the one-dimensional displacement-convexity mechanism emphasized, for instance, in [469].

Proposition 15.18 (One-dimensional interactions and energy-distance MMD)**.** *Let $\varphi : \mathbb{R}_+ \to \mathbb{R}$ be continuous with at most quadratic growth, and define on $\mathcal{P}_2(\mathbb{R})$*

$$\mathcal{I}_\varphi(\alpha) := \frac{1}{2} \iint \varphi(|x-y|) \mathrm{d}\alpha(x) \mathrm{d}\alpha(y).$$

Then $\mathcal{I}_\varphi$ is geodesically convex for $\mathcal{W}_2$ on $\mathcal{P}_2(\mathbb{R})$ if and only if φ is convex on $\mathbb{R}_+$.

Fix also $\beta \in \mathcal{P}_2(\mathbb{R})$. For the conditionally positive kernel $k(x, y) = -|x-y|$ from Definition 7.7, the squared MMD of Definition 7.9, equivalently the squared energy distance,

$$\mathcal{E}_\beta(\alpha) := \mathrm{MMD}_k^2(\alpha, \beta) = \iint -|x-y|\, \mathrm{d}(\alpha-\beta)(x) \mathrm{d}(\alpha-\beta)(y),$$

is geodesically convex as a function of α.

Proof. Let Q_0, Q_1 be the quantile functions of α_0, α_1, and let $Q_t = (1-t)Q_0 + tQ_1$ be the quantile function of the one-dimensional $\mathcal{W}_2$ geodesic α_t. For $r > s$, monotonicity gives $Q_i(r) - Q_i(s) \geqslant 0$ for $i = 0, 1$, and hence

$$Q_t(r) - Q_t(s) = (1-t)(Q_0(r) - Q_0(s)) + t(Q_1(r) - Q_1(s)).$$

Using the symmetry of the double integral and the quantile parametrization, one can write

$$\mathcal{I}_\varphi(\alpha_t) = \int_0^1 \int_0^r \varphi(Q_t(r) - Q_t(s))\,\mathrm{d}s\,\mathrm{d}r.$$

Convexity of φ on $\mathbb{R}_+$ gives the desired convexity after integration.

Conversely, test geodesic convexity on two-point measures $\alpha_i = \frac{1}{2}(\delta_0 + \delta_{a_i})$, with $a_i \geqslant 0$. Their monotone geodesic is $\alpha_t = \frac{1}{2}(\delta_0 + \delta_{(1-t)a_0 + t a_1})$. Since

$$\mathcal{I}_\varphi(\alpha_t) = \frac{1}{4}\varphi(0) + \frac{1}{4}\varphi((1-t)a_0 + t a_1),$$

and the identical $\frac{1}{4}\varphi(0)$ terms cancel from the geodesic-convexity inequality, one obtains the ordinary convexity inequality for φ on $\mathbb{R}_+$.

For the energy distance, the function $\varphi(r) = -r$ is affine on $\mathbb{R}_+$, so the self-interaction $-\iint |x-y|\mathrm{d}\alpha(x)\mathrm{d}\alpha(y)$ is affine along one-dimensional Wasserstein geodesics even though $z \mapsto -|z|$ is not convex on $\mathbb{R}$. Write Q_β for the quantile function of β. Expanding the MMD and dropping the term depending only on β gives

$$\mathcal{E}_\beta(\alpha) = 2\int_0^1\int_0^1 |Q_\alpha(r) - Q_\beta(s)|\,\mathrm{d}r\,\mathrm{d}s - \int_0^1\int_0^1 |Q_\alpha(r) - Q_\alpha(s)|\,\mathrm{d}r\,\mathrm{d}s + \text{constant}.$$

Here the constant is independent of α. The positive cross-distance term is convex in Q_α, since the absolute value is convex and Q_t is affine in t. The self-distance term is affine along quantile geodesics because

$$\int_0^1\int_0^1 |Q_t(r) - Q_t(s)|\,\mathrm{d}r\,\mathrm{d}s = 2\int_0^1\int_0^r (Q_t(r) - Q_t(s))\,\mathrm{d}s\,\mathrm{d}r.$$

Thus $\alpha \mapsto \mathcal{E}_\beta(\alpha)$ is geodesically convex. □

This one-dimensional result is the mechanism behind the favorable behavior of energy-distance particle flows discussed earlier in this chapter and in recent analyses of distance-kernel MMD flows [195]. It should not be read as a general MMD statement: for a typical positive definite kernel k, the decomposition

$$\mathrm{MMD}_k^2(\alpha,\beta) = \iint k\,\mathrm{d}\alpha\mathrm{d}\alpha - 2\iint k\,\mathrm{d}\alpha\mathrm{d}\beta + \text{constant}$$

contains neither a convex self-interaction nor a convex cross term along Wasserstein geodesics.

Internal energies require a different mechanism: convexity is hidden in the Jacobian determinant of the transport map, not in a pointwise convex integrand along particles. The precise criterion is McCann's displacement-convexity theorem [371].

Theorem 15.19 (McCann displacement convexity for internal energies)**.** *Let $g : [0,+\infty) \to \mathbb{R} \cup \{+\infty\}$ be convex with $g(0) = 0$, and define*

$$\psi(r) = r^d g(r^{-d}), \qquad r > 0.$$

If ψ is convex and nonincreasing on $(0,+\infty)$, then the internal energy

$$\mathcal{U}_g(\alpha) = \begin{cases} \displaystyle\int_{\mathbb{R}^d} g(\rho(x))\,\mathrm{d}x, & \alpha = \rho\mathrm{d}x, \\ +\infty, & \textit{otherwise}, \end{cases}$$

is geodesically convex on $\mathcal{P}_2(\mathbb{R}^d)$.

Proof. It is enough to prove the inequality for smooth positive compactly supported densities; the general case follows by approximation and lower semicontinuity. Let $T = \nabla\varphi$ be the Brenier map from $\alpha_0 = \rho_0\mathrm{d}x$ to $\alpha_1 = \rho_1\mathrm{d}x$, set $T_t = (1-t)\mathrm{Id} + tT$, and denote $J_t(x) = \det\big((1-t)\mathrm{Id} + t\nabla T(x)\big)$. The interpolation satisfies $\alpha_t = (T_t)_\sharp\alpha_0$ and $\rho_t(T_t(x))J_t(x) = \rho_0(x)$. By concavity of $\det^{1/d}$ on positive semidefinite matrices,

$$J_t(x)^{1/d} \geqslant (1-t) + t\,\det(\nabla T(x))^{1/d}.$$

Introduce $s_t(x) = (J_t(x)/\rho_0(x))^{1/d}$. Since $\rho_1(T(x))\det\nabla T(x) = \rho_0(x)$, this gives

$$s_t(x) \geqslant (1-t)\rho_0(x)^{-1/d} + t\rho_1(T(x))^{-1/d}.$$

Using the change of variables $y = T_t(x)$, one obtains

$$\int g(\rho_t(y))\,\mathrm{d}y = \int \rho_0(x)\,\psi(s_t(x))\,\mathrm{d}x.$$

Because ψ is nonincreasing and convex,

$$\psi(s_t(x)) \leqslant (1-t)\psi(\rho_0(x)^{-1/d}) + t\psi(\rho_1(T(x))^{-1/d}).$$

Multiplying by ρ_0 and integrating gives $\mathcal{U}_g(\alpha_t) \leqslant (1-t)\mathcal{U}_g(\alpha_0) + t\mathcal{U}_g(\alpha_1)$. □

The theorem applies to $g(s) = s\log s$, since then $\psi(r) = -d\log r$, to porous-medium powers $g(s) = s^m/(m-1)$, $m > 1$, and more generally to the fast-diffusion range $m \geqslant 1 - 1/d$ when the energy is well defined, with $m = 1$ understood as the entropy limit. The same criterion also clarifies which Csiszár divergences inherit displacement convexity from an internal-energy structure.

Proposition 15.20 (Geodesic convexity of power divergences)**.** *For $m > 0$, define*

$$\varphi_m(r) = \begin{cases} \dfrac{r^m - mr + m - 1}{m(m-1)}, & m \neq 1, \\ r\log r - r + 1, & m = 1. \end{cases}$$

Then the following geodesic-convexity statements hold.

1. *If $\Omega \subset \mathbb{R}^d$ is convex, $\beta = \rho_\beta \mathrm{d}x$ is uniform on Ω, and $m \geqslant 1 - 1/d$, then $\alpha \mapsto \mathcal{D}_{\varphi_m}(\alpha|\beta)$ is geodesically convex on $\mathcal{P}_2(\Omega)$.*
2. *If $\beta = Z^{-1}e^{-V}\mathrm{d}x$ on $\mathbb{R}^d$, V is convex, and $m \geqslant 1$, then $\alpha \mapsto \mathcal{D}_{\varphi_m}(\alpha|\beta)$ is geodesically convex on $(\mathcal{P}_2(\mathbb{R}^d), \mathcal{W}_2)$.*
3. *In the KL case $m = 1$, if V is λ-strongly convex, then $\alpha \mapsto \mathrm{KL}(\alpha|\beta) = \mathcal{D}_{\varphi_1}(\alpha|\beta)$ is λ-geodesically convex.*

All statements use the convention that the divergence is $+\infty$ when α is not absolutely continuous with respect to β.

Proof. For the flat reference case, write $\alpha = \rho\mathrm{d}x$ and let $\rho_\beta = b\,\mathbb{1}_\Omega$, where $b = 1/|\Omega|$. Up to affine terms in ρ, which only add constants on probability measures, the integrand $b\,\varphi_m(\rho/b)$ has non-affine part $b^{1-m}\rho^m/(m(m-1))$ for $m \neq 1$, and $\rho\log\rho$ for $m = 1$. For $m = 1$, McCann's criterion gives $\psi(r) = -d\log r$. For $m \neq 1$, the transformed non-affine part is, up to the irrelevant factor b^{1-m}, $\psi_m(r) = \frac{r^{d(1-m)}}{m(m-1)}$. If $m > 1$, this function is convex and nonincreasing. If $0 < m < 1$, it is convex and nonincreasing exactly when $d(1-m) \leqslant 1$, i.e. $m \geqslant 1 - 1/d$. This proves the first claim by Theorem 15.19.

Assume next that $m > 1$ and V is convex. Up to constants,

$$\mathcal{D}_{\varphi_m}(\alpha|\beta) = \frac{Z^{m-1}}{m(m-1)}\int \rho(x)^m e^{(m-1)V(x)}\,\mathrm{d}x.$$

Let $T = \nabla\varphi$ be the Brenier map from $\alpha_0 = \rho_0\mathrm{d}x$ to α_1, set $T_t = (1-t)\mathrm{Id} + tT$, and write $J_t(x) = \det((1-t)\mathrm{Id} + t\nabla T(x))$. Along the geodesic $\alpha_t = (T_t)_\sharp\alpha_0$,

$$\int \rho_t^m e^{(m-1)V}\mathrm{d}x = \int \rho_0(x)^m \exp\big((m-1)\big(V(T_t(x)) - \log J_t(x)\big)\big)\,\mathrm{d}x.$$

For each x, the function $t \mapsto V(T_t(x))$ is convex, and $t \mapsto -\log J_t(x)$ is convex because log det is concave on positive definite matrices. Since $m - 1 > 0$, the exponential of their sum is convex in t. Integrating gives geodesic convexity. The nonsmooth case follows by approximation.

Finally, for $m = 1$,

$$\mathrm{KL}(\alpha|\beta) = \int \rho\log\rho\,\mathrm{d}x + \int V\mathrm{d}\alpha + \log Z.$$

The entropy term is 0-geodesically convex by Theorem 15.19. The linear potential term is λ-geodesically convex when V is λ-strongly convex, by Proposition 15.17. Taking $\lambda = 0$ also covers the $m = 1$ case in the second claim. □

Remark 15.21 (Flat references and general φ-divergences). The Csiszár φ-divergences $\mathcal{D}_\varphi$ of Definition 7.16 behave like internal energies only in special reference geometries. If $\beta = b\,\mathbb{1}_\Omega \mathrm{d}x$ is proportional to Lebesgue measure on a bounded convex domain, then

$$\mathcal{D}_\varphi(\alpha|\beta) = \int_\Omega b\,\varphi(\rho(x)/b)\,\mathrm{d}x, \qquad \alpha = \rho \mathrm{d}x,$$

so McCann's theorem applies whenever the integrand $g(s) = b\,\varphi(s/b)$ satisfies the displacement-convexity condition. For a nonuniform target β, this reduction is no longer available. Proposition 15.20 gives a useful structured extension for the power family: log-concave targets preserve geodesic convexity for $m \geqslant 1$, while the KL endpoint is special in the stronger sense that λ-strong log-concavity gives λ-geodesic convexity. Outside such structured families, a generic φ-divergence is not expected to be geodesically convex.

The Hellinger integrand $\varphi_H(r) = (\sqrt{r} - 1)^2$ illustrates why the qualifier "flat reference" matters. Since $\varphi_H = \varphi_{1/2}/2$, Proposition 15.20 shows that the corresponding flat-reference internal energy is geodesically convex in dimension $d = 1$. The same φ-divergence can nevertheless fail to be geodesically convex once the reference is nonuniform.

A concrete Gaussian test shows the obstruction. Let $\beta = \mathcal{N}(0,1)$ on $\mathbb{R}$ and $\alpha_m = \mathcal{N}(m,1)$. The family $m \mapsto \alpha_m$ is a $\mathcal{W}_2$-geodesic family, since the geodesic between α_{m_0} and α_{m_1} is $\alpha_{(1-t)m_0+tm_1}$. Writing

$$r_m(x) = \frac{\mathrm{d}\alpha_m}{\mathrm{d}\beta}(x) = \exp(mx - m^2/2),$$

one obtains

$$\mathcal{D}_{\varphi_H}(\alpha_m|\beta) = \int (\sqrt{r_m} - 1)^2\,\mathrm{d}\beta = 2\left(1 - e^{-m^2/8}\right).$$

Its second derivative is

$$\frac{\mathrm{d}^2}{\mathrm{d}m^2}\mathcal{D}_{\varphi_H}(\alpha_m|\beta) = \frac{1}{2}e^{-m^2/8}\left(1 - \frac{m^2}{4}\right),$$

which is negative for $|m| > 2$. Thus the Hellinger φ-divergence to a Gaussian target is not even 0-geodesically convex. More generally, Gaussian translations reduce the question for any φ to ordinary convexity of

$$m \mapsto \mathbb{E}_{X\sim\mathcal{N}(0,1)}\left[\varphi\left(e^{mX-m^2/2}\right)\right],$$

which need not hold. The KL integrand is special because this function is $m^2/2$, up to the usual normalization.

Geodesically convex constraints. Geodesic convexity is also useful for hard constraints. If a feasible set is geodesically convex, then adding its indicator to an energy does not destroy the variational structure used by minimizing movements and convergence arguments. The density cap from (15.11) is the model example: it is numerically delicate, because one must compute a projection or a pressure, but it is geometrically compatible with Wasserstein descent.

Proposition 15.22 (Geodesic convexity of density caps)**.** *Assume that $\Omega \subset \mathbb{R}^d$ is convex and let $\rho_{\max} > 0$. The set $\mathcal{K}_{\rho_{\max}}$ in (15.11) is geodesically convex in $\mathcal{P}_2(\Omega)$: if $\alpha_0, \alpha_1 \in \mathcal{K}_{\rho_{\max}}$, then the constant-speed $\mathcal{W}_2$ geodesic $(\alpha_t)_{t\in[0,1]}$ between them also belongs to $\mathcal{K}_{\rho_{\max}}$.*

Proof. Assume first that the endpoints have smooth positive densities $\alpha_i = \rho_i \mathrm{d}x$ with $\rho_i \leqslant \rho_{\max}$, and let $T = \nabla\varphi$ be the Brenier map from α_0 to α_1. Since Ω is convex, $T_t = (1-t)\mathrm{Id} + tT$ maps Ω into Ω. The interpolation is $\alpha_t = (T_t)_\sharp \alpha_0$ and

$$\rho_t(T_t(x)) = \frac{\rho_0(x)}{\det((1-t)\mathrm{Id} + t\nabla T(x))}.$$

The concavity of $\det^{1/d}$ on positive semidefinite matrices and the identity $\rho_1(T(x))\det\nabla T(x) = \rho_0(x)$ give

$$\rho_t(T_t(x))^{-1/d} \geqslant (1-t)\rho_0(x)^{-1/d} + t\rho_1(T(x))^{-1/d} \geqslant \rho_{\max}^{-1/d}.$$

Hence $\rho_t \leqslant \rho_{\max}$ for all t. The general case follows by approximation and lower semicontinuity of the L^∞ bound, as in McCann's displacement-convexity theory [371]. □

Example 15.23 (Squared Wasserstein distance need not be geodesically convex). The squared distance to a fixed measure is not geodesically convex in general. This is a sharp contrast with Hilbert spaces, where $x \mapsto \|x-y\|^2$ is convex along line segments. In $\mathcal{P}_2(\mathbb{R}^d)$, $d \geqslant 2$, take

$$\beta = \frac{1}{2}\delta_{(-1/2,-1/2)} + \frac{1}{2}\delta_{(1/2,1/2)}, \quad \alpha_0 = \frac{1}{2}\delta_{(-1,0)} + \frac{1}{2}\delta_{(1,0)}, \quad \alpha_1 = \frac{1}{2}\delta_{(0,-1)} + \frac{1}{2}\delta_{(0,1)}.$$

One optimal coupling between α_0 and α_1 matches $(-1,0)$ with $(0,1)$ and $(1,0)$ with $(0,-1)$. The associated geodesic has midpoint

$$\alpha_{1/2} = \frac{1}{2}\delta_{(-1/2,1/2)} + \frac{1}{2}\delta_{(1/2,-1/2)}.$$

Then

$$\mathcal{W}_2^2(\alpha_0,\beta) = \mathcal{W}_2^2(\alpha_1,\beta) = \frac{1}{2}, \qquad \mathcal{W}_2^2(\alpha_{1/2},\beta) = 1,$$

which violates geodesic convexity. Thus objectives built from squared Wasserstein distances can be non-convex in the Wasserstein geometry itself. This is one reason why optimal quantization, which minimizes $\mathcal{W}_2^2(\alpha,\nu)$ over m-atomic measures ν, leads to non-convex algorithms such as Lloyd's method; see Section 6.4 and Algorithm 6.5.

Remark 15.24 (Target discrepancies are usually semiconcave). The preceding example is a useful warning. Losses of the form $\alpha \mapsto D(\alpha,\beta)$, where β is a fixed target distribution, are rarely geodesically convex in the transport geometry. This is unfortunate for generative AI and generative modeling more broadly: training a distribution to match data by decreasing $\mathcal{W}_2^2(\alpha,\beta)$, $\mathrm{SW}_2^2(\alpha,\beta)$, or similar discrepancies should not be expected to inherit the convex behavior of Euclidean least squares.
The squared Wasserstein loss has the opposite one-sided curvature: it is 2-geodesically semiconcave along $\mathcal{W}_2$ geodesics. More precisely, if $(\alpha_t)_{t\in[0,1]}$ is a constant-speed $\mathcal{W}_2$ geodesic, then

$$\mathcal{W}_2^2(\alpha_t,\beta) \geqslant (1-t)\,\mathcal{W}_2^2(\alpha_0,\beta) + t\,\mathcal{W}_2^2(\alpha_1,\beta) - t(1-t)\,\mathcal{W}_2^2(\alpha_0,\alpha_1).$$

This is the standard 2-geodesic semiconcavity convention for squared distances, often informally called "2-concavity": the graph may fall below the chord, but only by the quadratic correction $t(1-t)\,\mathcal{W}_2^2(\alpha_0,\alpha_1)$. Equivalently, $-\mathcal{W}_2^2(\cdot,\beta)$ is (-2)-geodesically convex in the convention of Definition 15.16.
To prove the estimate, write $\alpha_t = ((1-t)P_0 + tP_1)_\sharp \pi_{0,1}$, where $\pi_{0,1}$ is optimal between α_0 and α_1. Fix t, put $Z = (1-t)X_0 + tX_1$, take an optimal coupling between α_t and β, and glue it with a disintegration of $\pi_{0,1}$ over Z. This gives random variables (X_0, X_1, Y) such that (Z,Y) is optimal between α_t and β, while (X_0,X_1) is optimal between α_0 and α_1. Since (X_i,Y) are admissible couplings between α_i and β,

$$\mathcal{W}_2^2(\alpha_i,\beta) \leqslant \mathbb{E}\|X_i - Y\|^2, \qquad i = 0,1.$$

The Euclidean identity

$$(1-t)\|X_0-Y\|^2 + t\|X_1-Y\|^2 = \|Z-Y\|^2 + t(1-t)\|X_0-X_1\|^2$$

then gives the claim after taking expectations.
The sliced loss has the same flavor, but one must keep track of which geometry is used. The projection of a $\mathcal{W}_2$ geodesic need not be the monotone one-dimensional geodesic between the projected endpoint measures. Fix a direction θ and project $\pi_{0,1}$ onto the pair $(\langle\theta,\,X_0\rangle, \langle\theta,\,X_1\rangle)$. This projected endpoint coupling generates the projected curve, although it need not be optimal. At time t, couple the projected α_t optimally to the projected β, then glue this coupling with a disintegration of the projected endpoint coupling over $(1-t)\langle\theta,\,X_0\rangle + t\langle\theta,\,X_1\rangle$. The same Euclidean identity therefore applies, with correction cost $\mathbb{E}|\langle\theta,\,X_0-X_1\rangle|^2$. Integrating in θ with the normalization of Definition 11.22 gives, along the original $\mathcal{W}_2$ geodesic,

$$\mathrm{SW}_2^2(\alpha_t,\beta) \geqslant (1-t)\,\mathrm{SW}_2^2(\alpha_0,\beta) + t\,\mathrm{SW}_2^2(\alpha_1,\beta) - \frac{t(1-t)}{d}\,\mathcal{W}_2^2(\alpha_0,\alpha_1),$$

because the projection cost satisfies $\int_{\mathbb{S}^{d-1}} |\langle\theta,\,x_0-x_1\rangle|^2\,\mathrm{d}\sigma(\theta) = \|x_0-x_1\|^2/d$ and $\pi_{0,1}$ is $\mathcal{W}_2$-optimal. Thus this is a semiconcavity estimate for the sliced discrepancy along $\mathcal{W}_2$ geodesics, with a $\mathcal{W}_2$-controlled correction. If one instead works in the idealized Hilbert embedding by projected quantile

functions, straight sliced chords satisfy the exact identity with the last term $\mathrm{SW}_2^2(\alpha_0,\alpha_1)$. The caveat, discussed in the paragraph on intrinsic sliced length in Section 11.2, is that such straight chords need not be realized by actual measures on $\mathbb{R}^d$.

Dissipation and finite metric length. Before asking for convergence rates, one should first extract the basic information already contained in energy dissipation. For the classical gradient flow of a smooth function $h:\mathbb{R}^d\to\mathbb{R}$, $\dot{x}_t=-\nabla h(x_t)$, the chain rule gives

$$\frac{\mathrm{d}}{\mathrm{d}t}h(x_t)=-\|\nabla h(x_t)\|^2=-\|\dot{x}_t\|^2.$$

Thus, for $0\leqslant s<t$, Cauchy's inequality gives

$$\int_s^t\|\dot{x}_r\|\,\mathrm{d}r\leqslant\sqrt{t-s}\left(\int_s^t\|\dot{x}_r\|^2\mathrm{d}r\right)^{1/2}=\sqrt{t-s}\,\big(h(x_s)-h(x_t)\big)^{1/2}. \tag{15.22}$$

The Wasserstein analogue uses the metric derivative $|\dot{\alpha}_t|_{\mathcal{W}_2}$ of an absolutely continuous curve. The following estimate isolates the only ingredient needed at this stage: energy dissipation controls the metric length of the path.

Proposition 15.25 (Finite metric length from dissipation)**.** *Let $(\alpha_r)_{r\in[0,T]}$ be an absolutely continuous curve in $(\mathcal{P}_2(\mathbb{R}^d),\mathcal{W}_2)$. Assume that, for all $0\leqslant s<t\leqslant T$, it satisfies the exact energy-dissipation identity*

$$f(\alpha_s)-f(\alpha_t)=\int_s^t|\dot{\alpha}_r|_{\mathcal{W}_2}^2\mathrm{d}r.$$

Then

$$\mathrm{Length}_{\mathcal{W}_2}\big((\alpha_r)_{r\in[s,t]}\big)=\int_s^t|\dot{\alpha}_r|_{\mathcal{W}_2}\,\mathrm{d}r\leqslant\sqrt{t-s}\,\big(f(\alpha_s)-f(\alpha_t)\big)^{1/2}. \tag{15.23}$$

In the smooth Wasserstein gradient-flow setting with velocity $v_r=-\nabla_{\mathcal{W}}f(\alpha_r)$, the dissipation identity reads equivalently

$$f(\alpha_s)-f(\alpha_t)=\int_s^t|\dot{\alpha}_r|_{\mathcal{W}_2}^2\mathrm{d}r=\int_s^t\int\|v_r(x)\|^2\mathrm{d}\alpha_r(x)\mathrm{d}r.$$

If one assumes only the metric energy-dissipation inequality for a curve of maximal slope, then the same length estimate holds with the right-hand side multiplied by $\sqrt{2}$.

Proof. Cauchy's inequality gives

$$\int_s^t|\dot{\alpha}_r|_{\mathcal{W}_2}\,\mathrm{d}r\leqslant\sqrt{t-s}\left(\int_s^t|\dot{\alpha}_r|_{\mathcal{W}_2}^2\mathrm{d}r\right)^{1/2}.$$

The exact energy-dissipation identity substitutes the last integral by $f(\alpha_s)-f(\alpha_t)$, which proves (15.23). In the metric theory of curves of maximal slope [12, 469], the energy-dissipation inequality gives instead

$$\frac{1}{2}\int_s^t|\dot{\alpha}_r|_{\mathcal{W}_2}^2\mathrm{d}r\leqslant f(\alpha_s)-f(\alpha_t),$$

and the same argument yields the additional factor $\sqrt{2}$. □

Proposition 15.25 is weaker than a convergence rate: it does not identify a minimizer and it gives no decay law. It does, however, rule out escape to infinity in finite metric time whenever the energy is bounded below on finite time intervals. Indeed the trajectory segment remains in the $\mathcal{W}_2$-ball centered at α_s with radius given by (15.23). It also gives a concrete modulus of continuity. If $m_T\leqslant f(\alpha_r)$ for $0\leqslant r\leqslant T$, then for $0\leqslant s<t\leqslant T$,

$$\mathcal{W}_2(\alpha_s,\alpha_t)\leqslant\mathrm{Length}_{\mathcal{W}_2}\big((\alpha_r)_{r\in[s,t]}\big)\leqslant\sqrt{f(\alpha_0)-m_T}\,\sqrt{t-s},$$

with an additional factor $\sqrt{2}$ under the energy-dissipation inequality. Thus an energy-dissipating trajectory is 1/2-Hölder continuous as a curve in Wasserstein space on every time interval where the energy remains bounded from below. Non-coercivity of the objective means that sublevel sets need not be compact, but an energy-dissipating curve still cannot travel an infinite $\mathcal{W}_2$-distance in finite time without an infinite energy drop. The convex estimate below adds a first quantitative conclusion under geodesic convexity; the PL viewpoint then gives sharper coercive rates toward minimizers.

Proposition 15.26 (Energy decay for convex Wasserstein flows)**.** *Assume formally that f is geodesically convex, admits a smooth first variation, and has a minimizer $\alpha^\star$. Let $(\alpha_t)_t$ be a smooth solution of the Wasserstein gradient flow*

$$\partial_t \alpha_t + \operatorname{div}(\alpha_t v_t) = 0, \qquad v_t = -\mathbb{W} f(\alpha_t).$$

Then

$$\frac{\mathrm{d}}{\mathrm{d}t} f(\alpha_t) = -\int \|\mathbb{W} f(\alpha_t)(x)\|^2 \,\mathrm{d}\alpha_t(x) \leqslant 0.$$

If T_t is the optimal map from α_t to $\alpha^\star$, then

$$f(\alpha_t) - f(\alpha^\star) \leqslant -\frac{\mathrm{d}}{\mathrm{d}t} \frac{1}{2} \mathcal{W}_2^2(\alpha_t, \alpha^\star),$$

and consequently

$$f(\alpha_t) - f(\alpha^\star) \leqslant \frac{\mathcal{W}_2^2(\alpha_0, \alpha^\star)}{2t}.$$

Proof. The chain rule and Proposition 15.7 give

$$\frac{\mathrm{d}}{\mathrm{d}t} f(\alpha_t) = \int \langle \mathbb{W} f(\alpha_t)(x),\, v_t(x) \rangle \,\mathrm{d}\alpha_t(x) = -\int \|\mathbb{W} f(\alpha_t)(x)\|^2 \,\mathrm{d}\alpha_t(x).$$

Geodesic convexity along the geodesic $((1-s)\mathrm{Id} + sT_t)_\sharp \alpha_t$ gives

$$f(\alpha^\star) - f(\alpha_t) \geqslant \int \langle \mathbb{W} f(\alpha_t)(x),\, T_t(x) - x \rangle \,\mathrm{d}\alpha_t(x).$$

Since $v_t = -\mathbb{W} f(\alpha_t)$, this reads

$$f(\alpha_t) - f(\alpha^\star) \leqslant \int \langle v_t(x),\, T_t(x) - x \rangle \,\mathrm{d}\alpha_t(x).$$

The standard first-variation formula for the squared Wasserstein distance gives

$$\frac{\mathrm{d}}{\mathrm{d}t} \frac{1}{2} \mathcal{W}_2^2(\alpha_t, \alpha^\star) = \int \langle x - T_t(x),\, v_t(x) \rangle \,\mathrm{d}\alpha_t(x),$$

which proves the differential inequality. Integrating it from 0 to t and using the monotonicity of $s \mapsto f(\alpha_s)$ gives

$$t\big(f(\alpha_t) - f(\alpha^\star)\big) \leqslant \int_0^t \big(f(\alpha_s) - f(\alpha^\star)\big)\,\mathrm{d}s \leqslant \frac{1}{2} \mathcal{W}_2^2(\alpha_0, \alpha^\star).$$

□

Convergence: the Wasserstein–PL viewpoint. In general, analyzing (15.4) is delicate. Geodesic convexity gives the familiar convex-gradient-flow picture, but rates are driven more directly by a first-order coercivity inequality. The relevant quantity is the squared Wasserstein slope

$$|\partial f|^2(\alpha) := \int \|\mathbb{W} f(\alpha)(x)\|^2 \,\mathrm{d}\alpha(x)$$

in the smooth formal setting. Along a smooth gradient flow, this is both the squared metric speed and the energy-dissipation term. The terminology mirrors the finite-dimensional Polyak–Łojasiewicz condition introduced by Polyak [438] and now standard in nonconvex optimization [296]. It is also part of the broader family of Łojasiewicz and Kurdyka–Łojasiewicz gradient inequalities used to prove convergence of dissipative dynamics [21, 270, 181]. The Wasserstein version simply replaces the Euclidean gradient norm by the metric slope associated with $\mathcal{W}_2$.

Definition 15.27 (Wasserstein–Polyak–Łojasiewicz inequality)**.** Let $f_\star := \inf_{\alpha \in \mathcal{P}_2(\mathbb{R}^d)} f(\alpha) > -\infty$. The functional f satisfies the Wasserstein–PL inequality with constant $\kappa > 0$ if

$$|\partial f|^2(\alpha) \geqslant 2\kappa\big(f(\alpha) - f_\star\big) \tag{15.24}$$

for every α in the domain of f. In the smooth Otto notation this reads

$$\int \|\mathbb{W} f(\alpha)\|^2 \, \mathrm{d}\alpha \geqslant 2\kappa\big(f(\alpha) - f_\star\big).$$

Theorem 15.28 (Wasserstein–PL convergence). *Assume that $f_\star > -\infty$, that f satisfies the Wasserstein–PL inequality (15.24) with constant $\kappa > 0$, and that $(\alpha_t)_{t\geqslant 0}$ is a global Wasserstein gradient flow satisfying the full energy-dissipation identity*

$$\frac{\mathrm{d}}{\mathrm{d}t} f(\alpha_t) = -|\dot\alpha_t|^2 = -|\partial f|^2(\alpha_t) \qquad \text{for a.e. } t > 0. \tag{15.25}$$

Set $E(t) \coloneqq f(\alpha_t) - f_\star$. Then, for $0 \leqslant s \leqslant t$,

$$E(t) \leqslant e^{-2\kappa(t-s)} E(s),$$

and the length of the flow segment satisfies

$$\mathrm{Length}_{\mathcal{W}_2}\big((\alpha_r)_{r\in[s,t]}\big) \leqslant \sqrt{\frac{2}{\kappa}}\big(\sqrt{E(s)} - \sqrt{E(t)}\big).$$

If, in addition, f is lower semicontinuous for $\mathcal{W}_2$-convergence, then α_t converges in $\mathcal{W}_2$ to a minimizer $\alpha_\infty \in \operatorname{Argmin} f$, and

$$\mathcal{W}_2(\alpha_t, \alpha_\infty) \leqslant \sqrt{\frac{2}{\kappa}}\sqrt{E(t)} \leqslant \sqrt{\frac{2}{\kappa}}\sqrt{E(0)}\, e^{-\kappa t}.$$

In particular,

$$\operatorname{dist}_{\mathcal{W}_2}(\alpha_t, \operatorname{Argmin} f) \leqslant \sqrt{\frac{2}{\kappa}}\sqrt{E(0)}\, e^{-\kappa t}.$$

Proof. The energy estimate is immediate from (15.25) and (15.24):

$$E'(t) = -|\partial f|^2(\alpha_t) \leqslant -2\kappa E(t)$$

for almost every t. Gronwall's lemma gives $E(t) \leqslant e^{-2\kappa(t-s)}E(s)$.

The distance estimate uses the same two identities in a slightly different way. On the set where $E(r) > 0$, the PL inequality gives

$$|\partial f|(\alpha_r) \geqslant \sqrt{2\kappa E(r)}.$$

Using the full dissipation identity and $|\dot\alpha_r| = |\partial f|(\alpha_r)$,

$$\begin{aligned}
\mathrm{Length}_{\mathcal{W}_2}\big((\alpha_r)_{r\in[s,t]}\big) &= \int_s^t |\dot\alpha_r| \, \mathrm{d}r = \int_s^t |\partial f|(\alpha_r) \, \mathrm{d}r \\
&= \int_s^t \frac{|\partial f|^2(\alpha_r)}{|\partial f|(\alpha_r)} \, \mathrm{d}r \leqslant \frac{1}{\sqrt{2\kappa}} \int_s^t \frac{|\partial f|^2(\alpha_r)}{\sqrt{E(r)}} \, \mathrm{d}r \\
&= \frac{1}{\sqrt{2\kappa}} \int_s^t \frac{-E'(r)}{\sqrt{E(r)}} \, \mathrm{d}r = \sqrt{\frac{2}{\kappa}}\big(\sqrt{E(s)} - \sqrt{E(t)}\big).
\end{aligned}$$

If E vanishes on a subinterval, then the flow is stationary there by (15.25), so the same estimate follows by applying the argument on the part where $E > 0$. Since Wasserstein distance is bounded by metric length, the same upper bound controls $\mathcal{W}_2(\alpha_s, \alpha_t)$.

Letting $t \to +\infty$, the energy estimate gives $E(t) \to 0$, and the length estimate shows that the tail of $(\alpha_t)_t$ is Cauchy. Since $(\mathcal{P}_2(\mathbb{R}^d), \mathcal{W}_2)$ is complete, there is a limit α_∞. Lower semicontinuity yields

$$f(\alpha_\infty) \leqslant \liminf_{t\to+\infty} f(\alpha_t) = f_\star,$$

hence $\alpha_\infty \in \operatorname{Argmin} f$. Sending the endpoint of the length estimate to $+\infty$ with $s = t$ fixed gives

$$\mathcal{W}_2(\alpha_t, \alpha_\infty) \leqslant \sqrt{\frac{2}{\kappa}}\sqrt{E(t)}.$$

Combining this with the exponential energy decay gives the stated distance rate. The distance-to-the-set bound follows because α_∞ is one minimizer. □

The exponential decay of the energy gap is the direct metric-gradient-flow analogue of the classical PL argument. The tail-length estimate used in Theorem 15.28 is a convenient way to turn this energy decay into a $\mathcal{W}_2$-distance-to-Argmin f bound; this exact formulation was communicated to us by Raphaël Barboni. More refined accounts of KL/PL-type inequalities, global convergence and rates for gradient flows and proximal point sequences in general metric spaces are given by Hauer and Mazón [270] and by Dello Schiavo, Maas and Pedrotti [181].

Geodesic strong convexity. Strong geodesic convexity is the cleanest geometric assumption behind the Wasserstein–PL inequality: it turns displacement convexity into a quantitative lower bound on the metric slope, so that the abstract convergence theorem above applies automatically.

Proposition 15.29 (Strong geodesic convexity implies Wasserstein–PL)**.** *Assume that f is λ-geodesically convex with $\lambda > 0$, that $f_\star$ is attained at some $\alpha^\star$, and that the Wasserstein slope is finite at α. Then*

$$|\partial f|^2(\alpha) \geqslant 2\lambda\big(f(\alpha) - f_\star\big).$$

Thus positive Wasserstein strong convexity implies the Wasserstein–PL inequality with the same constant.

Proof. Fix α and a minimizer $\alpha^\star$. Set

$$g = f(\alpha) - f_\star, \qquad r = \mathcal{W}_2(\alpha, \alpha^\star), \qquad a = |\partial f|(\alpha).$$

If $r = 0$, then $\alpha = \alpha^\star$ and there is nothing to prove. Otherwise let $(\alpha_s)_{s\in[0,1]}$ be a constant-speed geodesic from α to $\alpha^\star$. By λ-geodesic convexity,

$$f(\alpha_s) \leqslant (1-s)f(\alpha) + s f_\star - \frac{\lambda}{2}s(1-s)r^2.$$

Therefore

$$f(\alpha) - f(\alpha_s) \geqslant s g + \frac{\lambda}{2}s(1-s)r^2.$$

Dividing by $\mathcal{W}_2(\alpha, \alpha_s) = sr$ and letting $s \downarrow 0$ in the definition of the descending metric slope yields $a \geqslant \frac{g}{r} + \frac{\lambda}{2}r$. Equivalently, $g \leqslant ar - \lambda r^2/2$. Maximizing the right-hand side over $r \geqslant 0$ gives $g \leqslant a^2/(2\lambda)$, which is the desired PL inequality. □

Together, Proposition 15.29 and Theorem 15.28 recover the exponential energy rate $e^{-2\lambda t}$ for positively geodesically convex Wasserstein gradient flows, and also give the distance rate $e^{-\lambda t}$ toward the minimizer selected by the flow. The PL formulation is useful because it separates the first-order energy-dissipation mechanism from the stronger curvature requirement of geodesic convexity.

> **Remark 15.30 (Convergence rates for classical examples).** The examples needed to apply this implication have already been isolated above. Proposition 15.17 gives λ-geodesic convexity for linear energies generated by λ-strongly convex potentials, and for the polynomial and pairwise interaction energies covered there when the corresponding finite-dimensional integrand is strongly convex. Theorem 15.19 gives displacement convexity of internal energies, including entropy-type examples, and Proposition 15.20 adds the important nonflat case $\mathrm{KL}(\cdot|\beta)$ when $\mathrm{d}\beta = Z^{-1}e^{-V}\mathrm{d}x$ and V is strongly convex. Whenever these criteria yield a positive convexity constant, Proposition 15.29 turns it into a Wasserstein–PL inequality, and Theorem 15.28 gives the corresponding linear convergence rates.

Wasserstein Kurdyka–Łojasiewicz inequalities. The PL condition is the quadratic case of a broader slope–energy principle. The classical starting point is the gradient inequality of Łojasiewicz for real-analytic functions [353], later extended by Kurdyka to functions definable in o-minimal structures [321]. Modern nonsmooth optimization uses the same idea as the Kurdyka–Łojasiewicz property to prove convergence of descent algorithms and subgradient flows; see, for instance, Bolte, Daniilidis, Ley and Mazet [74] and Attouch, Bolte, Redont and Soubeyran [21]. For displacement-convex functionals, Bolte and Blanchet [73] develop a family of Łojasiewicz-type functional inequalities in Wasserstein space and relate them to convergence of the associated gradient dynamics. In the metric-space setting relevant here, the Euclidean gradient norm is replaced by the metric slope, as in the general theory of curves of maximal slope and recent KL/PL convergence results for metric gradient flows [270, 181].

The Kurdyka–Łojasiewicz viewpoint replaces the linear relation between the energy gap and the squared slope by a power law. It is useful for degenerate convex landscapes, homogeneous energies, and transport discrepancies whose minimizers form a flat set; in those cases the same energy-dissipation computation still gives convergence, but typically with polynomial rather than exponential rates. We use "KL" in this paragraph for Kurdyka–Łojasiewicz, not for Kullback–Leibler divergence.

Definition 15.31 (Wasserstein Kurdyka–Łojasiewicz inequality)**.** Let $f_\star = \inf f$, let $E(\alpha) = f(\alpha) - f_\star$, and fix $c > 0$, $\theta \in [0, 1)$. We say that f satisfies a power Wasserstein–KL inequality on a set $\mathcal{U} \subset \mathcal{P}_2(\mathbb{R}^d)$ if

$$|\partial f|(\alpha) \geqslant c\, E(\alpha)^\theta, \qquad \alpha \in \mathcal{U}, \quad 0 < E(\alpha) < +\infty.$$

Equivalently, the desingularizing function

$$\psi(s) = \frac{s^{1-\theta}}{c(1-\theta)}$$

satisfies $\psi'(E(\alpha))|\partial f|(\alpha) \geqslant 1$. The Wasserstein–PL inequality (15.24) is the special case $\theta = 1/2$ with $c = \sqrt{2\kappa}$. The regime $\theta > 1/2$ gives sublinear rates, while $\theta = 1/2$ is exactly the exponential PL regime.

Theorem 15.32 (Sublinear convergence under Wasserstein–KL)**.** *Let $(\alpha_t)_{t\geqslant 0}$ be a Wasserstein gradient flow of f satisfying the full energy-dissipation identity (15.25). Assume that $f_\star > -\infty$, that $E(t) = f(\alpha_t) - f_\star$ is finite, and that the trajectory stays in a region where the Wasserstein–KL inequality of Definition 15.31 holds with $\theta \in (1/2, 1)$. If $E(0) > 0$, then*

$$E(t) \leqslant \Big(E(0)^{1-2\theta} + c^2(2\theta - 1)t\Big)^{-1/(2\theta-1)}.$$

Moreover, for every $t \geqslant 0$,

$$\mathrm{Length}_{\mathcal{W}_2}\big((\alpha_s)_{s\geqslant t}\big) \leqslant \frac{E(t)^{1-\theta}}{c(1-\theta)}.$$

If f is lower semicontinuous on $(\mathcal{P}_2(\mathbb{R}^d), \mathcal{W}_2)$, then α_t converges to some $\alpha_\infty \in \operatorname{Argmin} f$ and

$$\mathcal{W}_2(\alpha_t, \alpha_\infty) \leqslant \frac{E(t)^{1-\theta}}{c(1-\theta)} = O\Big(t^{-(1-\theta)/(2\theta-1)}\Big).$$

Proof. If E reaches zero at some time, then the flow is already at a global minimizer. Indeed, the integrated energy-dissipation identity forces both $|\dot\alpha_t|$ and $|\partial f|(\alpha_t)$ to vanish afterwards. We therefore argue on intervals where $E > 0$. For a.e. t,

$$-E'(t) = |\partial f|^2(\alpha_t) = |\dot\alpha_t|^2.$$

The Wasserstein–KL inequality yields $E'(t) \leqslant -c^2 E(t)^{2\theta}$. Since $1 - 2\theta < 0$,

$$\frac{\mathrm{d}}{\mathrm{d}t} E(t)^{1-2\theta} = (1-2\theta)E(t)^{-2\theta}E'(t) \geqslant c^2(2\theta - 1),$$

and integration from 0 to t gives the announced polynomial bound.

For the length estimate, use again $|\dot\alpha_t| = |\partial f|(\alpha_t)$. For $T > t$,

$$\begin{aligned}\int_t^T |\dot\alpha_s|\,\mathrm{d}s &= \int_t^T \frac{|\partial f|^2(\alpha_s)}{|\partial f|(\alpha_s)}\,\mathrm{d}s \leqslant \frac{1}{c}\int_t^T -E'(s)E(s)^{-\theta}\,\mathrm{d}s \\ &= \frac{E(t)^{1-\theta} - E(T)^{1-\theta}}{c(1-\theta)} \leqslant \frac{E(t)^{1-\theta}}{c(1-\theta)}.\end{aligned}$$

Letting $T \to +\infty$ shows that the tail has finite length. Hence $(\alpha_t)_t$ is Cauchy in $(\mathcal{P}_2(\mathbb{R}^d), \mathcal{W}_2)$, which is complete, and converges to some α_∞. Since the energy bound gives $E(t) \to 0$, lower semicontinuity gives $f(\alpha_\infty) = f_\star$. The distance to α_∞ is bounded by the remaining tail length, which proves the last estimate. □

Theorem 15.32 is the Wasserstein analogue of the standard KL-rate argument for curves of maximal slope in metric spaces; see Attouch, Bolte and Svaiter [21], Hauer and Mazón [270], and Dello Schiavo, Maas and Pedrotti [181]. The endpoint $\theta = 1/2$ is the PL theorem above. Exponents $\theta > 1/2$ correspond to flatter landscapes and slower polynomial rates; formally, $\theta < 1/2$ would instead lead to finite-time convergence in the scalar comparison inequality.

Proposition 15.33 (Powers of PL energies are KL). *Let $g \geqslant 0$ satisfy $\inf g = 0$ and the Wasserstein–PL inequality*

$$|\partial g|^2(\alpha) \geqslant 2\kappa g(\alpha)$$

on a region where $g > 0$. Fix $r \geqslant 1$ and set $f(\alpha) = g(\alpha)^r$. Whenever the metric-slope chain rule $|\partial(g^r)|(\alpha) = r g(\alpha)^{r-1}|\partial g|(\alpha)$ holds, f satisfies the Wasserstein–KL inequality

$$|\partial f|(\alpha) \geqslant r\sqrt{2\kappa}\, f(\alpha)^{1-1/(2r)}.$$

Thus f has KL exponent $\theta = 1-1/(2r)$. For $r = 1$ this is PL; for $r > 1$, Theorem 15.32 gives $f(\alpha_t) = O(t^{-r/(r-1)})$ along the f-gradient flow, under its hypotheses.

Proof. By the chain rule and the PL inequality for g,

$$|\partial f| = r g^{r-1}|\partial g| \geqslant r\sqrt{2\kappa}\, g^{r-1/2} = r\sqrt{2\kappa}\, f^{1-1/(2r)}.$$

The exponent and rate follow by substituting $\theta = 1 - 1/(2r)$ into Theorem 15.32. □

Two recurring homogeneous families make the Wasserstein–KL mechanism explicit.

Remark 15.34 (Potential and distance powers). Let V be smooth and bounded below, set $V_\star = \inf V$, and define $f_V(\alpha) = \int V \,\mathrm{d}\alpha$, whose formal slope is $|\partial f_V|^2(\alpha) = \int |\nabla V|^2 \,\mathrm{d}\alpha$. For $\theta \in [1/2, 1)$,

$$|\nabla V(x)| \geqslant c(V(x) - V_\star)^\theta \quad \Longleftrightarrow \quad |\partial f_V|(\alpha) \geqslant c(f_V(\alpha) - V_\star)^\theta$$

for every probability measure α. Jensen proves the forward implication, and testing $\alpha = \delta_x$ proves the converse.
For $V(x) = |x|^q$, $q \geqslant 2$, the moment energy $f_q(\alpha) = \int |x|^q \,\mathrm{d}\alpha$ satisfies

$$|\partial f_q|^2(\alpha) = q^2 \int |x|^{2q-2} \,\mathrm{d}\alpha \geqslant q^2 f_q(\alpha)^{2(q-1)/q}.$$

Thus its KL exponent is $\theta = (q-1)/q$: $q = 2$ is PL, while $q > 2$ gives $f_q(\alpha_t) = O(t^{-q/(q-2)})$. The same calculation applies to distance powers: $d(\cdot, x_\star)^q$ has slope $q d^{q-1}$. Hence $\mathcal{W}_q^q(\cdot, \beta)$ has this exponent in the $\mathcal{W}_q$ geometry, and $\mathcal{W}_2^q(\cdot, \beta)$ does so in the present $\mathcal{W}_2$ geometry.

Remark 15.35 (Homogeneous interaction energies). For $q \geqslant 2$, consider

$$f_q^{\mathrm{int}}(\alpha) = \frac{1}{2} \iint |x - y|^q \,\mathrm{d}\alpha(x)\mathrm{d}\alpha(y).$$

Its minimizers are the Dirac masses. Its Wasserstein gradient is $G_\alpha(x) = q \int |x - y|^{q-2}(x - y) \,\mathrm{d}\alpha(y)$. Writing $\bar{x}_\alpha = \int x \,\mathrm{d}\alpha(x)$, symmetrization gives $\int G_\alpha \cdot (x - \bar{x}_\alpha) \,\mathrm{d}\alpha = q f_q^{\mathrm{int}}(\alpha)$, while Jensen gives $\int |x - \bar{x}_\alpha|^q \,\mathrm{d}\alpha \leqslant 2 f_q^{\mathrm{int}}(\alpha)$. Hölder and Cauchy–Schwarz therefore yield

$$|\partial f_q^{\mathrm{int}}|(\alpha) = \|G_\alpha\|_{L^2(\alpha)} \geqslant q\, 2^{-1/q} f_q^{\mathrm{int}}(\alpha)^{(q-1)/q}.$$

Thus f_q^{int} has KL exponent $(q-1)/q$ relative to the manifold of Dirac minimizers. Again, $q = 2$ is PL and $q > 2$ gives the rate $O(t^{-q/(q-2)})$.

Convexity and curvature. The same language is not restricted to subsets of $\mathbb{R}^d$. If $(\mathcal{X}, d, \mathfrak{m})$ is a geodesic metric-measure space, $\mathcal{W}_2$ geodesics can be defined by transporting each pair of endpoints along metric geodesics, or more intrinsically by dynamical optimal plans on path space, as discussed in Section 4.1. Given a reference measure $\mathfrak{m}$, the entropy relative to $\mathfrak{m}$ is

$$\mathrm{Ent}_{\mathfrak{m}}(\alpha) := \begin{cases} \displaystyle\int_{\mathcal{X}} \rho \log \rho \,\mathrm{d}\mathfrak{m}, & \text{if } \alpha = \rho\,\mathfrak{m}, \\ +\infty, & \text{otherwise.} \end{cases}$$

On a smooth Riemannian manifold (M, g), the Ricci curvature tensor Ric_g is the trace of the Riemann curvature tensor; the lower bound $\mathrm{Ric}_g \geqslant \lambda g$ means that $\mathrm{Ric}_g(v, v) \geqslant \lambda |v|_g^2$ for every tangent vector v.

The fundamental link between curvature and optimal transport is that this tensor lower bound is exactly encoded by geodesic convexity of entropy.

Theorem 15.36 (Ricci curvature and entropy convexity)**.** *Let (M, g) be a smooth compact connected Riemannian manifold without boundary, and let $\mathfrak{m} = \mathrm{vol}_g$. For $\lambda \in \mathbb{R}$, the lower Ricci bound $\mathrm{Ric}_g \geqslant \lambda g$ holds if and only if $\mathrm{Ent}_{\mathfrak{m}}$ is λ-geodesically convex on $(\mathcal{P}_2(M), \mathcal{W}_2)$.*

This equivalence was developed in the smooth Riemannian setting by Cordero-Erausquin, McCann and Schmuckenschläger and by von Renesse and Sturm [156, 543]; it is a central theme of the optimal-transport approach to curvature in Villani's monograph [539]. Lott–Villani and Sturm then used the same entropy-convexity principle to define synthetic lower Ricci curvature bounds on metric-measure spaces [354, 507, 508]. Outside this convex, curvature-controlled regime, such as in the mean-field neural-network example below, the flow may still be informative but its convergence analysis requires problem-specific arguments.

15.3 Functional Inequalities via Optimal Transport

Optimal transport proves inequalities by turning comparison into geometry. One transports a density by a monotone or Brenier map, interpolates along the resulting rays, and converts concavity of a Jacobian determinant or convexity of an energy into an integral estimate. The first two examples below illustrate this geometric mechanism. The remainder concerns entropy, where a functional inequality comparing the energy gap with its squared Wasserstein slope is exactly a Wasserstein–PL inequality and therefore yields convergence rates through Theorem 15.28. We work in the smooth Euclidean setting so that the main calculation remains visible; the standard approximation arguments recover the usual Borel and Sobolev formulations. Systematic accounts include McCann's displacement-convexity principle, the Riemannian interpolation inequalities of Cordero-Erausquin, McCann and Schmuckenschläger, and Villani's treatment of concentration and functional inequalities [371, 156, 539].

Brunn–Minkowski and isoperimetry. The most geometric use of OT is to prove that transporting two uniform measures and looking at the interpolated support increases volume at least concavely. The perimeter inequality is then obtained by differentiating this volume estimate after adding a small ball.

Proposition 15.37 (Brunn–Minkowski and Euclidean isoperimetry)**.** *Let $A, B \subset \mathbb{R}^d$ be bounded Borel sets with positive Lebesgue measure. For $t \in [0,1]$, set*

$$(1-t)A + tB := \{(1-t)x + ty \ : \ x \in A,\ y \in B\}.$$

Then

$$\mathrm{Vol}((1-t)A + tB)^{1/d} \geqslant (1-t)\mathrm{Vol}(A)^{1/d} + t\,\mathrm{Vol}(B)^{1/d}.$$

Consequently, if A is smooth and bounded and $\omega_d = \mathrm{Vol}(B_1)$, then

$$\mathrm{Per}(A) \geqslant d\,\omega_d^{1/d}\mathrm{Vol}(A)^{(d-1)/d}.$$

The constant is sharp, with equality for Euclidean balls.

Proof. We first give the argument for regular sets, the general Borel case following by approximation from inside and outside. Let α and β be the uniform probability measures on A and B, and let $T = \nabla\varphi$ be the Brenier map with $T_\sharp\alpha = \beta$. The interpolating map $T_t = (1-t)\mathrm{Id} + tT$ sends α to a measure α_t supported in $(1-t)A + tB$. At points where T is differentiable, $\det DT_t = \det((1-t)\mathrm{Id} + tDT)$. Since DT is symmetric positive semidefinite and $\det^{1/d}$ is concave on the positive semidefinite cone, $\det(DT_t)^{1/d} \geqslant (1-t) + t\det(DT)^{1/d}$. The change-of-variables identity for the uniform measures gives $\det(DT) = \mathrm{Vol}(B)/\mathrm{Vol}(A)$ almost everywhere. If ρ_t denotes the density of α_t, a second application of the same formula gives

$$\rho_t(T_t(x))^{-1/d} = \mathrm{Vol}(A)^{1/d}\det(DT_t(x))^{1/d} \geqslant c_t,$$

where $c_t = (1-t)\mathrm{Vol}(A)^{1/d} + t\mathrm{Vol}(B)^{1/d}$. Thus $\rho_t \leqslant c_t^{-d}$ almost everywhere. Since α_t has total mass one and is supported in $(1-t)A + tB$, this gives $\mathrm{Vol}((1-t)A + tB) \geqslant c_t^d$, which is the displayed Brunn–Minkowski inequality.

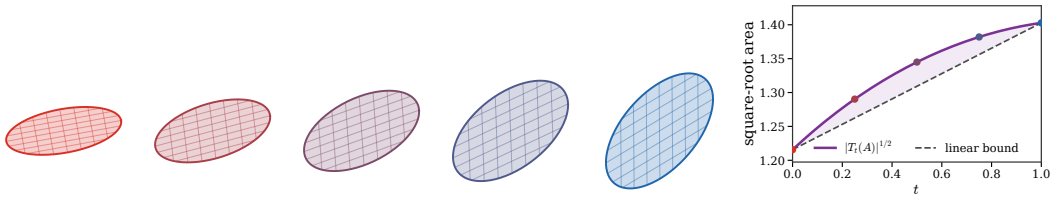


Figure 15.10: Brunn–Minkowski through an affine optimal-transport interpolation between two ellipses. For ellipses, the Brenier map is affine, so the transported support $T_t(A)$ remains an ellipse and its area is computed exactly from the determinant of $(1-t)\mathrm{Id}+tDT$. The right panel shows that $|T_t(A)|^{1/2}$ lies above the linear interpolation of the endpoint square-root areas, which is the two-dimensional determinant concavity behind the proof of Brunn–Minkowski.

For isoperimetry, use the homogeneity of Brunn–Minkowski and apply it to A and εB_1. Writing $A+\varepsilon B_1$ for the parallel set gives

$$\mathrm{Vol}(A+\varepsilon B_1)^{1/d} \geqslant \mathrm{Vol}(A)^{1/d} + \varepsilon \omega_d^{1/d}.$$

Using $\mathrm{Vol}(A+\varepsilon B_1) = \mathrm{Vol}(A) + \varepsilon\,\mathrm{Per}(A) + o(\varepsilon)$ and differentiating at $\varepsilon = 0$ yields the claimed perimeter bound. Balls attain equality, so the constant is optimal. □

Figure 15.10 isolates the determinant-concavity step in the special case where the optimal map is affine.

Prékopa–Leindler. The functional analogue of Brunn–Minkowski replaces sets by densities. The transport proof is the same argument with the Jacobian no longer constant.

Proposition 15.38 (Prékopa–Leindler inequality)**.** *Let $u, v, w : \mathbb{R}^d \to [0,+\infty)$ be integrable, and let $t \in [0,1]$. Assume that*

$$w((1-t)x+ty) \geqslant u(x)^{1-t}v(y)^t \qquad \text{for all } x, y \in \mathbb{R}^d.$$

Then

$$\int_{\mathbb{R}^d} w \geqslant \left(\int_{\mathbb{R}^d} u\right)^{1-t}\left(\int_{\mathbb{R}^d} v\right)^t.$$

Proof. The endpoint cases are immediate, so fix $t \in (0,1)$. Set $U = \int u$ and $V = \int v$. If $U = 0$ or $V = 0$, the conclusion is trivial; hence assume $U, V > 0$. Let $\alpha = (u/U)\mathrm{d}x$ and $\beta = (v/V)\mathrm{d}x$, and let T be the Brenier map from α to β. Put $T_t = (1-t)\mathrm{Id} + tT$. On a smooth positive regularization, T_t is the gradient of a strictly convex function and is therefore one-to-one almost everywhere. The area formula gives $\int w \geqslant \int w(T_t(x))\det(DT_t(x))\mathrm{d}x$. For each eigenvalue $s \geqslant 0$ of DT, weighted arithmetic–geometric mean gives $(1-t)+ts \geqslant s^t$. Hence $\det((1-t)\mathrm{Id}+tDT) \geqslant \det(DT)^t$. Combining this determinant estimate with the assumption on w yields

$$\int w \geqslant \int u(x)^{1-t}v(T(x))^t \det(DT(x))^t \mathrm{d}x.$$

Since $T_\sharp\alpha = \beta$, the Monge–Ampère identity gives $\det(DT(x)) = \frac{u(x)V}{v(T(x))U}$ almost everywhere. Substitution yields

$$\int w \geqslant \left(\frac{V}{U}\right)^t \int u(x)\mathrm{d}x = U^{1-t}V^t.$$

Approximation removes the smoothness and positivity assumptions. □

Logarithmic Sobolev inequalities. The logarithmic Sobolev inequality, introduced by Gross [256], turns entropy into a quantitative relaxation principle by controlling it with its dissipation. The two quantities are fixed by the following definition.

Definition 15.39 (Relative Fisher information and logarithmic Sobolev inequality)**.** Let $\beta \in \mathcal{P}(\mathbb{R}^d)$

have a positive density. For a probability measure $\alpha = h\beta$, define

$$\mathcal{I}(\alpha|\beta) := 4\int \|\nabla\sqrt{h}\|^2 \,\mathrm{d}\beta = \int \|\nabla \log h\|^2 \,\mathrm{d}\alpha, \tag{15.26}$$

whenever $\sqrt{h}$ has a weak gradient and the integral is finite, and set $\mathcal{I}(\alpha|\beta) = +\infty$ otherwise, including when $\alpha \not\ll \beta$. The measure β satisfies a logarithmic Sobolev inequality with constant $\lambda > 0$ if

$$\mathrm{KL}(\alpha|\beta) \leqslant \frac{1}{2\lambda}\mathcal{I}(\alpha|\beta) \qquad \text{for every probability measure } \alpha. \tag{15.27}$$

This is a coercivity property of the reference measure, not a universal inequality. A standard sufficient condition is positive curvature of its potential, in the sense of the Bakry–Emery criterion [27, 28]. The transport proof below uses the HWI inequality stated later in this section.

Proposition 15.40 (Curvature implies logarithmic Sobolev)**.** *Let $\beta = Z^{-1}e^{-V}\mathrm{d}x$ be a smooth probability measure on $\mathbb{R}^d$. If $\nabla^2 V \geq \lambda \mathrm{Id}$ for some $\lambda > 0$, then β satisfies (15.27) with constant λ.*

Proof. Apply the HWI inequality of Proposition 15.50. Under the curvature assumption, for every absolutely continuous α,

$$\mathrm{KL}(\alpha|\beta) \leqslant \mathcal{W}_2(\alpha,\beta)\sqrt{\mathcal{I}(\alpha|\beta)} - \frac{\lambda}{2}\mathcal{W}_2^2(\alpha,\beta).$$

For fixed $s = \sqrt{\mathcal{I}(\alpha|\beta)}$, maximize the right-hand side over $r = \mathcal{W}_2(\alpha,\beta) \geqslant 0$. The elementary inequality $rs - \lambda r^2/2 \leqslant s^2/(2\lambda)$ gives (15.27). Approximation yields the standard nonsmooth statement. This is the Otto–Villani transport route from the Bakry–Emery curvature bound to logarithmic Sobolev; the original Γ_2 calculus gives the same criterion directly [27, 412, 28]. □

Remark 15.41 (Logarithmic Sobolev as Wasserstein–PL). Under the usual smoothness assumptions, for $f(\alpha) = \mathrm{KL}(\alpha|\beta)$ the Wasserstein gradient is

$$\mathbb{W}f(\alpha) = \nabla \log \frac{\mathrm{d}\alpha}{\mathrm{d}\beta},$$

and its squared Wasserstein slope is precisely the relative Fisher information,

$$|\partial f|^2(\alpha) = \mathcal{I}(\alpha|\beta).$$

Since the unique minimizer is β and $f(\beta) = 0$, (15.27) is exactly

$$|\partial f|^2(\alpha) \geqslant 2\lambda\big(f(\alpha) - f(\beta)\big),$$

that is, the Wasserstein–PL inequality of Definition 15.27. Thus logarithmic Sobolev is the functional-inequality incarnation of the convergence mechanism studied in Section 15.2, and Theorem 15.28 converts it into exponential decay of both entropy and Wasserstein distance to equilibrium.

Proposition 15.42 (Logarithmic Sobolev inequality implies entropy decay)**.** *Let $\beta = e^{-V}\mathrm{d}x/Z$ be a smooth probability measure on $\mathbb{R}^d$ satisfying the logarithmic Sobolev inequality (15.27) with constant $\lambda > 0$. Let α_t solve the Wasserstein gradient flow of $f(\alpha) = \mathrm{KL}(\alpha|\beta)$, namely*

$$\partial_t \rho_t = \nabla\cdot(\rho_t \nabla V) + \Delta\rho_t, \qquad \alpha_t = \rho_t \mathrm{d}x.$$

If $\mathrm{KL}(\alpha_0|\beta) < +\infty$, then

$$\mathrm{KL}(\alpha_t|\beta) \leqslant e^{-2\lambda t}\,\mathrm{KL}(\alpha_0|\beta).$$

Moreover, if $\alpha_0, \beta \in \mathcal{P}_2(\mathbb{R}^d)$, then

$$\mathcal{W}_2(\alpha_t,\beta) \leqslant \sqrt{\frac{2}{\lambda}\,\mathrm{KL}(\alpha_0|\beta)}\, e^{-\lambda t}.$$

Proof. The first variation of f is $\log(\rho/\rho_\beta)$, up to an additive constant. Along the smooth Wasserstein gradient flow,

$$\frac{\mathrm{d}}{\mathrm{d}t}\,\mathrm{KL}(\alpha_t|\beta) = -\int \|\nabla \log \frac{\mathrm{d}\alpha_t}{\mathrm{d}\beta}\|^2 \,\mathrm{d}\alpha_t = -\mathcal{I}(\alpha_t|\beta).$$

The logarithmic Sobolev inequality bounds the Fisher information below by $2\lambda \operatorname{KL}(\alpha_t|\beta)$. Hence

$$\frac{\mathrm{d}}{\mathrm{d}t}\operatorname{KL}(\alpha_t|\beta) \leqslant -2\lambda \operatorname{KL}(\alpha_t|\beta),$$

and Gronwall's lemma gives the entropy estimate. The Wasserstein estimate follows from the length estimate in the Wasserstein–PL theorem: apply Theorem 15.28 to $f = \operatorname{KL}(\cdot|\beta)$, whose unique minimizer is β, and use Remark 15.41. □

Gaussian logarithmic Sobolev inequality. The standard Gaussian is the model case where the optimal constant is known exactly. With the convention of (15.27), $\gamma_d = \mathcal{N}(0, \mathrm{Id})$ satisfies the logarithmic Sobolev inequality with $\lambda = 1$, and this value is sharp. The following transport proof, due to Cordero-Erausquin [154], bounds the Jacobian equation for the Brenier map by $\log\det A \leqslant \operatorname{tr}(A - \mathrm{Id})$, then converts the trace term into Fisher information by Gaussian integration by parts.

Proposition 15.43 (Gaussian logarithmic Sobolev inequality)**.** *Let γ_d be the standard Gaussian measure on $\mathbb{R}^d$. If $\rho > 0$ is smooth, has sufficient decay, and $\int \rho \,\mathrm{d}\gamma_d = 1$, then*

$$\operatorname{Ent}_{\gamma_d}(\rho) \coloneqq \int \rho \log \rho \,\mathrm{d}\gamma_d \leqslant \frac{1}{2}\int \frac{\|\nabla \rho\|^2}{\rho}\,\mathrm{d}\gamma_d.$$

Proof. Let $\alpha = \rho\gamma_d$, and let $T = \nabla\varphi$ be the Brenier map with $T_\sharp \alpha = \gamma_d$. Write $T(x) = x + \xi(x)$. The Monge–Ampère equation reads

$$\rho(x) e^{-\|x\|^2/2} = e^{-\|T(x)\|^2/2}\det(DT(x)).$$

Thus

$$\log \rho(x) = -\langle x,\, \xi(x)\rangle - \frac{1}{2}\|\xi(x)\|^2 + \log\det(\mathrm{Id} + D\xi(x)).$$

Since DT is positive semidefinite, $\log\det(\mathrm{Id} + D\xi) \leqslant \operatorname{tr}(D\xi) = \operatorname{div}\xi$. Multiplying by ρ and integrating with respect to γ_d gives

$$\operatorname{Ent}_{\gamma_d}(\rho) \leqslant \int \rho(\operatorname{div}\xi - \langle x,\, \xi\rangle)\,\mathrm{d}\gamma_d - \frac{1}{2}\int \rho\|\xi\|^2\,\mathrm{d}\gamma_d.$$

Gaussian integration by parts gives

$$\int \rho(\operatorname{div}\xi - \langle x,\, \xi\rangle)\,\mathrm{d}\gamma_d = -\int \langle \nabla\rho,\, \xi\rangle\,\mathrm{d}\gamma_d.$$

Completing the square pointwise,

$$-\langle \nabla\rho,\, \xi\rangle - \frac{1}{2}\rho\|\xi\|^2 \leqslant \frac{1}{2}\frac{\|\nabla\rho\|^2}{\rho}.$$

This proves the inequality. The general Sobolev-class statement follows by approximation. □

The constant cannot be improved. For a translation tilt

$$\rho_a(x) = \exp\!\left(\langle a,\, x\rangle - \frac{1}{2}\|a\|^2\right), \qquad a \in \mathbb{R}^d,$$

one has $\rho_a\gamma_d = \mathcal{N}(a, \mathrm{Id})$. A direct computation gives

$$\int \rho_a \log\rho_a\,\mathrm{d}\gamma_d = \frac{1}{2}\|a\|^2, \qquad \frac{1}{2}\int \frac{\|\nabla\rho_a\|^2}{\rho_a}\,\mathrm{d}\gamma_d = \frac{1}{2}\|a\|^2.$$

Thus equality is attained along translations, and the sharp logarithmic Sobolev constant of the standard Gaussian is $\lambda = 1$. Proposition 15.42 then gives the classical exponential convergence of the Ornstein–Uhlenbeck flow.

Poincaré as a linearized Wasserstein–PL inequality. Many quadratic functional inequalities are the infinitesimal form of nonlinear slope inequalities near equilibrium. Assume that β minimizes an energy f, and let ξ be a smooth bounded density modulation with $\int \xi \, \mathrm{d}\beta = 0$. For $|\varepsilon|$ small enough, set

$$\beta_\varepsilon = (1 + \varepsilon\xi)\beta, \qquad \beta_0 = \beta.$$

Define the second-order energy and dissipation forms of f at β in the direction ξ by

$$\mathsf{H}^f_\beta(\xi) := \lim_{\varepsilon\to 0} \frac{2(f(\beta_\varepsilon) - f(\beta))}{\varepsilon^2}, \qquad \mathsf{D}^f_\beta(\xi) := \lim_{\varepsilon\to 0} \frac{|\partial f|^2(\beta_\varepsilon)}{\varepsilon^2},$$

whenever these finite limits exist. The first is the quadratic part of the energy gap and the second is the quadratic part of the squared Wasserstein slope. Their comparison is the linearized Wasserstein–PL inequality.

Proposition 15.44 (Linearizing a Wasserstein–PL inequality)**.** *Let β be a minimizer of an energy f, and assume that f satisfies*

$$|\partial f|^2(\alpha) \geqslant 2\kappa\big(f(\alpha) - f(\beta)\big)$$

in a neighborhood of β. Let ξ be a smooth bounded perturbation with $\int \xi \, \mathrm{d}\beta = 0$, and set $\beta_\varepsilon = (1 + \varepsilon\xi)\beta$. If $\mathsf{H}^f_\beta(\xi)$ and $\mathsf{D}^f_\beta(\xi)$ exist, then

$$\kappa \, \mathsf{H}^f_\beta(\xi) \leqslant \mathsf{D}^f_\beta(\xi).$$

Proof. Apply the PL inequality to β_ε:

$$|\partial f|^2(\beta_\varepsilon) \geqslant 2\kappa\big(f(\beta_\varepsilon) - f(\beta)\big).$$

Divide by ε^2 and let $\varepsilon \to 0$. The definitions of H^f_β and D^f_β give the claim. □

The two forms above can be evaluated explicitly under the indicated smoothness and integrability assumptions. Density divergences provide the local model.

Example 15.45 (Linearized φ-divergences). Suppose that β has a smooth positive density and that φ is twice continuously differentiable near 1, with $\varphi(1) = 0$ and $\varphi''(1) > 0$. For the Csiszár divergence $f(\alpha) = \mathcal{D}_\varphi(\alpha|\beta)$, one has

$$\mathsf{H}^f_\beta(\xi) = \varphi''(1) \int \xi^2 \, \mathrm{d}\beta, \qquad \mathsf{D}^f_\beta(\xi) = \varphi''(1)^2 \int \|\nabla\xi\|^2 \, \mathrm{d}\beta.$$

The affine part of φ is immaterial on probability measures, so only the curvature $\varphi''(1)$ enters the linearization. For KL, $\varphi(r) = r\log r - r + 1$, hence $\varphi''(1) = 1$, and the linearized PL inequality becomes the Poincaré inequality.

Kernel discrepancies instead produce nonlocal quadratic forms.

Example 15.46 (Linearized maximum mean discrepancy). For $f(\alpha) = \mathrm{MMD}^2_k(\alpha, \beta)$, use the convention

$$\mathrm{MMD}^2_k(\alpha, \beta) = \iint k \, \mathrm{d}(\alpha - \beta)\mathrm{d}(\alpha - \beta).$$

Assume that k is symmetric, differentiable in its first variable, and positive semidefinite on signed measures of zero total mass, and that the displayed integrals are finite. Then

$$\mathsf{H}^f_\beta(\xi) = 2 \iint k(x, y)\xi(x)\xi(y) \, \mathrm{d}\beta(x)\mathrm{d}\beta(y),$$

whereas

$$\mathsf{D}^f_\beta(\xi) = 4 \int \Big\| \int \nabla_x k(x, y)\xi(y) \, \mathrm{d}\beta(y)\Big\|^2 \mathrm{d}\beta(x).$$

Thus the energy Hessian is the kernel quadratic form, whereas the slope form differentiates the kernel witness spatially.

The transport discrepancy reveals the negative Sobolev geometry obtained by linearizing $\mathcal{W}_2$.

Example 15.47 (Linearized squared Wasserstein distance). For $f(\alpha) = \mathcal{W}_2^2(\alpha, \beta)$, the calculation is naturally written in the tangent space of $\mathcal{W}_2$. If $\beta = \rho_\beta \mathrm{d}x$ is smooth and positive and the weighted Poisson problem

$$-\nabla \cdot (\rho_\beta \nabla u_\xi) = \xi \rho_\beta$$

has a finite-energy solution u_ξ, then

$$\mathsf{H}_\beta^f(\xi) = 2 \int \|\nabla u_\xi\|^2 \, \mathrm{d}\beta, \qquad \mathsf{D}_\beta^f(\xi) = 4 \int \|\nabla u_\xi\|^2 \, \mathrm{d}\beta.$$

The quadratic form $\int \|\nabla u_\xi\|^2 \mathrm{d}\beta$ is the negative Sobolev norm induced by the linearized Wasserstein metric.

For $f(\alpha) = \mathrm{KL}(\alpha|\beta)$, Proposition 15.44 says precisely that logarithmic Sobolev linearizes to Poincaré. Indeed, for $h_\varepsilon = 1 + \varepsilon\xi$,

$$\mathrm{KL}(h_\varepsilon \beta|\beta) = \frac{\varepsilon^2}{2} \int \xi^2 \, \mathrm{d}\beta + o(\varepsilon^2), \qquad \mathcal{I}(h_\varepsilon \beta|\beta) = \varepsilon^2 \int \|\nabla \xi\|^2 \, \mathrm{d}\beta + o(\varepsilon^2).$$

Proposition 15.48 (Poincaré from logarithmic Sobolev). *Let β be a smooth probability measure on $\mathbb{R}^d$. Assume that β satisfies the logarithmic Sobolev inequality with constant $\lambda > 0$,*

$$\mathrm{KL}(\alpha|\beta) \leqslant \frac{1}{2\lambda} \mathcal{I}(\alpha|\beta), \qquad \mathcal{I}(\alpha|\beta) \coloneqq \int \|\nabla \log \frac{\mathrm{d}\alpha}{\mathrm{d}\beta}\|^2 \, \mathrm{d}\alpha.$$

Then, for every smooth ξ with $\int \xi \, \mathrm{d}\beta = 0$,

$$\lambda \int \xi^2 \, \mathrm{d}\beta \leqslant \int \|\nabla \xi\|^2 \, \mathrm{d}\beta.$$

In particular, if $\mathrm{d}\beta = Z^{-1} e^{-V} \mathrm{d}x$ and $\nabla^2 V \geq \lambda \mathrm{Id}$, so that β is λ-strongly log-concave, then the conclusion holds by the Bakry–Emery criterion [27, 28].

Proof. For $h_\varepsilon = 1 + \varepsilon\xi$, with $|\varepsilon|$ small enough, one has $\beta_\varepsilon = h_\varepsilon \beta$ and

$$\mathrm{KL}(\beta_\varepsilon|\beta) = \int h_\varepsilon \log h_\varepsilon \, \mathrm{d}\beta = \frac{\varepsilon^2}{2} \int \xi^2 \, \mathrm{d}\beta + o(\varepsilon^2),$$

because $\int \xi \, \mathrm{d}\beta = 0$. The Fisher information expands as

$$\mathcal{I}(\beta_\varepsilon|\beta) = \int \|\nabla \log h_\varepsilon\|^2 h_\varepsilon \, \mathrm{d}\beta = \varepsilon^2 \int \|\nabla \xi\|^2 \, \mathrm{d}\beta + o(\varepsilon^2).$$

Inserting these expansions in the logarithmic Sobolev inequality and letting $\varepsilon \to 0$ gives the Poincaré inequality. The last assertion is the usual Bakry–Emery implication from λ-convexity of V to the logarithmic Sobolev inequality. □

If $\mathrm{d}\beta = \rho_\beta \mathrm{d}x = Z^{-1} e^{-V} \mathrm{d}x$, the Poincaré inequality is equivalently a spectral gap for the β-weighted Laplacian

$$L_\beta \xi \coloneqq \Delta \xi - \langle \nabla V, \nabla \xi \rangle = \rho_\beta^{-1} \nabla \cdot (\rho_\beta \nabla \xi).$$

Indeed L_β is symmetric in $L^2(\beta)$ and

$$\int \|\nabla \xi\|^2 \, \mathrm{d}\beta = \langle \xi, -L_\beta \xi \rangle_{L^2(\beta)}.$$

Thus Poincaré says that the bottom of the spectrum of $-L_\beta$ on mean-zero functions is at least λ. When the resolvent is compact, this is the first nonzero eigenvalue; on a noncompact space the spectral gap need not be attained by an eigenfunction. This interpretation, classical for reversible Markov diffusion generators [331, 28], is the infinitesimal shadow of the nonlinear Wasserstein–PL/log-Sobolev inequality.

This interpretation can be made exact by freezing the Wasserstein mobility at equilibrium. For a mean-zero density perturbation $\dot h$, define

$$\|\dot h\|_{-1,\beta}^2 := \inf_v \left\{ \int \|v\|^2 \, \mathrm{d}\beta \ : \ \dot h = -\rho_\beta^{-1} \nabla \cdot (\rho_\beta v) \right\}.$$

For $E_2(h) = \frac{1}{2}\chi^2(h\beta|\beta) = \frac{1}{2}\int (h-1)^2 \mathrm{d}\beta$, integration by parts shows that its squared slope in this $H^{-1}(\beta)$ geometry is $\int \|\nabla h\|^2 \mathrm{d}\beta$. Poincaré is therefore exactly the PL inequality for E_2 in the linearized Wasserstein geometry. It is not the full $\mathcal{W}_2$-PL inequality for the same energy: the nonlinear Wasserstein mobility is $h\beta$, and the corresponding squared slope is $\int h\|\nabla h\|^2 \mathrm{d}\beta$.

Talagrand's transport-entropy inequality. The distance estimate in Theorem 15.28 already reveals the general principle: when $\mathrm{KL}(\cdot|\beta)$ satisfies logarithmic Sobolev with constant λ, the entropy flow yields the transport-entropy bound $\mathcal{W}_2^2(\alpha,\beta) \leqslant 2\,\mathrm{KL}(\alpha|\beta)/\lambda$. This is the Otto–Villani implication from logarithmic Sobolev to Talagrand's T_2 inequality [412]. For the standard Gaussian, the same Jacobian estimate used above gives a direct sharp proof of Talagrand's original result [514, 154, 539].

Proposition 15.49 (Gaussian T_2 inequality)**.** *For every $\alpha \in \mathcal{P}_2(\mathbb{R}^d)$,*

$$\frac{1}{2}\mathcal{W}_2^2(\alpha,\gamma_d) \leqslant \mathrm{KL}(\alpha|\gamma_d),$$

with the convention that the right-hand side is $+\infty$ when $\alpha \not\ll \gamma_d$.

Proof. If $\mathrm{KL}(\alpha|\gamma_d) = +\infty$, the claim is immediate. Assume first that $\alpha = \rho\gamma_d$ with ρ smooth and positive. Let $T = \nabla\varphi$ be the Brenier map with $T_\sharp\gamma_d = \alpha$, and write $T(x) = x + \xi(x)$. The change of variables gives

$$\rho(T(x))e^{-\|T(x)\|^2/2}\det(DT(x)) = e^{-\|x\|^2/2}.$$

Hence

$$\log\rho(T(x)) = \langle x,\, \xi(x)\rangle + \frac{1}{2}\|\xi(x)\|^2 - \log\det(\mathrm{Id} + D\xi(x)).$$

Using again $\log\det(\mathrm{Id}+D\xi) \leqslant \operatorname{div}\xi$ and integrating with respect to γ_d,

$$\mathrm{KL}(\alpha|\gamma_d) \geqslant \frac{1}{2}\int \|\xi\|^2 \,\mathrm{d}\gamma_d + \int (\langle x,\,\xi\rangle - \operatorname{div}\xi)\,\mathrm{d}\gamma_d.$$

The last integral vanishes by Gaussian integration by parts. Since T is optimal,

$$\int \|\xi\|^2 \,\mathrm{d}\gamma_d = \mathcal{W}_2^2(\gamma_d,\alpha),$$

which proves the claim. Approximation gives the general case. Equality holds for translations $\alpha = \mathcal{N}(a,\mathrm{Id})$, so the constant is sharp. □

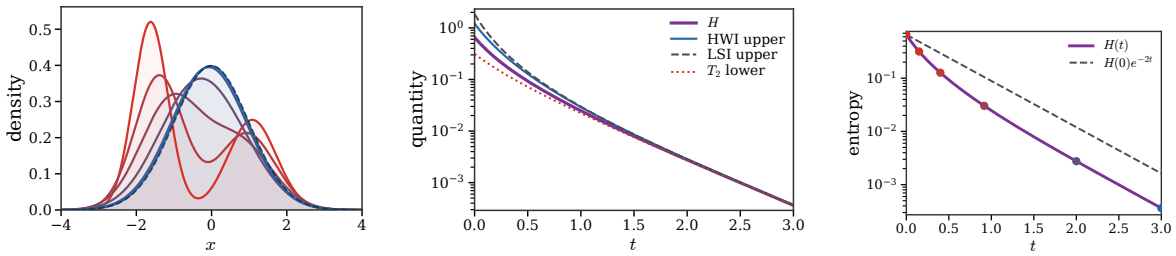
The HWI bridge. The Gaussian logarithmic Sobolev and transport-entropy inequalities are not isolated facts. The HWI inequality of Otto and Villani [412] compares three quantities at once: the entropy H, the Wasserstein distance W, and the Fisher information I. It is one of the cleanest examples where displacement convexity turns into a functional inequality.

Proposition 15.50 (Otto–Villani HWI inequality)**.** *Let $\beta = \rho_\beta \mathrm{d}x = e^{-V}\mathrm{d}x/Z$ be a smooth probability measure in $\mathcal{P}_2(\mathbb{R}^d)$, and assume that $\nabla^2 V \geq \lambda\mathrm{Id}$ for some $\lambda \in \mathbb{R}$. For $\alpha \in \mathcal{P}_2(\mathbb{R}^d)$, define the relative Fisher information by*

$$\mathcal{I}(\alpha|\beta) := \int \|\nabla \log\frac{\rho_\alpha}{\rho_\beta}\|^2 \,\mathrm{d}\alpha,$$

when $\alpha = \rho_\alpha \mathrm{d}x$ and the expression is finite, and set $\mathcal{I}(\alpha|\beta) = +\infty$ otherwise. Then

$$\mathrm{KL}(\alpha|\beta) \leqslant \mathcal{W}_2(\alpha,\beta)\sqrt{\mathcal{I}(\alpha|\beta)} - \frac{\lambda}{2}\mathcal{W}_2^2(\alpha,\beta).$$



Ornstein–Uhlenbeck densities static inequalities entropy decay

Figure 15.11: Functional inequalities along the Ornstein–Uhlenbeck flow from a one-dimensional Gaussian mixture to the standard Gaussian. The left panel shows the density relaxation, with the Gaussian target as a dashed curve. The middle panel compares $H = \mathrm{KL}(\alpha_t|\gamma_1)$ with the HWI upper bound $W\sqrt{I} - W^2/2$, the logarithmic Sobolev upper bound $I/2$, and the Talagrand lower bound $W^2/2$, where $W = \mathcal{W}_2(\alpha_t, \gamma_1)$ and $I = \mathcal{I}(\alpha_t|\gamma_1)$. The right panel displays the dynamic consequence of logarithmic Sobolev, namely $H(t) \leqslant H(0)e^{-2t}$.

Proof. If $\mathcal{I}(\alpha|\beta) = +\infty$, there is nothing to prove. Assume first that ρ_α is smooth and positive. Let T be the Brenier map from α to β, set $v = T - \mathrm{Id}$, and let $\alpha_t = ((1-t)\mathrm{Id} + tT)_\sharp\alpha$. By Theorem 15.19 and Proposition 15.17, $f(\eta) = \mathrm{KL}(\eta|\beta)$ is λ-geodesically convex. Thus the one-dimensional function $e(t) = f(\alpha_t)$ satisfies

$$e(1) \geqslant e(0) + e'(0) + \frac{\lambda}{2}\mathcal{W}_2^2(\alpha,\beta).$$

Since $e(1) = f(\beta) = 0$, it remains to compute the initial derivative. The continuity equation for the geodesic gives $\partial_t\alpha_t + \nabla\cdot(\alpha_t v_t) = 0$ and $v_0 = v$. Hence

$$e'(0) = \int \langle \nabla \log\frac{\rho_\alpha}{\rho_\beta}, v\rangle \,\mathrm{d}\alpha.$$

Therefore

$$\mathrm{KL}(\alpha|\beta) \leqslant -\int \langle \nabla\log\frac{\rho_\alpha}{\rho_\beta}, T-\mathrm{Id}\rangle\,\mathrm{d}\alpha - \frac{\lambda}{2}\mathcal{W}_2^2(\alpha,\beta).$$

Cauchy–Schwarz and $\int \|T-\mathrm{Id}\|^2\mathrm{d}\alpha = \mathcal{W}_2^2(\alpha,\beta)$ give the announced bound. The general case follows by standard approximation and lower semicontinuity of entropy, Fisher information, and $\mathcal{W}_2$. □

When $\lambda > 0$, the HWI inequality is a bridge from curvature to PL. It is not itself a PL inequality because it still contains the distance term $\mathcal{W}_2(\alpha,\beta)$. Optimizing the right-hand side, or simply using Young's inequality $rs - \lambda r^2/2 \leqslant s^2/(2\lambda)$, gives the logarithmic Sobolev estimate

$$\mathrm{KL}(\alpha|\beta) \leqslant \frac{1}{2\lambda}\mathcal{I}(\alpha|\beta).$$

Since $\mathcal{I}(\cdot|\beta) = |\partial\,\mathrm{KL}(\cdot|\beta)|^2$, this is the Wasserstein–PL inequality with constant λ. This is precisely the curvature-to-logarithmic-Sobolev implication used in Proposition 15.40; Proposition 15.42 then converts the same estimate into exponential convergence of the Fokker–Planck flow.

Figure 15.11 shows these inequalities on an exact one-dimensional Ornstein–Uhlenbeck relaxation, where all quantities can be evaluated accurately by grid quadrature and quantile inversion.

15.4 Training Two-Layer MLPs as Wasserstein Flows

Mean-field limits recast the training of wide neural networks as transport of a distribution of neurons. This section shows how the particle ODE of gradient descent becomes a Wasserstein flow in parameter space. This viewpoint, often summarized as treating parameters as interacting particles, was developed in closely related forms by Mei, Montanari and Nguyen [373], Rotskoff and Vanden-Eijnden [456], and Chizat and Bach [138]. These works pass from the finite-width empirical neuron dynamics to a limiting nonlinear PDE for the neuron law.

Wasserstein training of two-layer MLPs. The key modeling step is to forget the ordering of the hidden neurons and to regard the network weights as a probability measure on parameter space. A finite-width network is then an empirical measure, and the population loss becomes a functional of this measure. Gradient descent on the particle positions is precisely the finite-dimensional discretization of a Wasserstein gradient flow for this functional, with the Wasserstein metric acting on neuron parameters rather than on data samples. This is the sense in which training a two-layer mean-field MLP becomes a transport problem for the law of its neurons.

We use $z \in \mathbb{R}^d$ for the input data and $y \in \mathbb{R}^{d'}$ for the label. A neuron is a particle

$$x = (u, v) \in \mathbb{R}^d \times \mathbb{R}^{d'},$$

where u is the inner weight and v is the outer vector weight. For a scalar nonlinearity σ, define the vector-valued feature

$$\psi(x, z) = v\, \sigma(\langle u,\, z\rangle) \in \mathbb{R}^{d'}.$$

The width-n network and its mean-field version are

$$G_X(z) = \frac{1}{n}\sum_{i=1}^{n} \psi(x_i, z), \qquad G_\alpha(z) = \int \psi(x, z)\mathrm{d}\alpha(x), \qquad \alpha = \frac{1}{n}\sum_i \delta_{x_i}.$$

This formulation removes the artificial ordering of neurons and allows α to be a continuous distribution of infinitely many neurons.

Let ζ be a probability distribution on data-label pairs $(z, y) \in \mathbb{R}^d \times \mathbb{R}^{d'}$. The population risk is

$$f(\alpha) = \int \ell(G_\alpha(z), y)\mathrm{d}\zeta(z, y),$$

and the empirical risk is the special case $\zeta = \zeta_N := N^{-1}\sum_{k=1}^{N} \delta_{(z_k, y_k)}$. Since $\alpha \mapsto G_\alpha$ is linear, f is convex as a function of α whenever $\ell(\cdot, y)$ is convex. For the empirical neuron law $\alpha_X = n^{-1}\sum_i \delta_{x_i}$, the Wasserstein metric induces on particles the rescaled metric $n^{-1}\sum_i \|\dot{x}_i\|^2$. The corresponding particle flow is

$$\dot{x}_i = -n\nabla_{x_i} F(X), \qquad F(X) = f\left(\frac{1}{n}\sum_i \delta_{x_i}\right),$$

which is the gradient flow of $F(X) = f(\alpha_X)$ for this Wasserstein particle metric, equivalently Euclidean gradient descent with the time scale multiplied by n. It gives a particle discretization of (15.4).

Assume that ℓ is differentiable in its first variable. The first variation is

$$\delta f(\alpha)(x) = \int \langle \nabla_1 \ell(G_\alpha(z), y),\, \psi(x, z)\rangle \mathrm{d}\zeta(z, y), \tag{15.28}$$

and the Wasserstein gradient in parameter space is

$$\mathbb{W} f(\alpha)(x) = \nabla_x \delta f(\alpha)(x) = \int [D_x \psi(x, z)]^\top \nabla_1 \ell(G_\alpha(z), y)\mathrm{d}\zeta(z, y).$$

For the squared Euclidean loss $\ell(s, y) = \frac{1}{2}\|s - y\|^2$, the energy is the sum of a quadratic interaction and a linear potential:

$$f(\alpha) = \frac{1}{2}\iint k(x, x')\mathrm{d}\alpha(x)\mathrm{d}\alpha(x') + \int g(x)\mathrm{d}\alpha(x) + \frac{1}{2}\int \|y\|^2 \mathrm{d}\zeta(z, y), \tag{15.29}$$

with

$$k(x, x') = \int \langle \psi(x, z),\, \psi(x', z)\rangle \mathrm{d}\zeta(z, y), \qquad g(x) = -\int \langle y,\, \psi(x, z)\rangle \mathrm{d}\zeta(z, y). \tag{15.30}$$

Thus

$$\delta f(\alpha)(x) = \int k(x, x')\mathrm{d}\alpha(x') + g(x), \qquad \mathbb{W} f(\alpha)(x) = \int \nabla_x k(x, x')\mathrm{d}\alpha(x') + \nabla_x g(x).$$

These kernels are generally not convex in the particle variable, so the geodesic-convex convergence theory above does not apply directly.

Figure 15.12 illustrates the resulting transport of neurons for a homogeneous ReLU model.

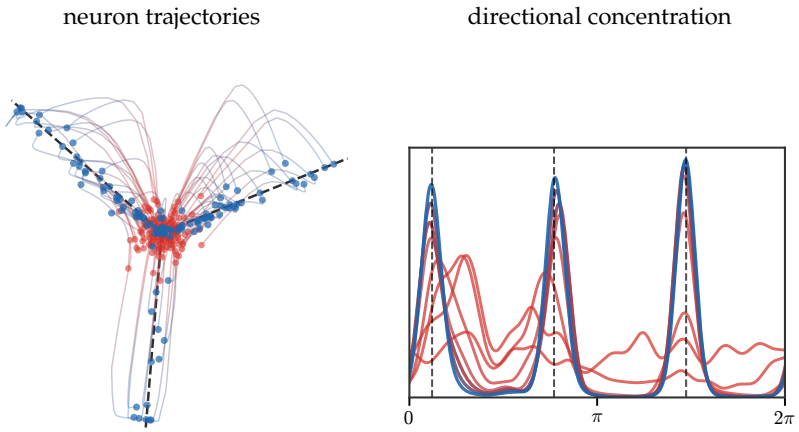


Figure 15.12: Mean-field training of a homogeneous two-layer model as transport in neuron space. The left panel shows the Wasserstein particle gradient flow in the reduced homogeneous coordinates $(|u|v_1, |u|v_2)$, with black dashed rays marking the teacher directions. The right panel shows the weighted angular density along a front-loaded sequence of times, colored from red to blue, so that the early concentration of neuron directions is visible. The display follows the rendering of the auxiliary MLP experiment but keeps only the $\mathcal{W}_2$ flow, not the spectral-flow comparison.

Classical convexity and stationarity. Before using the specific homogeneity mechanism of Chizat and Bach, it is useful to isolate a simpler convex-analytic principle behind many mean-field arguments. Consider an energy of the form

$$f(\alpha) = \frac{1}{2}\iint k(x,x')\mathrm{d}\alpha(x)\mathrm{d}\alpha(x') + \int V(x)\mathrm{d}\alpha(x) + C,$$

on probability measures over a parameter domain. Assume that the quadratic part is convex in the classical sense, namely convex for the affine structure of measures:

$$Q((1-s)\alpha + s\beta) \leqslant (1-s)Q(\alpha) + sQ(\beta), \qquad Q(\alpha) = \frac{1}{2}\iint k\mathrm{d}\alpha\mathrm{d}\alpha.$$

This is ordinary convexity of the functional on the convex set of measures, not displacement convexity along $\mathcal{W}_2$ geodesics.

Proposition 15.51 (Affine convexity and stationary positive densities)**.** *Let $\Omega \subset \mathbb{R}^d$ be connected, let $f = Q + \int V\,\mathrm{d}\alpha + C$ be as above on $\mathcal{P}(\Omega)$, and assume that Q is classically convex. Suppose that a Wasserstein gradient flow α_t of f converges to $\alpha_\infty = \rho_\infty \mathrm{d}x$, where $\rho_\infty > 0$ almost everywhere on Ω, and that $\inf_{t\geqslant 0} f(\alpha_t) > -\infty$. Assume that $\delta f(\alpha_\infty) \in C^1(\Omega)$ and that the squared Wasserstein slope is lower semicontinuous along the convergence, in the sense that*

$$\int_\Omega \|\nabla \delta f(\alpha_\infty)\|^2 \mathrm{d}\alpha_\infty \leqslant \liminf_{n\to\infty} \int_\Omega \|\nabla \delta f(\alpha_{t_n})\|^2 \mathrm{d}\alpha_{t_n}$$

whenever $t_n \to \infty$. Then α_∞ is a global minimizer of f.

Proof. The energy-dissipation identity and the lower bound on f along the convergent trajectory imply

$$\int_0^\infty \int_\Omega \|\nabla \delta f(\alpha_t)\|^2 \mathrm{d}\alpha_t \mathrm{d}t < +\infty.$$

Hence there exists $t_n \to \infty$ for which the inner integral tends to zero. The assumed lower semicontinuity gives

$$\int_\Omega \|\nabla \delta f(\alpha_\infty)\|^2 \mathrm{d}\alpha_\infty = 0.$$

Because $\rho_\infty > 0$ almost everywhere and $\nabla \delta f(\alpha_\infty)$ is continuous, this gradient vanishes throughout Ω. Connectedness then implies that $\delta f(\alpha_\infty)$ is constant. Consequently, for every $\beta \in \mathcal{P}(\Omega)$,

$$\int_\Omega \delta f(\alpha_\infty)\mathrm{d}(\beta - \alpha_\infty) = 0.$$

Classical convexity of f in the affine variable α gives the subgradient inequality

$$f(\beta) \geqslant f(\alpha_\infty) + \int \delta f(\alpha_\infty) \mathrm{d}(\beta - \alpha_\infty) = f(\alpha_\infty).$$

Thus no competitor has smaller energy. Strict positivity is essential to this argument: without it, stationarity controls the first variation only on the support explored by the limiting measure. For square-loss two-layer mean-field models, (15.29) is exactly of this quadratic-plus-linear form, and positive semidefiniteness of the induced kernel k is the classical convexity assumption. □

The convergence theory then separates two regimes. Chizat and Bach prove global convergence for the unregularized, noiseless Wasserstein flow of positively homogeneous models [138]. Mei, Montanari and Nguyen prove convergence for the noisy, regularized dynamics [373]: at the PDE level the noise adds a diffusion, or Laplacian, term, so an initially singular neuron law is immediately smoothed and acquires a density. Rotskoff and Vanden-Eijnden emphasize the parameters-as-particles interpretation, long-time convergence and asymptotic error scaling with the network width [456]. The following formal statement isolates the noiseless Chizat–Bach mechanism and ignores the technical issues due to ReLU non-smoothness, support propagation and compactness.

Proposition 15.52 (Formal global optimality for two-homogeneous mean-field flows)**.** *Assume that the feature is positively two-homogeneous in the neuron variable,*

$$\psi(\lambda x, z) = \lambda^2 \psi(x, z) \qquad (\lambda > 0),$$

and that $f(\alpha) = J(G_\alpha)$ with J convex and differentiable as a functional of the predictor. Let α be a smooth stationary point of the Wasserstein flow, so that $\nabla_x \delta f(\alpha)(x) = 0$ on $\mathrm{supp}(\alpha)$. Assume also full directional support: for every unit direction ω, the support of α intersects the ray $\{\lambda\omega : \lambda > 0\}$. Then α is a global minimizer of f over the mean-field model class.

Proof. Choose the first-variation representative

$$h_\alpha(x) = \delta f(\alpha)(x) = \langle \nabla J(G_\alpha), \psi(x, \cdot) \rangle_\rho .$$

Additive constants in first variations do not affect the Wasserstein gradient or first-order inequalities between probability measures; this normalization is useful because it inherits the homogeneity of ψ. By two-homogeneity of ψ, one has

$$h_\alpha(\lambda x) = \lambda^2 h_\alpha(x), \qquad \lambda > 0.$$

Taking $x = 0$ and any $\lambda \neq 1$ in this identity gives $h_\alpha(0) = 0$. We first record that stationarity forces h_α to vanish on $\mathrm{supp}(\alpha)$. Indeed, if $x = r\omega \in \mathrm{supp}(\alpha)$ with $r > 0$ and $\|\omega\| = 1$, then

$$0 = \langle \nabla h_\alpha(r\omega),\, \omega \rangle = \frac{\mathrm{d}}{\mathrm{d}s} h_\alpha(s\omega)\Big|_{s=r} = 2r h_\alpha(\omega),$$

so $h_\alpha(\omega) = 0$, and hence $h_\alpha(x) = r^2 h_\alpha(\omega) = 0$. The case $x = 0$ has already been covered, so $h_\alpha = 0$ on $\mathrm{supp}(\alpha)$ and $\int h_\alpha \, \mathrm{d}\alpha = 0$.

We now argue by contradiction. Suppose that a competitor β satisfies $f(\beta) < f(\alpha)$. Since

$$G_\beta - G_\alpha = \int \psi(x, \cdot) \mathrm{d}(\beta - \alpha)(x),$$

convexity of J gives

$$f(\beta) - f(\alpha) \geqslant \int h_\alpha(x) \mathrm{d}(\beta - \alpha)(x).$$

The left-hand side is negative, while $\int h_\alpha \, \mathrm{d}\alpha = 0$, so $\int h_\alpha \, \mathrm{d}\beta < 0$. Since $h_\alpha(0) = 0$, this strict negativity must occur away from the origin: there exists a nonzero point $x = r\omega$ with $r > 0$, $\|\omega\| = 1$, and $h_\alpha(x) < 0$. Equivalently $h_\alpha(\omega) = h_\alpha(x)/r^2 < 0$. By full directional support, there is $r_\omega > 0$ such that $r_\omega \omega \in \mathrm{supp}(\alpha)$. At this support point,

$$\langle \nabla h_\alpha(r_\omega \omega),\, \omega \rangle = 2 r_\omega h_\alpha(\omega) < 0,$$

which contradicts the stationarity condition $\nabla h_\alpha = 0$ on $\mathrm{supp}(\alpha)$. Thus no competitor has smaller risk.

The rigorous Chizat–Bach theorem replaces the full directional support assumption by propagation and overparameterization hypotheses ensuring that any negative first-variation direction is represented by the evolving support. The same radial-derivative contradiction, applied after this support-propagation step, then rules out non-optimal stationary limits. □

15.5 Generalized Dynamic Wasserstein Flows

This section explains how changing the metric geometry changes the associated descent PDE.

Generalized Wasserstein flows. The JKO construction is not tied to the quadratic Wasserstein distance. Once a space of probability measures is equipped with a metric or extended metric d, it gives the following generalized descent construction.

Definition 15.53 (Generalized action gradient flow)**.** Let $d = \mathsf{D}_{\mathbb{A}}$ be the dynamic distance induced by the action $\mathbb{A}$ in Definition 14.8. Given $\alpha^{(0)} \in \mathcal{P}(\mathbb{R}^d)$ and $\tau > 0$, define

$$\alpha^{(\ell+1)} \in \operatorname*{argmin}_{\alpha \in \mathcal{P}(\mathbb{R}^d)} \left\{ \frac{1}{2\tau} d^2(\alpha^{(\ell)}, \alpha) + f(\alpha) \right\}. \tag{15.31}$$

Any continuous-time limit of its piecewise-constant interpolants is called a generalized action gradient flow of f for $\mathbb{A}$.

This is the metric-gradient-flow framework of Ambrosio, Gigli and Savaré [12, 11]: under compactness, coercivity and lower-semicontinuity assumptions adapted to d, the piecewise-constant interpolants of (15.31) can converge, as $\tau \to 0$, to a curve of maximal slope. A nonmetric discrepancy may still be inserted in the same variational update, but the metric theory does not then apply automatically. In the local mass-preserving metrics considered below, a smooth limiting curve is again represented by a continuity equation of the type introduced in (14.2) and used for Wasserstein flows in (15.4),

$$\partial_t \alpha_t + \operatorname{div}(\alpha_t v_{\alpha_t}) = 0, \tag{15.32}$$

but the velocity v_α is selected by the local steepest-descent rule associated with the geometry induced by d, not necessarily by the classical $\mathcal{W}_2$ rule. Unbalanced variants replace this continuity equation by a balance law, while nonlocal variants replace vector fields by jump fluxes.

The metric ingredient needed below is the squared-speed action $\mathbb{A}(\alpha, w)$ introduced in Section 14.3. Its length distance is already defined in (14.15); the present section only uses the local cost in the infinitesimal variational step defined below. Some actions are pointwise integrals, such as ordinary $\mathcal{W}_2$ with $\mathbb{A}(\alpha, w) = \int \|w\|^2 \mathrm{d}\alpha$. Others are built from a homogeneous local density but then squared at the metric level: for $\mathcal{W}_p$, the local densities are $A_p(a, w) = a\|w\|^p$ and $J_p(a, m) = \|m\|^p / a^{p-1}$, while the action paired with the JKO penalty is the squared Finsler speed

$$\mathbb{A}_p(\alpha, w) = \left(\int \|w\|^p \mathrm{d}\alpha \right)^{2/p}.$$

Spectral geometries use the nonlocal covariance action $\mathbb{A}_\gamma$ from (14.23). In the $\mathcal{W}_2$ case, all objects reduce to the familiar Benamou–Brenier ones:

$$A_2(a, w) = a\|w\|^2, \qquad J_2(a, m) = \frac{\|m\|^2}{a}, \qquad \mathsf{D}_{\mathbb{A}} = \mathcal{W}_2 .$$

Definition 15.54 (Penalized minimization oracle)**.** Let $\mathbb{A}(\alpha, \cdot)$ be the local squared-speed action at α. For a cotangent field u for which the pairing below is finite, usually $u = \mathbb{W} f(\alpha)$, the penalized minimization oracle associated with this action is

$$\mathrm{PMO}_{\mathbb{A},\alpha}(u) \in \operatorname*{argmin}_{w} \left\{ \int \langle u(x), w(x) \rangle \mathrm{d}\alpha(x) + \frac{1}{2} \mathbb{A}(\alpha, w) \right\}, \tag{15.33}$$

where the minimization is over admissible finite-action velocity representatives for the chosen geometry. In Hilbertian examples $u \in L^2(\alpha; \mathbb{R}^d)$; for $\mathcal{W}_p$, the natural dual space is $L^q(\alpha; \mathbb{R}^d)$, with $q = p/(p-1)$.

For an energy f, the formal small-step expansion of (15.31), when the distance is generated by the squared-speed action $\mathbb{A}$ and the JKO displacement admits an admissible finite-action tangent representative, selects the PMO direction

$$v_\alpha = \mathrm{PMO}_{\mathbb{A},\alpha}(\mathbb{W} f(\alpha)). \tag{15.34}$$

Equivalently, the formal evolution is

$$\partial_t \alpha_t + \operatorname{div}(\alpha_t \ \mathrm{PMO}_{\mathbb{A},\alpha_t}(\mathbb{W} f(\alpha_t))) = 0. \tag{15.35}$$

If $w \mapsto \mathbb{A}(\alpha, w)$ has zero-action directions, the minimization in (15.33) is understood on the quotient tangent space, or after projecting onto finite-action directions modulo the kernel of $\mathbb{A}(\alpha, \cdot)$. Along smooth solutions of (15.35), whenever the minimizer is attained and $w \mapsto \mathbb{A}(\alpha, w)$ is 2-homogeneous, one obtains the dissipation identity

$$\frac{\mathrm{d}}{\mathrm{d}t} f(\alpha_t) = \langle \mathbb{W} f(\alpha_t), v_t \rangle_{\alpha_t} = -\mathbb{A}(\alpha_t, v_t), \qquad v_t = \mathrm{PMO}_{\mathbb{A},\alpha_t}(\mathbb{W} f(\alpha_t)).$$

For $\mathcal{W}_2$, $\mathbb{A}(\alpha, w) = \int \|w\|^2 \mathrm{d}\alpha$ and $\mathrm{PMO}_{\mathbb{A},\alpha}(u) = -u$. Hence (15.35) reduces to the standard Wasserstein gradient-flow equation

$$\partial_t \alpha_t = \operatorname{div}(\alpha_t \mathbb{W} f(\alpha_t)),$$

which is the convention used throughout the preceding sections.

In the restricted Riemannian case of (14.16), where

$$\mathbb{A}(\alpha, w) = \langle Q_\alpha w, w \rangle_{L^2(\alpha)},$$

the PMO becomes the linear preconditioning rule

$$\mathrm{PMO}_{\mathbb{A},\alpha}(u) = -Q_\alpha^{-1} u, \qquad v_\alpha = -Q_\alpha^{-1} \mathbb{W} f(\alpha). \tag{15.36}$$

The linear inverse formula is special to Hilbertian actions. A basic non-Hilbertian example is $\mathcal{W}_p$: for $1 < p < +\infty$, the squared Finsler tangent action associated with the distance is

$$\mathbb{A}_p(\alpha, w) = \left(\int \|w\|^p \mathrm{d}\alpha \right)^{2/p}.$$

Writing $q = p/(p-1)$ and $U = \|u\|_{L^q(\alpha)}$, its PMO is the duality-map formula

$$\mathrm{PMO}_{\mathbb{A}_p,\alpha}(u)(x) = -U^{2-q} \|u(x)\|^{q-2} u(x), \tag{15.37}$$

with the conventions that the right-hand side is 0 when $U = 0$, and that $\|u(x)\|^{q-2} u(x) = 0$ when $u(x) = 0$. This is nonlinear in u, and it reduces to $-u$ only for $p = 2$. Endpoint cases such as $p = 1$ are interpreted through subgradients of the dual norm.

The examples below are organized by the same dictionary. Concave-mobility flows use $\mathbb{A}_{\theta,\lambda}(\alpha, v)$ from Section 14.3; spectral flows use $\mathbb{A}_\gamma(\alpha, v)$ from (14.23); kernelized Benamou–Brenier flows use $\mathbb{A}_k(\alpha, v) = \|v\|^2_{\mathcal{H}_k^d}$ with a restricted RKHS tangent space. Nonlocal logarithmic-mean geometries also have quadratic actions, but their tangent variables are edge potentials or jump velocities rather than vector fields in $\mathbb{R}^d$; their flows are treated separately in Section 15.6. WFR, defined by (14.41) and studied separately in Section 15.7, is another quadratic geometry, but its tangent variable is a pair (v, g) combining displacement and mass modulation.

The generalized distances of Section 14.3 are useful precisely because they change what *descent* means. The energy functional can be the same, while the local metric tensor, Finsler norm, mobility, spectral gauge or RKHS constraint changes the PDE selected by the minimizing movement. The rest of this section focuses on local or vector-field mass-preserving geometries. Nonlocal pair-space geometries are separated in Section 15.6, and the transport–reaction case is separated in Section 15.7, because their tangent variables are structurally different.

General mobility flows. The concave-mobility distances of Section 14.3 replace the scalar linear mobility a by a concave mobility $\theta(a)$. For a density field $a = \rho_t(x)$, this changes the Onsager operator of the gradient flow while keeping a continuity equation.

Let λ be Lebesgue measure, write $\alpha = \rho\lambda$, and let $u = \nabla(\delta f/\delta\rho)$. For the velocity action $\mathbb{A}_{\theta,\lambda}$, the PMO introduced in Definition 15.54 is the pointwise minimizer of

$$\int \langle u, w \rangle \rho \, \mathrm{d}x + \frac{1}{2} \int \frac{\rho}{\theta(\rho)} \|w\|^2 \rho \, \mathrm{d}x,$$

so, wherever $\rho > 0$ and $\theta(\rho) > 0$,

$$\mathrm{PMO}_{\mathbb{A}_{\theta,\lambda},\rho\lambda}(u) = -\frac{\theta(\rho)}{\rho}u. \tag{15.38}$$

Thus the flux ρw is $-\theta(\rho)\nabla(\delta f/\delta\rho)$. For a smooth density ρ_t and an energy f, the formal gradient flow associated with the pointwise momentum action $J_\theta(a,m) = |m|^2/\theta(a)$ is

$$\partial_t \rho_t = \nabla\cdot\left(\theta(\rho_t)\nabla\frac{\delta f}{\delta\rho}(\rho_t)\right). \tag{15.39}$$

If

$$f(\rho) = \int U(\rho(x))\,\mathrm{d}x + \int V(x)\rho(x)\,\mathrm{d}x,$$

then (15.39) becomes

$$\partial_t\rho = \nabla\cdot\big(\theta(\rho)\big(U''(\rho)\nabla\rho + \nabla V\big)\big).$$

Thus $\theta(a) = a$ and $U(a) = a\log a$ gives the usual heat or Fokker–Planck equation, while $\theta(a) = a(1-a/M)$ gives a volume-filling drift–diffusion whose mobility vanishes at saturation. Power internal energies and power mobilities tune nonlinear diffusion: for instance $\theta(a) = a$ and $U(a) = a^m/(m-1)$ yield $\partial_t\rho = \Delta(\rho^m)$, whereas the entropy with $\theta(a) = a^\gamma$, $0 < \gamma < 1$, gives the fast-diffusion-type equation $\partial_t\rho = (1/\gamma)\Delta(\rho^\gamma)$. This viewpoint is useful for nonlinear diffusions, exclusion models, volume-filling models and finite-volume discretizations [189]; related vector-valued and dissipative extensions are discussed in Section 13.1.

Dynamic spectral Wasserstein flows. The spectral tangent action (14.23), and the dynamic distance (14.24) it generates, normalize the whole velocity covariance through a monotone spectral gauge. Using this geometry for gradient descent replaces the pointwise Wasserstein direction by a globally preconditioned direction. This is close to many large-scale optimizers: normalized gradient methods solve norm-constrained linear minimization problems [423]; related ideas appear in normalized SGD [394, 164], tensor-aware optimizers such as Shampoo [258], and Muon-type spectral normalizations [293, 344]. We now describe the mean-field version developed in [427]. The trace gauge gives the classical $\mathcal{W}_2$ flow, while the operator gauge $\gamma(M) = \lambda_{\max}(M)$ gives the idealized Muon geometry.

Specializing Definition 15.54 to the spectral action $\mathbb{A}_\gamma$, the descent direction associated with a field $g \in L^2(\alpha;\mathbb{R}^d)$ is simply $\mathrm{PMO}_{\mathbb{A}_\gamma,\alpha}(g)$. The field g should be read as the Wasserstein gradient $g = \mathbb{W}f(\alpha) = \nabla\delta f(\alpha)$. Since $v \mapsto \mathbb{A}_\gamma(\alpha,v)$ is 2-homogeneous, fixed-speed normalized algorithms keep the same ray as this PMO and only change the scalar normalization; equivalently, they minimize the linear pairing over the unit ball $\mathbb{A}_\gamma(\alpha,w) \leqslant 1$. In the trace case, $\mathbb{A}_{\mathrm{tr}}(\alpha,v) = \int\|v\|^2\mathrm{d}\alpha$ and $\mathrm{PMO}_{\mathbb{A}_{\mathrm{tr}},\alpha}(g) = -g$. For other gauges, the velocity is globally preconditioned by the covariance of g under α.

Proposition 15.55 (Polar formula for the spectral direction)**.** *Let*

$$S_\alpha(g) \coloneqq \int g(x)g(x)^\top\mathrm{d}\alpha(x).$$

Assume that the inverse-trace problem admits a positive definite minimizer

$$A_\alpha^\star \in \operatorname*{argmin}_{A\in\mathcal{B}_\gamma\cap\mathbb{S}^d_{++}} \mathrm{tr}\Big(A^{-1}S_\alpha(g)\Big).$$

Then

$$\mathrm{PMO}_{\mathbb{A}_\gamma,\alpha}(g)(x) = -(A_\alpha^\star)^{-1}g(x) \tag{15.40}$$

is the spectral PMO direction. If the minimizer lies on the boundary of $\mathcal{B}_\gamma$, the same formula is understood by a limiting argument along positive definite minimizers, or equivalently by replacing the inverse by the corresponding inverse on the range of $S_\alpha(g)$.

Proof. Using the polar formula $\gamma(M) = \sup_{A\in\mathcal{B}_\gamma}\mathrm{tr}(AM)$, the PMO objective (15.33) with $\mathbb{A} = \mathbb{A}_\gamma$ can be written as

$$\sup_{A\in\mathcal{B}_\gamma}\left\{\int\langle g,\,v\rangle\mathrm{d}\alpha + \frac{1}{2}\int v^\top Av\mathrm{d}\alpha\right\}.$$

For a fixed $A \succ 0$, the quadratic lower bound

$$\int \langle g,\, v\rangle \mathrm{d}\alpha + \frac{1}{2}\int v^\top A v \mathrm{d}\alpha$$

is minimized by $v_A = -A^{-1}g$, with value $-\frac{1}{2}\operatorname{tr}(A^{-1}S_\alpha(g))$. Let $v_\star = -(A_\alpha^\star)^{-1}g$ and

$$M_\star = \int v_\star(x)v_\star(x)^\top \mathrm{d}\alpha(x) = (A_\alpha^\star)^{-1}S_\alpha(g)(A_\alpha^\star)^{-1}.$$

Since $A_\alpha^\star$ minimizes $A \mapsto \operatorname{tr}(A^{-1}S_\alpha(g))$ over the convex polar set, its first-order optimality condition gives

$$\operatorname{tr}(AM_\star) \leqslant \operatorname{tr}(A_\alpha^\star M_\star) \qquad \text{for all } A \in \mathcal{B}_\gamma.$$

Thus $A_\alpha^\star$ is active in the polar representation of $\gamma(M_\star)$, and

$$\gamma(M_\star) = \operatorname{tr}(A_\alpha^\star M_\star) = \operatorname{tr}((A_\alpha^\star)^{-1}S_\alpha(g)).$$

For any v, the polar formula gives

$$\int \langle g,\, v\rangle \mathrm{d}\alpha + \frac{1}{2}\mathbb{A}_\gamma(\alpha, v) \geqslant \int \langle g,\, v\rangle \mathrm{d}\alpha + \frac{1}{2}\int v^\top A_\alpha^\star v \mathrm{d}\alpha,$$

and the right-hand side is minimized at $v_\star$. The activity identity above shows that equality holds at $v_\star$, hence $v_\star$ is the spectral PMO direction. The boundary case follows by approximation with positive definite matrices in the polar set. □

For the trace gauge, $\mathcal{B}_\gamma = \{A\ ;\ 0 \preceq A \preceq \mathrm{Id}\}$ and the minimizer is $A_\alpha^\star = \mathrm{Id}$. For the operator gauge $\gamma(M) = \lambda_{\max}(M)$, one has $\mathcal{B}_\gamma = \{A \succeq 0\ ;\ \operatorname{tr}(A) \leqslant 1\}$; if $S_\alpha(g) \succ 0$, then

$$A_\alpha^\star = \frac{S_\alpha(g)^{1/2}}{\operatorname{tr}(S_\alpha(g)^{1/2})}, \qquad \mathrm{PMO}_{\mathbb{A}_\gamma,\alpha}(g)(x) = -\operatorname{tr}(S_\alpha(g)^{1/2})\, S_\alpha(g)^{-1/2}g(x).$$

This is the continuum covariance-normalization formula that becomes the Muon polar factor for empirical measures.

Proposition 14.16 identifies the static spectral distance $\mathcal{W}_\gamma$ with the length distance generated by $\mathbb{A}_\gamma$. The infinitesimal descent direction is therefore defined directly by the general PMO problem (15.33) specialized to this action. This action-based statement avoids an important overinterpretation: static/dynamic equality alone does not imply $\mathcal{W}_\gamma^2(\alpha, (\mathrm{Id}+\tau v)_\sharp\alpha) = \tau^2\mathbb{A}_\gamma(\alpha, v) + o(\tau^2)$ for every smooth field v, because v need not be the minimal finite-action representative of the induced tangent perturbation. Whenever a minimizing-movement scheme for $\mathcal{W}_\gamma$ converges to a smooth curve whose metric derivative is represented by this minimal action, it produces the same evolution described below.

Proposition 15.56 (Formal normalized spectral gradient flow)**.** *Assume that f has a smooth first variation and that the spectral PMO is attained along a smooth curve. The formal gradient flow generated by the dynamic spectral action $\mathbb{A}_\gamma$ solves*

$$\partial_t\alpha_t + \operatorname{div}\Big(\alpha_t\ \mathrm{PMO}_{\mathbb{A}_\gamma,\alpha_t}(\mathbb{W}f(\alpha_t))\Big) = 0. \tag{15.41}$$

Equivalently, if $A_t^\star$ solves the inverse-trace problem for

$$S_t = \int \mathbb{W}f(\alpha_t)(x)\mathbb{W}f(\alpha_t)(x)^\top \mathrm{d}\alpha_t(x),$$

then the velocity is

$$v_t(x) = -(A_t^\star)^{-1}\mathbb{W}f(\alpha_t)(x).$$

Moreover, along smooth solutions,

$$\frac{\mathrm{d}}{\mathrm{d}t}f(\alpha_t) = -\mathbb{A}_\gamma(\alpha_t, v_t).$$

If the minimizing movements (15.31) for $d = \mathcal{W}_\gamma$ converge to a smooth curve with this minimal-action metric derivative, that limit satisfies the same equation.

Proof. The action-based steepest-descent rule (15.34) gives $v_t = \mathrm{PMO}_{\mathbb{A}_\gamma,\alpha_t}(\mathbb{W}f(\alpha_t))$. Inserting this velocity into the continuity equation (15.32) gives (15.41). Proposition 15.55 gives the explicit inverse-trace formula. Finally, the variation formula gives $\frac{\mathrm{d}}{\mathrm{d}t} f(\alpha_t) = \int \langle \mathbb{W}f(\alpha_t), v_t\rangle \mathrm{d}\alpha_t$. Since v_t minimizes a 2-homogeneous penalized objective, the one-dimensional rescaling $s \mapsto sv_t$ has derivative zero at $s = 1$, hence

$$\int \langle \mathbb{W}f(\alpha_t), v_t\rangle \mathrm{d}\alpha_t + \mathbb{A}_\gamma(\alpha_t, v_t) = 0,$$

which proves the dissipation identity. □

Remark 15.57 (Empirical and Muon limits). For empirical measures $\alpha_X = n^{-1}\sum_i \delta_{x_i}$, stack the velocities $v(x_i)$ into a matrix $V \in \mathbb{R}^{n\times d}$ and the gradients $g(x_i)$ into a matrix $G \in \mathbb{R}^{n\times d}$. By 1-homogeneity of γ,

$$\mathbb{A}_\gamma(\alpha_X, v) = \gamma\left(\frac{1}{n}V^\top V\right) = \frac{1}{n}\gamma(V^\top V), \qquad \int \langle g, v\rangle \mathrm{d}\alpha_X = \frac{1}{n}\operatorname{tr}(G^\top V).$$

Thus the empirical specialization of the PMO problem (15.33) is, up to the harmless factor $1/n$, the finite-dimensional matrix descent problem. Writing

$$(G_X(t))_i = \mathbb{W}f(\alpha_{X(t)})(x_i(t)),$$

the empirical flow is

$$\dot{X}(t) \in \operatorname*{argmin}_{V\in\mathbb{R}^{n\times d}} \left\{\operatorname{tr}(G_X(t)^\top V) + \frac{1}{2}\gamma(V^\top V)\right\}. \tag{15.42}$$

Here G_X is the gradient for the empirical Wasserstein metric $\|\dot{X}\|_n^2 = n^{-1}\sum_i \|\dot{x}_i\|^2$, not the ordinary Euclidean gradient of $f(\alpha_X)$. If one rewrites the same rule with the unweighted matrix convention, it corresponds to the lift $X \mapsto nf(\alpha_X)$; using $X \mapsto f(\alpha_X)$ only rescales time by the factor n. If $G = U\operatorname{diag}(\sigma_i)W^\top$ and $\gamma(M) = \lambda_{\max}(M)$, then

$$\operatorname*{argmin}_{V\in\mathbb{R}^{n\times d}} \left\{\operatorname{tr}(G^\top V) + \frac{1}{2}\lambda_{\max}(V^\top V)\right\} \ni -\|G\|_{S_1}UW^\top = -\operatorname{tr}\Big((G^\top G)^{1/2}\Big)\, G(G^\top G)^{\dagger/2}. \tag{15.43}$$

The ray of this direction is the polar factor $-UW^\top$, and the scalar factor $\|G\|_{S_1}$ can be absorbed into a time or step-size normalization. This is the exact-polar, full-batch, continuous-time idealization of Muon [293, 427]. Practical Muon implementations replace the polar factor by a few Newton–Schulz iterations for GPU efficiency [344]; this polynomial approximation fits the same spectral-normalization philosophy, although adaptive rescaling and finite momentum make the practical algorithm more than a position-only metric gradient flow.

The effect of this spectral normalization is visible already in the homogeneous two-layer ReLU model discussed in Section 15.4. Figure 15.13 reproduces, in the style of the book figures, the numerical comparison from [427]: the same teacher network, same initialization and same empirical first variation are evolved either by the $\mathcal{W}_2$ particle flow or by the operator-gauge Muon direction.

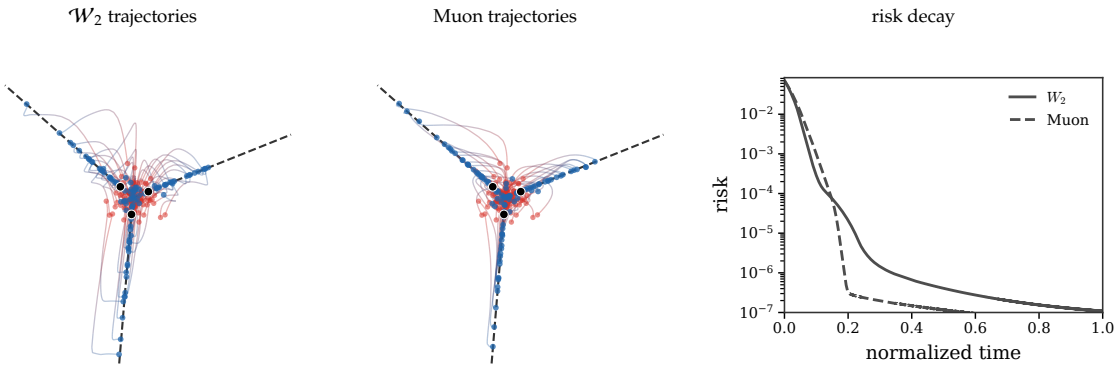


Figure 15.13: Wasserstein versus Muon mean-field training of a homogeneous ReLU model. The first two panels display trajectories in the reduced homogeneous coordinates $(|u|v_1, |u|v_2)$; black dashed rays are the teacher directions. The right panel compares the empirical square-loss decay, with the solid curve for the $\mathcal{W}_2$ flow and the dashed curve for the operator-gauge Muon flow. Both methods use the same particles and first variation; the spectral normalization reaches the low-risk regime earlier in normalized time.

15.6 Nonlocal Wasserstein Flows

The nonlocal distances of Section 14.4 replace the local vector-field tangent model by pairwise exchanges. This separate section records the corresponding gradient-flow interpretation in the same order as the distance constructions: first continuum nonlocal Wasserstein flows and fractional PDEs, then discrete Wasserstein flows on Markov chains. In the continuum jump-kernel case, entropy descent gives nonlocal diffusion and, for power-law kernels, fractional PDEs. On a finite state space, the same logarithmic-mean edge calculus makes a reversible Markov chain exactly the entropy gradient flow of the discrete Wasserstein distance. Both examples are mass-preserving, but they are not local Benamou–Brenier flows: their tangent variables live on pairs or graph edges rather than at individual points.

Nonlocal Wasserstein flows and fractional PDEs. The continuum nonlocal distance $\mathcal{W}_K$ of Section 14.4 replaces local velocities by antisymmetric fluxes across jumps selected by a reversible kernel K. Its logarithmic-mean mobility is chosen so that entropy dissipation produces the Markov generator. This turns jump processes, fractional heat equations and some heavy-tailed stochastic models into gradient flows for nonlocal transport geometries.

Proposition 15.58 (Entropy gradient flow for reversible jump kernels)**.** *Under the regularity and irreducibility assumptions used in Proposition 14.21, the Markov semigroup generated by the closure of*

$$L\psi(x) = \int \big(\psi(y) - \psi(x)\big)K(x, \mathrm{d}y)$$

is the gradient flow, for the distance $\mathcal{W}_K$ defined in (14.29), of the relative entropy

$$\mathrm{Ent}_{\mathfrak{m}}(\alpha) = \int \rho \log \rho \, \mathrm{d}\mathfrak{m}, \qquad \alpha = \rho\mathfrak{m}.$$

Equivalently, in weak form, the entropy flow is

$$\partial_t \rho_t = L\rho_t. \tag{15.44}$$

When K is singular, the integral defining L is understood in the principal-value sense.

Proof idea. The key identity is the logarithmic-mean relation $\theta(a, b)(\log a - \log b) = a - b$. The first variation of $\mathrm{Ent}_{\mathfrak{m}}$ is $\log \rho + 1$. With the sign convention of (14.28), the steepest descent velocity is

$$v(x, y) = -\bar{\nabla}(\log \rho + 1)(x, y) = \log \rho(x) - \log \rho(y).$$

Hence $\theta(\rho(x), \rho(y))v(x, y) = \rho(x) - \rho(y)$. Substituting this in (14.28) and using the symmetry of J gives

$$\frac{\mathrm{d}}{\mathrm{d}t} \int \varphi\rho \, \mathrm{d}\mathfrak{m} = \int \varphi(x) \int (\rho(y) - \rho(x))K(x, \mathrm{d}y)\, \mathfrak{m}(\mathrm{d}x),$$

which is the weak form of $\partial_t \rho = L\rho$. The metric properties needed to justify this computation as a gradient-flow identity are the distance and geodesic results recalled in Proposition 14.21. □

The construction becomes especially transparent for translation-invariant kernels. A power-law jump kernel turns the entropy flow into a fractional heat equation.

On $\mathcal{X} = \mathbb{R}^d$, let

$$K_s(x, \mathrm{d}y) = c_{d,s} \frac{\mathrm{d}y}{|x-y|^{d+s}}, \qquad 0 < s < 2.$$

The associated generator is, up to the normalizing constant,

$$L_s \psi(x) = \text{p.v.} \int_{\mathbb{R}^d} \big(\psi(y) - \psi(x)\big) \frac{c_{d,s}}{|x-y|^{d+s}} \, \mathrm{d}y = -(-\Delta)^{s/2} \psi(x),$$

where $(-\Delta)^{s/2}$ is the fractional Laplacian [467]. Hence the $\mathcal{W}_{K_s}$-gradient flow of the entropy is

$$\partial_t \rho_t = -(-\Delta)^{s/2} \rho_t. \tag{15.45}$$

Figure 15.14 illustrates the qualitative effect of lowering s. Classical heat diffusion, $s = 2$, smooths local discontinuities by nearby averaging. Fractional diffusion keeps a sharper memory of localized peaks but immediately creates algebraic long-range tails, because mass can jump over macroscopic distances.

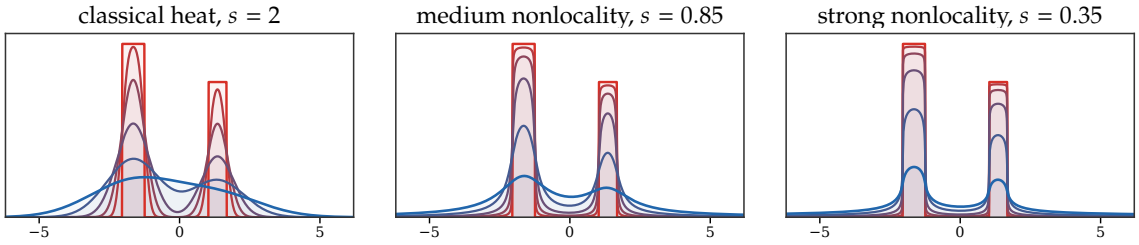


Figure 15.14: Classical and fractional heat flows from two localized indicator bumps. Each panel evolves the same normalized mixture of two intervals by the Fourier multiplier $e^{-t|\xi|^s}$, with time colored from red to blue. The classical heat flow quickly rounds the discontinuities and spreads by local Gaussian averaging. As s decreases, the diffusion becomes increasingly nonlocal: peaks remain more localized near the bumps, while heavier tails appear across the whole displayed window.

More generally, when a target-dependent jump kernel K_β satisfies detailed balance with a desired invariant measure $\beta = \rho_\beta \mathfrak{m}$, the same construction can be applied to $r_t = \mathrm{d}\alpha_t / \mathrm{d}\beta$. The gradient flow of $\mathrm{KL}(\alpha|\beta)$ is then $\partial_t r_t = L_\beta r_t$. This is the nonlocal analogue of the Fokker–Planck gradient flow of relative entropy, and it is one of the motivations behind more recent nonlocal diffusion gradient-flow frameworks [547].

Remark 15.59 (Lévy SDE viewpoint). The same operators appear probabilistically as generators of jump processes. This gives a complementary interpretation of nonlocal PDEs as laws of stochastic processes with heavy-tailed jumps.

For the kernel K_s, the process with generator L_s is the symmetric s-stable Lévy process $X_t = X_0 + L_t^{(s)}$, whose density solves (15.45) [18]. Adding a smooth confining drift gives the fractional Fokker–Planck equation

$$\partial_t \rho_t = \nabla \cdot (\rho_t \nabla V) - \sigma^s (-\Delta)^{s/2} \rho_t, \tag{15.46}$$

which is the forward equation of

$$\mathrm{d}X_t = -\nabla V(X_t)\, \mathrm{d}t + \sigma \, \mathrm{d}L_t^{(s)}.$$

Equation (15.46) should be distinguished from the exactly reversible entropy flow above unless the jump kernel is chosen to satisfy detailed balance with the target measure. Both viewpoints, however, share the same nonlocal mechanism: relaxation is driven not only by local drift and Brownian diffusion but also by rare long jumps.

Remark 15.60 (Connection with stochastic gradient dynamics). This nonlocal point of view is useful in machine learning because stochastic optimization noise need not be well approximated by

Brownian noise. Classical diffusion approximations model small-step SGD by a Brownian SDE near stationarity [366]. In deep-network regimes, empirical and theoretical work has shown that gradient noise and even stationary iterates can be heavy-tailed [487, 259]. A coarse continuous-time model for parameters Θ_t is then

$$\mathrm{d}\Theta_t = -\nabla \mathcal{L}(\Theta_t)\,\mathrm{d}t + \sigma\,\mathrm{d}L_t^{(s)},$$

whose law solves a fractional Fokker–Planck equation of the form (15.46). The jumps represent occasional large stochastic-gradient fluctuations, so they can move the dynamics between basins in a way that Brownian approximations tend to understate. This makes nonlocal Wasserstein geometries a useful language for linking heavy-tailed training noise, fractional PDE models, and measure-valued gradient-flow ideas. The correspondence remains a modeling approximation: the effective tail index, anisotropy, minibatch correlations and learning-rate schedule all influence whether a Lévy limit is a faithful description of a concrete training run.

Discrete Wasserstein flows on Markov chains. The finite-state distance $\mathcal{W}_K$ in Section 14.4 is designed so that the reversible Markov chain itself becomes an entropy gradient flow. This is the discrete counterpart of the fact that the heat equation is the Wasserstein gradient flow of Shannon entropy.

For masses $a_i = \pi_i \rho_i$, or equivalently relative densities $\rho_i = a_i/\pi_i$ with respect to the invariant law π, the entropy relative to π is

$$\mathrm{Ent}_\pi(\rho) := \sum_i \pi_i \rho_i \log \rho_i.$$

Applying the general minimizing movement (15.31) with $d = \mathcal{W}_K$ and $f = \mathrm{Ent}_\pi$, the small-τ first-order optimality condition gives

$$\frac{\rho^{(\ell+1)} - \rho^{(\ell)}}{\tau} = -\mathcal{K}_{\rho^{(\ell+1)}} \log \rho^{(\ell+1)} + o(1) = K\rho^{(\ell+1)} + o(1).$$

Here $(K\rho)_i = \sum_j K_{i,j}(\rho_j - \rho_i)$. Under smoothness and strict positivity, $\rho^{(\ell+1)} = \rho^{(\ell)} + O(\tau)$, so the last expression also equals $K\rho^{(\ell)} + O(\tau)$. Thus the discrete Wasserstein geometry is engineered so that the entropy gradient flow is the Markov semigroup; the proposition below makes this identity explicit.

Proposition 15.61 (Entropy gradient flow of a reversible Markov chain)**.** *Let K be reversible with invariant law π. The gradient flow of Ent_π for the metric $\mathcal{W}_K$ is the forward equation of the Markov chain,*

$$\dot\rho_i(t) = \sum_j K_{i,j}\big(\rho_j(t) - \rho_i(t)\big). \tag{15.47}$$

Equivalently, for the masses $a_i(t) = \pi_i\rho_i(t)$, this is $\dot a_i(t) = \sum_j (a_j(t)K_{j,i} - a_i(t)K_{i,j})$.

Proof. The first variation of the entropy, with respect to the weighted pairing $\sum_i \pi_i \xi_i \varphi_i$, is $\log \rho_i + 1$. Constants do not contribute to $\mathcal{K}_\rho$, hence the metric gradient-flow equation is $\dot\rho = -\mathcal{K}_\rho \log \rho$. Using the identity $\theta(a,b)(\log a - \log b) = a - b$, one obtains, componentwise,

$$\dot\rho_i = -\sum_j K_{i,j}\theta(\rho_i,\rho_j)(\log\rho_i - \log\rho_j) = \sum_j K_{i,j}(\rho_j - \rho_i),$$

which is the density form of the Kolmogorov forward equation. Multiplying by π_i and using detailed balance gives the equation for the probability masses. □

15.7 Dynamic Unbalanced OT and WFR Flows

The generalized dynamic Wasserstein flows of Section 15.5 are primarily mass-preserving: their tangent vectors are generated by continuity equations and are therefore velocity fields advecting the measure. Unbalanced transport changes the state space from probabilities to positive finite measures, where descent may both move and create or remove mass. The tangent variable is then a pair (v, g): the vector field v displaces mass, while the scalar field g is a relative growth rate. This transport–reaction geometry is the dynamic counterpart of the unbalanced distances of Section 14.5, especially the three-variable action A_κ and the balance-equation formula (14.41). It is best treated separately from the purely advective geometries above.

Definition 15.62 (WFR minimizing movement)**.** Let $f : \mathcal{M}_+(\mathbb{R}^d) \to \mathbb{R} \cup \{+\infty\}$ be an energy on positive finite measures and let WFR_κ be the Wasserstein–Fisher–Rao distance with growth scale $\kappa > 0$, defined by the dynamic balance-equation formulation (14.41) and its static equivalence in Proposition 14.26. The unbalanced JKO step is

$$\alpha^{(\ell+1)} \in \operatorname*{argmin}_{\alpha \in \mathcal{M}_+(\mathbb{R}^d)} \frac{1}{2\tau} \mathrm{WFR}_\kappa^2(\alpha^{(\ell)}, \alpha) + f(\alpha). \tag{15.48}$$

The continuous-time limit, when it exists, is called the WFR_κ gradient flow of f.

Formally, a smooth curve $\alpha_t = \rho_t \mathrm{d}x$ has tangent vectors of the form

$$\partial_t \rho_t + \operatorname{div}(\rho_t v_t) = g_t \rho_t,$$

with squared norm $\int (\|v_t\|^2 + \kappa^2 g_t^2)\rho_t \mathrm{d}x$, namely the integral of $A_\kappa(\rho_t, v_t, g_t)$. The vector field v_t accounts for displacement, while g_t is a Fisher–Rao growth rate. This is the infinitesimal version of the balance-equation action in Section 14.5.

Proposition 15.63 (Formal WFR gradient-flow PDE)**.** *Assume that f has a smooth first variation $\varphi_t = \delta f(\alpha_t)$ and that $\alpha_t = \rho_t \mathrm{d}x$ is smooth and positive. The formal WFR_κ gradient flow of f is*

$$\partial_t \rho_t = \operatorname{div}(\rho_t \nabla \varphi_t) - \frac{1}{\kappa^2} \varphi_t \rho_t. \tag{15.49}$$

Along such a smooth solution,

$$\frac{\mathrm{d}}{\mathrm{d}t} f(\alpha_t) = -\int \left(\|\nabla \varphi_t\|^2 + \frac{1}{\kappa^2} \varphi_t^2 \right) \rho_t \, \mathrm{d}x \leqslant 0. \tag{15.50}$$

Proof. For an infinitesimal tangent pair (v, g), the density variation is $\zeta = -\operatorname{div}(\rho v) + g\rho$. The first variation of f is

$$\int \varphi \, \zeta = \int \langle \nabla \varphi, v \rangle \rho \, \mathrm{d}x + \int \varphi g \rho \, \mathrm{d}x,$$

after integration by parts. For the inner product

$$\langle (v, g), (\tilde{v}, \tilde{g}) \rangle_\rho = \int \langle v, \tilde{v} \rangle \rho \, \mathrm{d}x + \kappa^2 \int g \tilde{g} \rho \, \mathrm{d}x,$$

the Riesz representative is $(\nabla \varphi, \varphi/\kappa^2)$. Steepest descent therefore uses $(v, g) = (-\nabla \varphi, -\varphi/\kappa^2)$, which gives (15.49). Substituting this pair in the first-variation formula gives

$$\frac{\mathrm{d}}{\mathrm{d}t} f(\alpha_t) = -\int \left(\|\nabla \varphi_t\|^2 + \frac{1}{\kappa^2} \varphi_t^2 \right) \rho_t \, \mathrm{d}x,$$

which is (15.50). □

The reaction term makes additive constants in the first variation meaningful: adding a constant to φ changes the global growth rate. This is not a defect but the infinitesimal signature of optimizing over positive finite measures, rather than on the probability simplex. The rigorous metric-space construction of Kantorovich–Fisher–Rao and WFR gradient flows is developed in [230, 143, 141]. Chizat's measure-optimization framework relates weight-changing gradient methods to mirror- and Bregman-type descents on positive measures [137, 136]. In machine learning, birth–death dynamics use the same reaction intuition to let a particle population reweight, remove and create neurons during training [455].

Example 15.64 (Relative entropy and birth–death relaxation). Let $\beta = \rho_\beta \mathrm{d}x$ and use the finite-measure relative entropy

$$f(\alpha) = \mathrm{KL}(\alpha|\beta) = \int \left(\rho \log \frac{\rho}{\rho_\beta} - \rho + \rho_\beta \right) \mathrm{d}x, \qquad \alpha = \rho \mathrm{d}x.$$

This is the Csiszár divergence on positive measures; it is nonnegative and is minimized, without

imposing a mass constraint, at $\alpha = \beta$. Its first variation is $\delta f(\alpha) = \log(\rho/\rho_\beta)$. The balanced $\mathcal{W}_2$ flow, when the total mass is fixed, is the Fokker–Planck equation

$$\partial_t \rho_t = \operatorname{div}\left(\rho_t \nabla \log \frac{\rho_t}{\rho_\beta}\right) = \Delta \rho_t - \operatorname{div}(\rho_t \nabla \log \rho_\beta),$$

whereas the WFR_κ flow is

$$\partial_t \rho_t = \Delta \rho_t - \operatorname{div}(\rho_t \nabla \log \rho_\beta) - \frac{1}{\kappa^2} \rho_t \log \frac{\rho_t}{\rho_\beta}. \tag{15.51}$$

The first two terms transport and diffuse mass, while the last term kills mass where $\rho_t > \rho_\beta$ and creates it where $\rho_t < \rho_\beta$.

Figure 15.15 contrasts this reaction-assisted relaxation with its mass-preserving Wasserstein counterpart.

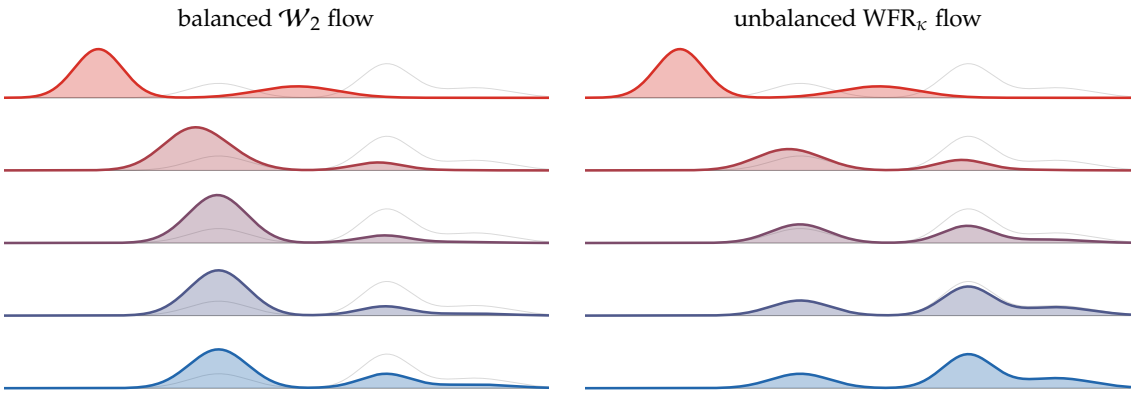


Figure 15.15: Balanced and unbalanced gradient flows of $\mathrm{KL}(\alpha|\beta)$ *for one-dimensional Gaussian mixtures.* In each panel the colored density stacks correspond to five increasing times, from red to blue, and the faint gray curve repeated on each row is the target mixture β. The balanced Wasserstein flow is the conservative Fokker–Planck equation, so mass must move through the low-density region between modes. The WFR_κ flow adds the birth–death reaction term in (15.51); it attenuates overrepresented regions and creates missing mass near target modes, reaching the target shape much faster.

15.8 Conditional Wasserstein Training of Infinite ResNets

Residual networks learn small residual updates around the identity [271]. This makes depth behave like time: very deep residual architectures can be interpreted as discretizations of differential equations, a viewpoint developed in stable architectures, neural ODEs and mean-field optimal-control models of deep learning [264, 355, 127, 200]. The conditional-OT formulation of Barboni, Peyré and Vialard [29] adds the width mean-field limit to this continuous-depth picture. This should be read alongside the broader conditional-transport literature reviewed in Section 11.6, especially fibered gradient flows [421], function-space conditional OT [279], and conditional OT flow-matching constructions [125, 300]. From the viewpoint of Section 15.5, it is a concrete instance of a generalized dynamic Wasserstein flow: the conditional Wasserstein distance has a geodesic dynamic structure obtained by integrating the usual Benamou–Brenier action over the conditioning variable. The point of the construction is to keep track of two independent symmetries: neurons can be relabelled inside each layer, but mass should not move from one depth to another. Conditional Wasserstein geometry implements exactly this rule. It transports neurons fiber by fiber in parameter space, while the depth variable remains fixed.

We follow the three limiting levels of the model. A finite ResNet is first a labelled particle system arranged by layers. Sending the width to infinity turns each layer into a probability law of neurons. Sending the depth step to zero then turns the sequence of layer laws into a conditional measure indexed by depth, on which training becomes a Wasserstein gradient flow.

Finite-depth finite-width ResNets. The starting point is the ordinary architecture: depth is discrete, width is finite, and each layer carries a labelled collection of neurons. Fix a parameter space $\Theta \subset \mathbb{R}^q$, a

residual feature $\psi : \Theta \times \mathbb{R}^d \to \mathbb{R}^d$, a data law ζ on input-label pairs (z, y), and a loss ℓ. With L layers, step size $\tau = 1/L$, widths n_r, and parameters $\theta_{r,i} \in \Theta$, a finite ResNet is the composition of the residual updates

$$z_{r+1} = z_r + \tau G_{r,n_r}(z_r), \qquad G_{r,n_r}(z) := \frac{1}{n_r} \sum_{i=1}^{n_r} \psi(\theta_{r,i}, z), \qquad z_0 = z, \qquad r = 0, \dots, L-1. \tag{15.52}$$

The associated population loss, with $n = (n_0, \dots, n_{L-1})$, is

$$f_{L,n}((\theta_{r,i})_{r,i}) = \int \ell(z_L^\theta(z), y) \mathrm{d}\zeta(z, y).$$

Equivalently, each layer carries an empirical neuron law

$$\alpha_r = \frac{1}{n_r} \sum_{i=1}^{n_r} \delta_{\theta_{r,i}}, \qquad G_{r,n_r}(z) = \int_\Theta \psi(\theta, z) \mathrm{d}\alpha_r(\theta).$$

Thus a finite ResNet is already a discrete conditional measure on depth and neuron space,

$$\alpha^{L,n}(\mathrm{d}s, \mathrm{d}\theta) = \frac{1}{L} \sum_{r=0}^{L-1} \delta_{s_r}(\mathrm{d}s) \alpha_r(\mathrm{d}\theta), \qquad s_r = r/L,$$

whose marginal on the depth variable is the discrete measure $\lambda_L = L^{-1} \sum_{r=0}^{L-1} \delta_{s_r}$.

Finite-depth mean-field limit. The first mean-field limit removes neuron labels inside each fixed layer. The residual depth grid stays discrete, but the widths tend to infinity. Each empirical law α_r is replaced by an arbitrary probability measure in $\mathcal{P}_2(\Theta)$, and each residual block becomes a mean-field residual block

$$z_{r+1} = z_r + \tau G_{\alpha_r}(z_r), \qquad z_0 = z, \qquad G_{\alpha_r}(z) := \int_\Theta \psi(\theta, z) \mathrm{d}\alpha_r(\theta).$$

The loss becomes a functional of the finite family of layer measures,

$$f_L(\alpha_0, \dots, \alpha_{L-1}) = \int \ell(z_L^\alpha(z), y) \mathrm{d}\zeta(z, y).$$

This is the finite-depth conditional Wasserstein setting with condition marginal λ_L: each fiber s_r carries a Wasserstein geometry on neuron laws.

Continuous-depth conditional geometry. The second limit turns the layer index into a conditioning variable. The model is no longer described by one neuron law, nor even by a finite list of them, but by a family of neuron laws indexed by depth. The state is therefore a conditional probability law

$$\alpha(\mathrm{d}s, \mathrm{d}\theta) = \alpha_s(\mathrm{d}\theta)\mathrm{d}s,$$

or, equivalently, a probability measure on $[0, 1] \times \Theta$ whose first marginal is Lebesgue measure. The general conditional Wasserstein distance is defined in Section 11.6. Here the condition is continuous depth, $s \in [0, 1]$, the condition law is $\lambda(\mathrm{d}s) = \mathrm{d}s$, and the fiber space is the neuron parameter space Θ. For two conditional neuron laws $\alpha(\mathrm{d}s, \mathrm{d}\theta) = \alpha_s(\mathrm{d}\theta)\mathrm{d}s$ and $\beta(\mathrm{d}s, \mathrm{d}\theta) = \beta_s(\mathrm{d}\theta)\mathrm{d}s$, the relevant distance is the depthwise quadratic Wasserstein metric

$$\mathcal{W}_{2,\lambda}^2(\alpha, \beta) = \int_0^1 \mathcal{W}_2^2(\alpha_s, \beta_s)\mathrm{d}s. \tag{15.53}$$

It compares neurons only at the same depth: transport in θ is allowed fiber by fiber, while mass is not transported between layers.

Infinite-depth mean-field ResNet. Once depth is continuous, the forward pass is a controlled ODE and the control is the conditional neuron law. A mean-field infinite-depth and infinite-width ResNet is described by a measurable family $s \mapsto \alpha_s \in \mathcal{P}_2(\Theta)$. Each slice is the neuron law of one infinitesimal residual block, and the feature $\psi(\theta, z)$ now acts as a vector field on the state variable. The forward pass is

$$\partial_s z_s = G_{\alpha_s}(z_s) := \int_\Theta \psi(\theta, z_s) \mathrm{d}\alpha_s(\theta), \qquad z_0 = z. \tag{15.54}$$

The population risk is

$$f_{\mathrm{Res}}(\alpha) = \int \ell(z_1^\alpha(z), y) \mathrm{d}\zeta(z, y).$$

Under the usual regularity assumptions ensuring well-posedness of the ODE, this defines a functional on the conditional Wasserstein space.

Conditional Wasserstein gradient flow. Training is now steepest descent for an objective on conditional laws. More generally, let (S, λ) be a probability space of conditions and let f be a differentiable functional on measures $\alpha(\mathrm{d}s, \mathrm{d}\theta) = \alpha_s(\mathrm{d}\theta)\lambda(\mathrm{d}s)$ with fixed first marginal λ. For signed perturbations $\xi(\mathrm{d}s, \mathrm{d}\theta) = \xi_s(\mathrm{d}\theta)\lambda(\mathrm{d}s)$ satisfying $\int_\Theta \mathrm{d}\xi_s = 0$ for λ-a.e. s, we use the fiberwise first-variation convention

$$\frac{\mathrm{d}}{\mathrm{d}\varepsilon} f(\alpha + \varepsilon\xi)\Big|_{\varepsilon=0} = \int_S \int_\Theta \frac{\delta f}{\delta \alpha_s}(\alpha)(\theta) \mathrm{d}\xi_s(\theta) \mathrm{d}\lambda(s).$$

The formal conditional Wasserstein gradient flow is the family of continuity equations

$$\partial_t \alpha_{t,s} + \mathrm{div}_\theta(\alpha_{t,s} v_{t,s}) = 0, \qquad v_{t,s}(\theta) = -\nabla_\theta \frac{\delta f}{\delta \alpha_s}(\alpha_t)(\theta), \tag{15.55}$$

for λ-a.e. condition s, whenever the first variation is differentiable in the parameter variable. In the ResNet case one takes $f = f_{\mathrm{Res}}$ and $S = [0, 1]$. The equation is fiberwise in its transport geometry, but not decoupled as a training dynamics: changing α_s affects the terminal loss through the forward ODE and the corresponding adjoint equation. Thus each depth performs a Wasserstein steepest descent in its own neuron space, while all depths interact through the state trajectory.

Particle discretization and layerwise matching. The finite network is recovered by discretizing both the condition variable and the fiber measures. With the discrete depth marginal λ_L, finite-depth finite-width ResNets are particle approximations of the conditional flow above. If two networks have the same depth grid and the same widths, and if their empirical layer laws are

$$\alpha_r = \frac{1}{n_r} \sum_{i=1}^{n_r} \delta_{\theta_{r,i}}, \qquad \beta_r = \frac{1}{n_r} \sum_{i=1}^{n_r} \delta_{\theta'_{r,i}},$$

then their squared conditional distance is

$$\mathcal{W}^2_{2,\lambda_L}(\alpha^{L,n}, \beta^{L,n}) = \frac{1}{L} \sum_{r=0}^{L-1} \mathcal{W}_2^2(\alpha_r, \beta_r).$$

Keeping neuron labels fixed gives the labelled parameter distance

$$\frac{1}{L} \sum_{r=0}^{L-1} \frac{1}{n_r} \sum_{i=1}^{n_r} \|\theta_{r,i} - \theta'_{r,i}\|^2,$$

which corresponds to a particular feasible coupling and therefore upper bounds the squared conditional Wasserstein distance. Optimizing over permutations inside each equal-weight layer gives the Wasserstein matching term above. In this sense, the shallow mean-field flow is the one-fiber case, whereas infinitely deep ResNets require a continuum of Wasserstein flows indexed by depth and coupled through the network state.

15.9 Second-Order Momentum Flows

The minimizing-movement viewpoint is intrinsically first order: a JKO step produces the next measure by trading off proximity to the previous measure and decrease of the energy. Momentum adds a memory of velocity. The state is therefore no longer only a measure, but a measure together with a tangent, or equivalently phase-space, variable. This makes a direct implicit JKO construction less natural than for first-order flows. We instead use an explicit construction, in the same spirit as optimization and machine-learning practice: momentum reduces zig-zagging across stiff directions, helps iterates keep a coherent direction through shallow valleys, and can accelerate convergence when the geometry is favorable [441, 512]. The goal is to write this inertial idea in a form compatible with Wasserstein particle flows.

Finite-dimensional momentum. Let $F : \mathbb{R}^N \to \mathbb{R}$ be a smooth finite-dimensional objective. Polyak's heavy-ball method [439] augments gradient descent with a velocity variable. With step size $h > 0$ and damping $\lambda_{\mathrm{fric}} > 0$, one convenient velocity form is

$$S^{(\ell+1)} = (1 - \lambda_{\mathrm{fric}} h) S^{(\ell)} - h \nabla F(X^{(\ell)}), \qquad X^{(\ell+1)} = X^{(\ell)} + h S^{(\ell+1)}. \tag{15.56}$$

If $S^{(\ell)}$ approximates $\dot{X}(\ell h)$, this is a semi-implicit, or symplectic-Euler, discretization of the first-order system: the velocity update is explicit, while the position uses the newly updated velocity. The limiting system is

$$\dot{X}(t) = S(t), \qquad \dot{S}(t) = -\lambda_{\mathrm{fric}} S(t) - \nabla F(X(t)), \tag{15.57}$$

or equivalently the damped second-order equation

$$\ddot{X}(t) + \lambda_{\mathrm{fric}} \dot{X}(t) + \nabla F(X(t)) = 0. \tag{15.58}$$

Nesterov's accelerated method [400] uses a closely related extrapolation but evaluates the gradient at a look-ahead point,

$$Y^{(\ell)} = X^{(\ell)} + \theta^{(\ell)} (X^{(\ell)} - X^{(\ell-1)}), \qquad X^{(\ell+1)} = Y^{(\ell)} - h \nabla F(Y^{(\ell)}), \tag{15.59}$$

with $\theta^{(\ell)} = (\ell - 1)/(\ell + r - 1)$ for the classical convex schedule $r = 3$. Its scaling limit is the time-dependent damping equation

$$\ddot{X}(t) + \frac{r}{t} \dot{X}(t) + \nabla F(X(t)) = 0. \tag{15.60}$$

This ODE interpretation is due to Su, Boyd and Candès [510]; see also the variational perspective of Wibisono, Wilson and Jordan [550], and the inertial dynamics with Hessian-driven damping studied by Attouch, Peypouquet and Redont [22]. High-resolution ODE limits refine this picture by keeping terms that distinguish heavy-ball and Nesterov dynamics beyond their leading continuous-time behavior [486]. For the Wasserstein extension below, the autonomous heavy-ball form (15.57) is the cleanest template.

Empirical Wasserstein lift. We now take a functional f on measures and lift it to particle configurations. For $X = (x_1, \dots, x_n) \in (\mathbb{R}^d)^n$, write

$$\alpha_X := \frac{1}{n} \sum_{i=1}^{n} \delta_{x_i}, \qquad F(X) := f(\alpha_X). \tag{15.61}$$

This is the same empirical lift as in (15.5), and is the particle viewpoint underlying mean-field analyses of wide neural networks and Wasserstein gradient methods [138, 373, 456]. The important point is that the configuration space is not used with the plain Euclidean metric, but with the empirical metric

$$\|\dot{X}\|_n^2 := \frac{1}{n} \sum_{i=1}^{n} \|\dot{x}_i\|^2,$$

which is the metric induced by $\mathcal{W}_2$ on equally weighted empirical measures with fixed labels. Moving only x_i to $x_i + \varepsilon \xi$ gives

$$\frac{\mathrm{d}}{\mathrm{d}\varepsilon}\Big|_{\varepsilon=0} F(x_1, \dots, x_i + \varepsilon \xi, \dots, x_n) = \frac{1}{n} \langle \nabla \delta f(\alpha_X)(x_i),\, \xi \rangle.$$

Thus, if $\nabla^{(n)}F$ denotes the gradient for the metric $\|\cdot\|_n$, then

$$(\nabla^{(n)}F(X))_i = n\nabla_{x_i}F(X) = \nabla\delta f(\alpha_X)(x_i) = \mathbb{W}f(\alpha_X)(x_i). \tag{15.62}$$

The first-order Wasserstein particle descent is therefore $\dot{x}_i = v_{\alpha_X}(x_i)$ with

$$v_\alpha(x) \coloneqq -\mathbb{W}f(\alpha)(x) = -\nabla\delta f(\alpha)(x). \tag{15.63}$$

Applying the heavy-ball equation to the empirical metric gives the second-order Wasserstein momentum system

$$\dot{x}_i(t) = s_i(t), \qquad \dot{s}_i(t) = -\lambda_{\text{fric}}s_i(t) + v_{\alpha_t}(x_i(t)), \qquad \alpha_t = \frac{1}{n}\sum_{i=1}^n \delta_{x_i(t)}. \tag{15.64}$$

Equivalently,

$$\ddot{x}_i(t) + \lambda_{\text{fric}}\dot{x}_i(t) = v_{\alpha_t}(x_i(t)) = -\mathbb{W}f(\alpha_t)(x_i(t)).$$

This is an explicit inertial particle method for the measure energy f: the Wasserstein steepest-descent field acts as an acceleration rather than as an instantaneous velocity. The statement is deliberately formal at this level; rigorous mean-field limits require stability and regularity assumptions on v_α, as in propagation-of-chaos arguments for classical Vlasov dynamics and in mean-field analyses of interacting learning systems.

Proposition 15.65 (Discrete mechanical energy dissipation)**.** *Assume that f and the particle solution are smooth enough to justify differentiating the identities above along* (15.64)*. Define*

$$\mathcal{H}_n(X,S) \coloneqq F(X) + \frac{1}{2n}\sum_{i=1}^n \|s_i\|^2, \qquad S = (s_1,\dots,s_n).$$

Then along (15.64)*,*

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{H}_n(X(t),S(t)) = -\frac{\lambda_{\text{fric}}}{n}\sum_{i=1}^n \|s_i(t)\|^2 \leqslant 0. \tag{15.65}$$

Proof. By (15.62) and the definition $v_\alpha = -\mathbb{W}f(\alpha)$, one has $\nabla_{x_i}F(X) = -v_{\alpha_X}(x_i)/n$. Hence

$$\frac{\mathrm{d}}{\mathrm{d}t}F(X(t)) = \sum_{i=1}^n \langle \nabla_{x_i}F(X(t)),\, s_i(t)\rangle = -\frac{1}{n}\sum_{i=1}^n \langle v_{\alpha_t}(x_i(t)),\, s_i(t)\rangle.$$

On the other hand,

$$\frac{\mathrm{d}}{\mathrm{d}t}\frac{1}{2n}\sum_i \|s_i(t)\|^2 = \frac{1}{n}\sum_i \langle s_i(t),\, -\lambda_{\text{fric}}s_i(t) + v_{\alpha_t}(x_i(t))\rangle.$$

The force terms cancel, leaving (15.65). □

The identity says that friction dissipates exactly the kinetic part of the mechanical energy. In the undamped case $\lambda_{\text{fric}} = 0$, the same computation gives conservation of $\mathcal{H}_n$.

Phase-space formulation. Because momentum is part of the state, the mean-field object is not merely a measure on positions. It is a law on phase space (x,s), whose spatial marginal is the current distribution of particles. Here s is a Lagrangian particle velocity; it should not be confused with the Eulerian momentum measure $\omega = \alpha v$ used in the Benamou–Brenier convex formulation (14.12); when $\alpha = \rho\,\mathrm{d}x$, this measure has density $m = \rho v$. The particle ODE then becomes a transport equation in phase space. This is the deterministic analogue of the kinetic mean-field viewpoint used for interacting particle systems; in learning, heavy-ball mean-field limits have recently been used to analyze wide networks trained with momentum [551].

Proposition 15.66 (Phase-space Liouville formulation)**.** *Let v_α be smooth enough that* (15.64) *has a classical solution. Define the empirical phase-space measure*

$$\eta_t^n \coloneqq \frac{1}{n}\sum_{i=1}^n \delta_{(x_i(t),s_i(t))} \in \mathcal{P}(\mathbb{R}^d_x \times \mathbb{R}^d_s), \qquad \alpha_t^n = (\pi_x)_\sharp \eta_t^n.$$

Then η_t^n solves, in the sense of distributions,

$$\partial_t \eta_t^n + \nabla_x \cdot (s\,\eta_t^n) + \nabla_s \cdot \big((-\lambda_{\mathrm{fric}} s + v_{\alpha_t^n}(x))\,\eta_t^n\big) = 0. \tag{15.66}$$

If, as $n \to \infty$, the empirical phase-space laws converge to a smooth density $\eta_t(x,s)$ and the nonlinear fields converge accordingly, the limiting kinetic equation is

$$\partial_t \eta_t + \nabla_x \cdot (s\,\eta_t) + \nabla_s \cdot \big((-\lambda_{\mathrm{fric}} s + v_{\alpha_t}(x))\,\eta_t\big) = 0, \qquad \alpha_t = (\pi_x)_\sharp \eta_t. \tag{15.67}$$

Proof. Let $\varphi \in C_c^\infty(\mathbb{R}_x^d \times \mathbb{R}_s^d)$. Along the particle system,

$$\frac{\mathrm{d}}{\mathrm{d}t} \int \varphi \,\mathrm{d}\eta_t^n = \frac{1}{n} \sum_{i=1}^n \left[\langle \nabla_x \varphi(x_i, s_i),\, s_i \rangle + \langle \nabla_s \varphi(x_i, s_i),\, -\lambda_{\mathrm{fric}} s_i + v_{\alpha_t^n}(x_i) \rangle \right].$$

Rewriting the right-hand side as an integral against η_t^n and integrating by parts gives (15.66). Passing formally to the mean-field limit gives (15.67). □

Remark 15.67 (Newtonian limit). If damping is removed, or if inertia dominates friction after rescaling, (15.64) becomes

$$\ddot{x}_i(t) = v_{\alpha_t}(x_i(t)).$$

This is Newton's law with acceleration field v_{α_t}. Thus first-order Wasserstein gradient flow reads the force as a velocity, while second-order momentum flow reads the same force as an acceleration. The price is that the state is no longer just the spatial measure α_t; it is the phase-space law of position and velocity. In this limit the flow is Hamiltonian rather than dissipative whenever the force comes from a smooth interaction energy and $\lambda_{\mathrm{fric}} = 0$.

Example 15.68 (Quadratic energies and gravitational attraction). The quadratic-plus-linear energies

$$f(\alpha) = \frac{1}{2} \iint k(x,y)\,\mathrm{d}\alpha(x)\mathrm{d}\alpha(y) + \int V(x)\,\mathrm{d}\alpha(x), \qquad k(x,y) = k(y,x),$$

are the interaction energies already discussed around (15.9), up to the harmless convention of inserting the factor $1/2$. The force computed there becomes, in the present inertial setting, an acceleration:

$$v_\alpha(x) = -\int \nabla_x k(x,y)\,\mathrm{d}\alpha(y) - \nabla V(x).$$

For the attractive Newtonian kernel in three dimensions, $k(x,y) = -G/\|x-y\|$, understood with a short-distance regularization if needed, this gives

$$v_\alpha(x) = G \int \frac{y-x}{\|x-y\|^3}\,\mathrm{d}\alpha(y) - \nabla V(x),$$

the usual mean-field gravitational acceleration. The phase-space equation (15.67) is then a damped Vlasov-type equation. This kinetic viewpoint is classical in statistical physics: Boltzmann introduced the description of gases through phase-space distributions, the collisionless mean-field limit gives Liouville–Vlasov equations, and adding collision operators leads to the Boltzmann equation [120, 537, 78].

The numerical illustration below uses the energy-distance kernel

$$k(x,y) = -\|x-y\|, \qquad f(\alpha) = \mathrm{MMD}_k^2(\alpha,\beta) = \iint k(x,y)\,\mathrm{d}(\alpha-\beta)(x)\mathrm{d}(\alpha-\beta)(y).$$

This is the usual squared energy distance between α and β. Away from the diagonal, its Wasserstein velocity is

$$v_\alpha(x) = 2 \int \frac{x-y}{\|x-y\|}\,\mathrm{d}(\alpha-\beta)(y).$$

For numerical stability, Figure 15.16 uses the smoothed distance $\sqrt{\|x-y\|^2+\delta^2}$ with a small δ. It compares the first-order flow with the undamped Newton lift $\ddot{x} = v_{\alpha_t}(x) = -\mathbb{W} f(\alpha_t)(x)$, initialized with

zero velocity. The initial measure is a single Gaussian located to the left of a two-Gaussian target mixture. Both the transported measure and the target are discretized by large empirical clouds, and the smooth density views are KDE renderings of these particles. The trajectory panels display only a representative subset of initial particles selected by farthest-point sampling.

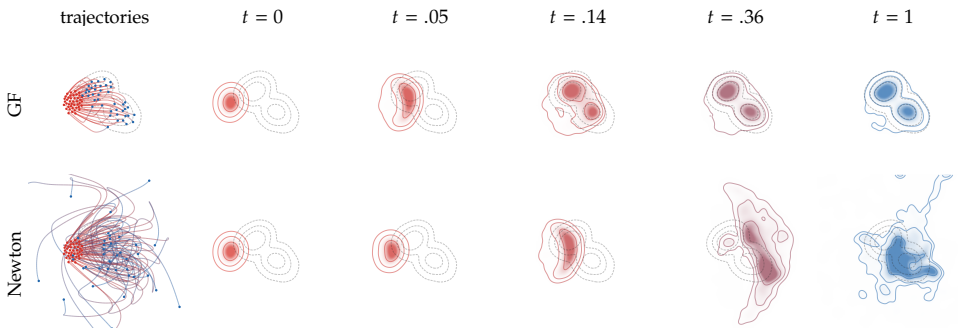


Figure 15.16: First-order and Newton Wasserstein particle flows for the squared MMD with the energy-distance kernel $k(x, y) = -\|x - y\|$. The first row follows the overdamped Wasserstein gradient flow. The second row follows the infinite-momentum Newton dynamics $\ddot{x} = -\mathbb{W} f(\alpha_t)(x)$, so the same Wasserstein force is read as an acceleration rather than a velocity. Both rows use the same large empirical source cloud and the same empirical two-Gaussian target cloud. Dashed gray contours show a KDE of the target particles, colored panels show KDEs of the evolving transported particles, and the left panels show farthest-point-subsampled trajectories.

Entropy without an empirical energy. The previous interaction example has a genuine value on empirical measures: once the particles are known, the double sum approximates the interaction energy. Entropy gives the complementary situation. Its Wasserstein gradient flow is the simplest diffusion equation, but the entropy itself is finite only on absolutely continuous measures. A particle method must therefore introduce a density, or score, closure before the inertial template can be applied.

Example 15.69 (Entropy-driven inertial flow). Let

$$f(\alpha) = \begin{cases} \displaystyle\int_{\mathbb{R}^d} \rho(x) \log \rho(x)\, \mathrm{d}x, & \text{if } \alpha = \rho \mathrm{d}x, \\ +\infty, & \text{otherwise.} \end{cases}$$

Thus $f(\alpha_X) = +\infty$ for an empirical measure $\alpha_X = n^{-1} \sum_i \delta_{x_i}$, and the finite-dimensional lift $F(X) = f(\alpha_X)$ is not meaningful. For $\alpha = \rho \mathrm{d}x$ with a smooth positive density, however,

$$\delta f(\alpha)(x) = \log \rho(x) + 1, \qquad v_\alpha(x) = -\nabla \log \rho(x).$$

The first-order Wasserstein gradient flow is therefore the heat equation $\partial_t \rho_t = \Delta \rho_t$. If the spatial marginal remains smooth and positive, the corresponding formal second-order phase-space equation is

$$\partial_t \eta_t + \nabla_x \cdot (s \eta_t) + \nabla_s \cdot \big((-\lambda_{\mathrm{fric}} s - \nabla \log \rho_t(x)) \eta_t\big) = 0, \qquad \rho_t(x) = \int_{\mathbb{R}^d} \eta_t(x, s)\, \mathrm{d}s. \tag{15.68}$$

This is a kinetic, inertial version of entropy diffusion: the entropy score acts as an acceleration rather than as an instantaneous velocity.

For the numerical illustration, we add the quadratic confinement $V(x) = \|x\|^2/2$. In one dimension the overdamped equation becomes the Ornstein–Uhlenbeck Fokker–Planck equation

$$\partial_t \rho_t = \partial_{xx} \rho_t + \partial_x (x \rho_t),$$

whose stationary law is the centered Gaussian $\mathcal{N}(0, 1)$. The Newton lift is discretized directly in phase space: for a density $\eta_t(x, s)$ with spatial marginal $\rho_t(x) = \int \eta_t(x, s)\, \mathrm{d}s$, it reads

$$\partial_t \eta_t + \partial_x (s \eta_t) + \partial_s \big((-\partial_x \log \rho_t(x) - x) \eta_t\big) = 0.$$

Both equations are solved by finite differences on fixed grids, with a mild grid smoothing only to evaluate the logarithmic derivative in the Newton row. The first row of Figure 15.17 shows only the

spatial density of the Wasserstein gradient flow. The second and third rows show, for the Newton equation, respectively the spatial marginal and the full phase-space density; the latter uses white for zero mass and black for high mass.

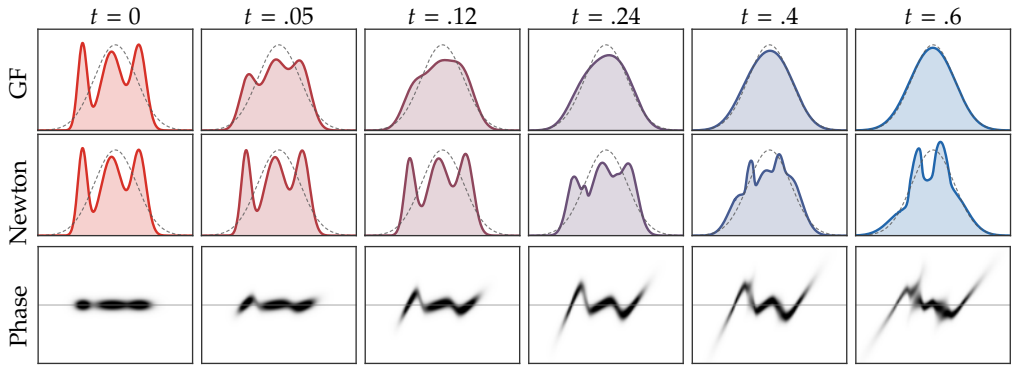


Figure 15.17: Finite-difference entropy and inertial entropy flows in one dimension. The initial spatial density is a three-Gaussian mixture with different widths and overlapping components. The top row solves the confined entropy Wasserstein gradient flow, i.e. the OU/Fokker–Planck equation, whose stationary density is the dashed centered Gaussian. The two lower rows solve the undamped Newton lift on a grid in (x, s), initialized with a narrow centered speed distribution: its spatial marginal is shown in color and the full phase-space density below in grayscale.

This chapter has moved from first-order minimizing movements to several increasingly structured dynamics. The common thread is that an energy alone does not determine an evolution: the metric or action selects the admissible tangent variables and converts the first variation into a velocity, jump flux, growth rate, fiberwise transport, or acceleration. Geodesic convexity and slope inequalities then distinguish the cases with quantitative convergence guarantees from the mean-field and inertial models that require additional problem-specific analysis.

CHAPTER 16

Generative Models via Transportation

The preceding gradient-flow calculus is variational. Modern machine-learning models often use the same transportation language more broadly: one may prescribe an interpolation and regress its velocity, fit a one-step generator to a descent field, or view network depth as a continuous transport of token measures. The examples below separate what is genuinely a Wasserstein gradient flow from what is a transportation dynamics with a useful geometric interpretation.

16.1 Generative Models via Flow Matching

Flow matching constructs a generative map by learning the velocity field of an interpolation. The key computational insight is that a constrained continuity-equation problem can be trained by an unconstrained regression.

Generative models aim to build a transportation map T between a reference distribution α (typically an isotropic Gaussian) and the target data distribution β. Since such reference measures are non-atomic, a measurable map with $T_\sharp \alpha = \beta$ exists on standard Borel spaces, for instance by identifying both probability spaces with the unit interval and using a quantile-type rearrangement. This abstract existence statement is much weaker than having an explicit and numerically stable construction of T. Optimal transport is one approach to achieving this, but it is computationally expensive and raises questions about how to estimate it from samples. A different route is to prescribe an interpolation between noise and data, learn its velocity, and obtain T by integrating a time-dependent vector field v_t. This point of view sits at the meeting point of two literatures, surveyed from a transport perspective in [426]. The diffusion branch builds on score matching [285], denoising score matching [540], nonequilibrium noising chains [494], denoising diffusion probabilistic models [277], score-based generative modeling [497], and the continuous-time score-SDE/probability-flow formulation [498]. The deterministic regression branch was introduced, essentially in parallel, under three closely related names: flow matching [343], rectified flow [347], and stochastic interpolants [3]. In all three cases, the computational object is a velocity field whose regression loss avoids simulating the learned ODE during training. This vector field v_t is obtained by constructing an interpolation α_t and then finding v_t using the least-squares formula (14.7). As we will explain, for a specific class of interpolation (obtained by a parametric push-forward), this v_t can be obtained by avoiding explicitly inverting a Laplacian and instead computing a simple conditional expectation. This conditional expectation can itself be estimated by solving another least-squares problem, but this time unconstrained, making the estimation feasible from finite samples of α and β.

Deterministic interpolants. The word "stochastic" can hide two different levels of randomness. We first use the simpler one: after drawing a latent variable $U \sim \pi$, the path $t \mapsto I_t(U)$ is deterministic and differentiable. The randomness only comes from the initial draw of U; after taking the push-forward law, $\alpha_t = (I_t)_\sharp \pi$ is a deterministic curve of measures and obeys an ordinary continuity equation. This is the setting behind the stochastic-interpolant construction recalled in Definition 16.4, and behind the flow-matching and rectified-flow regressions below. Genuine temporal noise, where the path itself has Brownian fluctuations, is different and is discussed in Remark 16.5.

The endpoint data, their coupling, and the path joining each coupled pair determine the basic object used below.

Definition 16.1 (Deterministic interpolant)**.** Let $\pi \in \mathcal{U}(\alpha_0, \alpha_1)$. A deterministic interpolant is a measurable family $I_t : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d$, differentiable in t, satisfying

$$I_0(x, y) = x, \qquad I_1(x, y) = y.$$

It induces $\alpha_t = (I_t)_\sharp \pi$ and has pathwise velocity $\partial_t I_t(x, y)$.

More generally, α_t can be obtained by pushing a latent distribution $\pi \in \mathcal{P}(\mathbb{R}^{d'})$ through a time-dependent

map $I_t : \mathbb{R}^{d'} \to \mathbb{R}^d$; the latent dimension d' may be larger than the data dimension d, i.e.

$$\forall t \in [0,1], \quad \alpha_t := (I_t)_\sharp \pi. \tag{16.1}$$

The basic two-endpoint construction already covers most flow-matching paths used in practice.

Example 16.2 (Linear two-endpoint deterministic interpolants). Set $d' = 2d$, write $(x, y) \in \mathbb{R}^d \times \mathbb{R}^d$, and choose $I_0(x, y) = x$ and $I_1(x, y) = y$. If π has marginals (α_0, α_1), then $\alpha_t = (I_t)_\sharp \pi$ interpolates between the two endpoint laws. The simplest choices are the independent, or trivial, coupling $\pi = \alpha_0 \otimes \alpha_1$ and the straight path

$$I_t(x, y) = (1 - t)x + ty.$$

With this linear path and an arbitrary coupling π, the regression below is the common core of flow matching and rectified flow: Lipman et al. emphasize conditional probability paths and simulation-free training of continuous normalizing flows, while rectified flow emphasizes straight couplings, reflow, and the possibility of reducing transport costs and discretization error [343, 347]. More complex constructions are possible when sampling from π remains simple. Static auxiliary randomness is still handled by enlarging the latent variable, while Brownian noise leads to the diffusion correction described below; this is the broader stochastic-interpolant viewpoint connecting deterministic flows, probability-flow ODEs and diffusion SDEs [3].

If $\pi = \alpha \otimes \beta$ and $\alpha = \frac{1}{n}\sum_i \delta_{x_i}$, $\beta = \frac{1}{m}\sum_j \delta_{y_j}$, then α_t consists of $n \times m$ Dirac masses

$$\alpha_t = \frac{1}{nm} \sum_{i,j} \delta_{I_t(x_i, y_j)}.$$

If $\pi = (\mathrm{Id}, T)_\sharp \alpha$ is a Brenier-type coupling, then $\alpha_t = ((1 - t)\mathrm{Id} + tT)_\sharp \alpha$ is the so-called McCann OT interpolation.

Figure 16.1 separates the effect of the endpoint coupling from the effect of the path joining each paired endpoint.

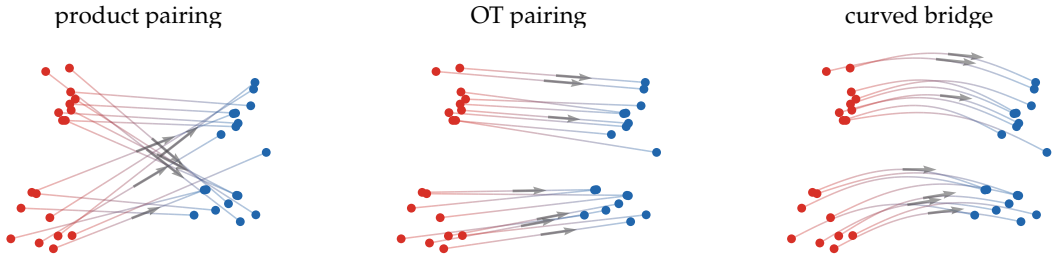


Figure 16.1: Flow matching interpolants between the same empirical source and target measures. A product-style random pairing produces crossing paths, an OT pairing gives direct displacement rays, and a curved bridge changes the path geometry while keeping the same endpoints. Gray arrows mark representative midpoint velocities $\partial_t I_t$.

Flow matching formula. This interpolation is not directly useful for sampling from β, but it can be used to define a flow field v_t so that the Eulerian advection equation (14.2) holds. This flow field is computed by solving an unconstrained least-squares problem, or equivalently, it is a conditional expectation.

Proposition 16.3 (Flow matching vector field)**.** *For each fixed t, assume $\partial_t I_t \in L^2(\pi; \mathbb{R}^d)$. Consider the flow-matching problem over measurable fields $v_t : \mathbb{R}^d \to \mathbb{R}^d$*

$$\min_{v_t} \int_{\mathbb{R}^{d'}} \|v_t(I_t(u)) - [\partial_t I_t](u)\|^2 \, \mathrm{d}\pi(u). \tag{16.2}$$

Its minimizer is characterized α_t-almost everywhere by the conditional expectation

$$v_t(z) = \mathbb{E}_{u \sim \pi}\big([\partial_t I_t](u) \,\big|\, z = I_t(u)\big). \tag{16.3}$$

Then the pair (α_t, v_t) satisfies the continuity equation (14.2).

Proof. We first recall the two equivalent ways of writing the interpolated measure. Formally, one may write

$$\alpha_t(z) = \int_{\mathbb{R}^{d'}} \delta(z - I_t(u))\, \mathrm{d}\pi(u),$$

while the rigorous meaning is that, for every smooth test function φ,

$$\int_{\mathbb{R}^d} \varphi(z)\, \mathrm{d}\alpha_t(z) = \int_{\mathbb{R}^{d'}} \varphi(I_t(u))\, \mathrm{d}\pi(u). \tag{16.4}$$

The minimizer in (16.2) is the orthogonal projection in $L^2(\pi; \mathbb{R}^d)$ of the latent velocity $\partial_t I_t(u)$ onto the closed subspace of functions that depend on u only through $I_t(u)$. This projection is the conditional expectation (16.3). Formally, this can be read as

$$v_t(z) = \frac{1}{\alpha_t(z)} \int_{\mathbb{R}^{d'}} \delta(z - I_t(u))[\partial_t I_t](u)\, \mathrm{d}\pi(u),$$

and rigorously it means that, for every smooth test vector field m,

$$\int \langle m(z),\, v_t(z)\rangle\, \mathrm{d}\alpha_t(z) = \int \langle m(I_t(u)),\, [\partial_t I_t](u)\rangle\, \mathrm{d}\pi(u). \tag{16.5}$$

We now prove that this field transports the curve $(\alpha_t)_t$. The weak form of $\partial_t \alpha_t + \operatorname{div}(\alpha_t v_t) = 0$ is that, for every smooth scalar test function φ,

$$\frac{\mathrm{d}}{\mathrm{d}t} \int \varphi(z)\, \mathrm{d}\alpha_t(z) - \int \langle v_t(z),\, \nabla\varphi(z)\rangle\, \mathrm{d}\alpha_t(z) = 0. \tag{16.6}$$

Using (16.4) and differentiating under the integral sign gives

$$\frac{\mathrm{d}}{\mathrm{d}t} \int \varphi(z)\mathrm{d}\alpha_t(z) = \int \langle \nabla\varphi(I_t(u)),\, [\partial_t I_t](u)\rangle \mathrm{d}\pi(u). \tag{16.7}$$

On the other hand, applying (16.5) with $m = \nabla\varphi$ gives

$$\int \langle v_t(z),\, \nabla\varphi(z)\rangle\, \mathrm{d}\alpha_t(z) = \int \langle \nabla\varphi(I_t(u)),\, [\partial_t I_t](u)\rangle \mathrm{d}\pi(u). \tag{16.8}$$

Comparing (16.7) and (16.8) yields (16.6), which is the desired continuity equation. □

The conditional expectation in (16.3) has a simple measure-theoretic meaning. Let $\alpha_t = (I_t)_\sharp \pi$ and define the vector-valued flux measure ω_t on $\mathbb{R}^d$ by

$$\int_{\mathbb{R}^d} \langle \psi(z),\, \mathrm{d}\omega_t(z)\rangle := \int_{\mathbb{R}^{d'}} \langle \psi(I_t(u)),\, [\partial_t I_t](u)\rangle \mathrm{d}\pi(u)$$

for every bounded continuous vector field ψ. Since $\alpha_t(A) = 0$ implies $\pi(I_t^{-1}(A)) = 0$, one has $\omega_t \ll \alpha_t$. The Radon–Nikodym decomposition of ω_t with respect to α_t is therefore

$$\mathrm{d}\omega_t(z) = v_t(z)\mathrm{d}\alpha_t(z), \qquad v_t = \frac{\mathrm{d}\omega_t}{\mathrm{d}\alpha_t}.$$

In the language of Lebesgue decomposition, ω_t has only an absolutely continuous part with respect to α_t and no singular part; the conditional expectation is precisely its density. This agrees with the flux notation used in the dynamic formulation of Section 14.2. Equivalently, disintegrating π with respect to the map I_t gives $\pi(\mathrm{d}u) = \pi_{t,z}(\mathrm{d}u)\alpha_t(\mathrm{d}z)$, where $\pi_{t,z}$ is supported on the fiber $\{u : I_t(u) = z\}$, and

$$v_t(z) = \int_{\{I_t(u)=z\}} [\partial_t I_t](u)\mathrm{d}\pi_{t,z}(u).$$

Thus the solution of (16.2) is the conditional expectation of the velocities $\partial_t I_t$: intuitively, $v_t(z)$ is the average velocity of all trajectories passing through z. Numerically, $(x, t) \to v_t(x)$ can be parameterized by a neural network (e.g., a U-Net for vision tasks) and estimated using stochastic gradient descent on the objective in (16.2).

For the exact field v_t, integrating the ODE $\dot{x} = v_t(x)$ defines a transport map T_t. If v_t is regular enough, or more generally if the continuity equation has a unique solution for this velocity, then $(T_t)_\sharp \alpha_0 = \alpha_t$. Thus the same interpolation as (16.1) is represented by a deterministic flow rather than by the original coupling. The sampling procedure first draws $X_0 \sim \alpha$ and then integrates the ODE $\dot{X}_t = v_t(X_t)$ starting with $X_{t=0} = X_0$. In the ideal exact-field limit, the resulting $X_{t=1}$ is distributed according to $\alpha_1 = \beta$.

Static-noise stochastic interpolants. In the terminology of Albergo–Boffi–Vanden-Eijnden [3], randomness can enter through a static auxiliary variable without turning the path into an SDE.

Definition 16.4 (Stochastic interpolant)**.** Let $(X_0, X_1) \sim \pi \in \mathcal{U}(\alpha_0, \alpha_1)$, and let Z be an auxiliary random variable independent of (X_0, X_1). A stochastic interpolant is

$$X_t = I_t(X_0, X_1, Z),$$

where $t \mapsto I_t(x_0, x_1, z)$ is differentiable and

$$I_0(x_0, x_1, z) = x_0, \qquad I_1(x_0, x_1, z) = x_1.$$

Its marginal curve is $\alpha_t = \mathrm{Law}(X_t)$.

A typical spatially linear example is

$$X_t = a(t)X_0 + b(t)X_1 + \gamma(t)Z, \qquad \gamma(0) = \gamma(1) = 0.$$

The noise Z is static: conditionally on (X_0, X_1, Z), the path $t \mapsto X_t$ is differentiable. Thus this construction is exactly the previous push-forward framework with latent variable $u = (X_0, X_1, Z)$, law $\pi = \mathrm{Law}(X_0, X_1, Z)$, and interpolant given by the displayed I_t. Its Eulerian velocity is therefore

$$v_t(x) = \mathbb{E}\big[\partial_t I_t(X_0, X_1, Z) \mid X_t = x\big],$$

and the interpolant density satisfies the continuity equation. The associated SDEs in the stochastic-interpolant framework are alternative sampling dynamics having the same one-time marginals; they are not the definition of the interpolant itself.

Remark 16.5 (Brownian realizations of interpolant marginals). One can also represent an interpolating marginal curve by Brownian-in-time dynamics. This is a different construction from the static-noise bridge of Definition 16.4. Let Z_t solve the Itô equation

$$\mathrm{d}Z_t = r_t(U, Z_t)\mathrm{d}t + \Sigma_t(U, Z_t)\mathrm{d}B_t, \qquad \alpha_t = \mathrm{Law}(Z_t),$$

where $U \sim \pi$ is static and B_t is Brownian motion. Define the Eulerian drift and diffusion matrix by conditioning on the observed state,

$$v_t(z) = \mathbb{E}\big[r_t(U, Z_t) \mid Z_t = z\big], \qquad D_t(z) = \mathbb{E}\big[\Sigma_t(U, Z_t)\Sigma_t(U, Z_t)^\top \mid Z_t = z\big].$$

Then, for smooth test functions φ,

$$\frac{\mathrm{d}}{\mathrm{d}t}\int \varphi \mathrm{d}\alpha_t = \int \langle \nabla\varphi,\, v_t\rangle \mathrm{d}\alpha_t + \frac{1}{2}\int \mathrm{tr}\big(D_t\nabla^2\varphi\big)\mathrm{d}\alpha_t,$$

or, in distributional form,

$$\partial_t \alpha_t + \mathrm{div}(\alpha_t v_t) = \frac{1}{2}\sum_{i,j}\partial^2_{i,j}\big((D_t)_{i,j}\alpha_t\big). \tag{16.9}$$

Thus the natural noisy analogue of (16.2) regresses the instantaneous drift,

$$\min_w \mathbb{E}\big[\|w_t(Z_t) - r_t(U, Z_t)\|^2\big],$$

and learns the drift term of a Fokker–Planck equation, not a pure continuity equation unless the diffusion tensor vanishes. When $\alpha_t = \rho_t \mathrm{d}x$ has a smooth positive density, the same marginal curve can be represented, at least formally, by a probability-flow ODE

$$\partial_t \rho_t + \mathrm{div}(\rho_t \bar{v}_t) = 0, \qquad \bar{v}_t = v_t - \frac{1}{2\rho_t}\,\mathrm{div}(\rho_t D_t),$$

where the divergence of the matrix field $\rho_t D_t$ is taken row-wise. In the scalar spatially homogeneous case $D_t = \sigma_t^2 \mathrm{Id}$, this reduces to

$$\bar{v}_t = v_t - \frac{\sigma_t^2}{2}\nabla \log \rho_t,$$

which is the familiar score correction relating diffusion SDEs to probability-flow ODEs.

Connection with diffusion models. In the special case where $I_t(x, y) = (1-t)x + ty$ is a linear interpolation and $\pi = \alpha \otimes \beta$, the curve α_t is a convolution of rescaled versions of α_0 and α_1. The flow-matching problem (16.2) becomes

$$\min_{(v_t)_t} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|v_t((1-t)x + ty) - (y-x)\|^2 \, \mathrm{d}\alpha_0(x) \mathrm{d}\alpha_1(y).$$

When one endpoint is an isotropic Gaussian, this construction is closely related to the probability-flow formulation of diffusion models, up to the usual change of time parametrization [498]. This is why flow matching can be viewed both as a deterministic alternative to diffusion training and as a common language for diffusion paths, OT-inspired paths, and rectified paths [343, 347, 3]. The next two propositions are written in the noising direction, from a data law α to a Gaussian; reversing time gives the corresponding sampling flow. They also give an explicit closed form for v_t and show that it is a gradient field. In this setting, v_t is also the solution of the constrained least-squares problem (14.7). The regression (16.2) is computationally simpler because the continuity equation has already been enforced by the chosen interpolant. To prove this, we rely on Tweedie's formula [202], which expresses the optimal Gaussian denoiser through the score, i.e. the gradient of the log-density.

Proposition 16.6 (Tweedie identity)**.** *Let W be a random vector in $\mathbb{R}^d$ with law $\beta \in \mathcal{P}_1(\mathbb{R}^d)$. For $\sigma > 0$, observe*

$$Z = W + \sigma \varepsilon, \quad \text{where } \varepsilon \sim \mathcal{N}(0, \mathrm{Id}) \text{ is independent of } W.$$

Denote by ρ_σ the smooth positive density of

$$\beta_\sigma = \beta * \mathcal{N}\big(0, \sigma^2 \mathrm{Id}\big),$$

which is the law of Z. Then the conditional mean admits the following everywhere-defined version:

$$\mathbb{E}\big[\, W \mid Z = z \,\big] = z + \sigma^2 \nabla \log \rho_\sigma(z) \qquad \text{for all } z \in \mathbb{R}^d.$$

Proof. Let φ_σ be the $\mathcal{N}(0, \sigma^2 \mathrm{Id})$ density. Bayes' rule gives

$$\mathbb{E}[W \mid Z = z] = \frac{1}{\rho_\sigma(z)} \int_{\mathbb{R}^d} w \, \varphi_\sigma(z - w) \, \mathrm{d}\beta(w).$$

Differentiating the Gaussian convolution under the integral sign and using $\nabla_z \varphi_\sigma(z-w) = -\sigma^{-2}(z-w)\,\varphi_\sigma(z-w)$ yields

$$\nabla \rho_\sigma(z) = \int \nabla_z \varphi_\sigma(z-w) \, \mathrm{d}\beta(w) = -\sigma^{-2}\Big(z - \mathbb{E}[W \mid Z = z]\Big)\, \rho_\sigma(z).$$

Rearranging finishes the proof. □

Proposition 16.7 (Gaussian-endpoint flow-matching field)**.** *Let $X \sim \alpha$ and $Y \sim \mathcal{N}(0, \mathrm{Id})$ be independent. For $t \in (0,1)$ set*

$$Z_t = (1-t)\, X + t\, Y, \qquad \alpha_t = \mathrm{Law}(Z_t) = \rho_t \, \mathrm{d}x.$$

The regression minimizer $v^\star : \mathbb{R}^d \times (0,1) \to \mathbb{R}^d$ of

$$\min_v \int_0^1 \iint_{\mathbb{R}^d \times \mathbb{R}^d} \big| y - x - v\big((1-t)x + ty, t\big)\big|^2 \, \mathrm{d}\alpha(x) \, \mathrm{d}\mathcal{N}(y) \, \mathrm{d}t$$

is

$$v^\star(x,t) = -\frac{1}{1-t}\, x \; - \; \frac{t}{1-t} \nabla \log \rho_t(x) \qquad (x \in \mathbb{R}^d,\ t \in (0,1)).$$

In particular, for each $t \in (0,1)$ this field is a gradient field,

$$v^\star(\cdot, t) = -\nabla \left(\frac{\|\cdot\|^2}{2(1-t)} + \frac{t}{1-t} \log \rho_t \right).$$

Proof. Fix $t \in (0,1)$ and write $W = (1-t)X$, $\sigma = t$, so that $Z_t = W + \sigma Y$ matches the setting of Proposition 16.6. Conditional expectations satisfy $v^\star(z,t) = \mathbb{E}[Y - X \mid Z_t = z] = \frac{1}{t}\,\mathbb{E}[Z_t - W \mid Z_t = z] - \frac{1}{1-t}\,\mathbb{E}[W \mid Z_t = z]$. Applying Proposition 16.6 to $\mathbb{E}[W \mid Z_t = z]$ and noting $\mathbb{E}[Y \mid Z_t = z] = -\,t\,\nabla \log \rho_t(z)$ gives the claimed formula. □

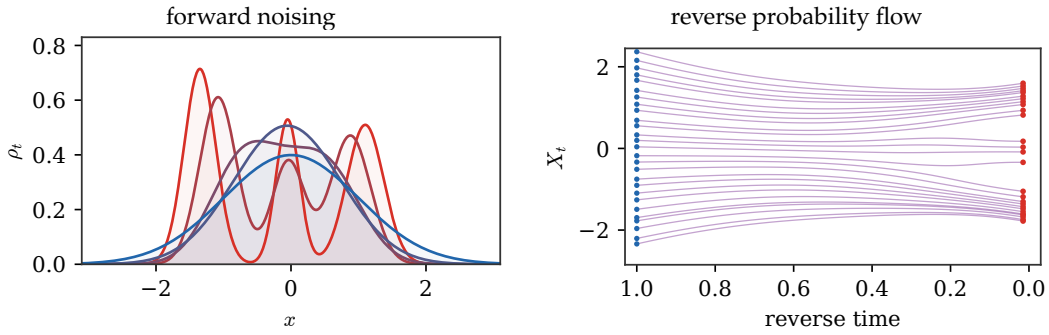


Figure 16.2: One-dimensional diffusion bridge for a Gaussian-mixture data law. The forward path $Z_t = (1-t)X + tY$ smooths the red data density toward a blue Gaussian endpoint. Reversing the probability-flow ODE transports a denser set of blue noise samples back toward the data modes, making the splitting of trajectories across mixture components visible.

Figure 16.2 shows both directions of this construction: the prescribed interpolation noising the data and the same probability-flow ODE integrated backward for sampling.

The same probability-flow intuition is visible in two dimensions. For a discrete data law, or more generally for a Gaussian mixture, the noising density is a Gaussian mixture whose score can be evaluated explicitly. This makes it possible to draw backward trajectories without training a neural network. In Figure 16.3, the Gaussian endpoint has covariance $\sigma^2\mathrm{Id}$ to keep the geometry visible at the scale of the three atoms. For a scalar noising schedule $Z_t = a_t X + b_t Y$, the intermediate law has component centers $a_t c_j$ and covariance $(b_t\sigma)^2\mathrm{Id}$. For the linear bridge, $p_t(z) = \sum_j w_j \mathcal{N}((1-t)c_j, (t\sigma)^2\mathrm{Id})$, with $s_t = \nabla \log p_t$, and the scaled version of Proposition 16.7 gives $v_t(z) = -(z + t\sigma^2 s_t(z))/(1-t)$.

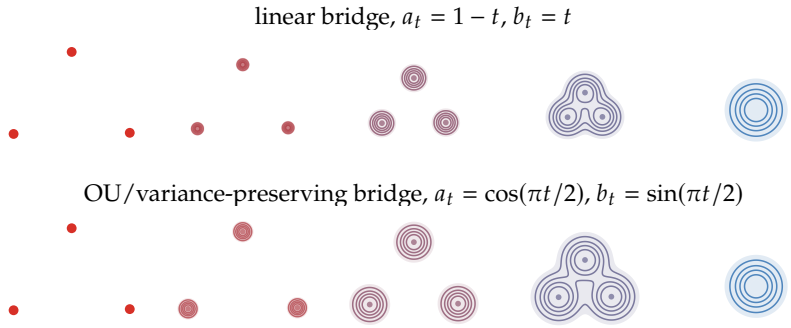


Figure 16.3: Two-dimensional noising paths from three Dirac masses to a single Gaussian. The top row shows the linear interpolation $Z_t = (1-t)X + tY$, whose component centers move linearly toward the origin and whose covariance grows like $(t\sigma)^2\mathrm{Id}$. The bottom row uses the variance-preserving Ornstein–Uhlenbeck coefficients $a_\tau = e^{-\tau}$ and $b_\tau = \sqrt{1-e^{-2\tau}}$, reparametrized by $\tau = -\log\cos(\pi t/2)$ so that $a_t = \cos(\pi t/2)$ and $b_t = \sin(\pi t/2)$. It has the same endpoints but a different speed of contraction and noising.

When is the induced map optimal? Integrating the learned velocity gives a deterministic map from α_0 to α_1, but this map is not automatically the Brenier optimal map. It is optimal only in special cases where the accumulated flow remains the gradient of a convex potential. The Gaussian product-coupling case already shows the precise obstruction: the interpolated covariances are simple, the velocity is affine, but the terminal map can contain a hidden rotational part. This phenomenon, and its extensions to rectified flows and mixtures, is analyzed in depth in [275].

Proposition 16.8 (Gaussian flow matching and optimality)**.** *Let $\Sigma_0, \Sigma_1 \succ 0$ and let $X_0 \sim \mathcal{N}(0,\Sigma_0)$ and $X_1 \sim \mathcal{N}(0,\Sigma_1)$ be independent. Consider the linear flow-matching interpolation*

$$Z_t = (1-t)X_0 + tX_1, \qquad \alpha_t = \mathrm{Law}(Z_t) = \mathcal{N}(0,\Sigma_t),$$

where

$$\Sigma_t = (1-t)^2\Sigma_0 + t^2\Sigma_1. \tag{16.10}$$

Then the exact flow-matching velocity is affine, $v_t(z) = A_t z$, with

$$A_t = \big(t\Sigma_1 - (1-t)\Sigma_0\big)\Sigma_t^{-1}. \tag{16.11}$$

The induced flow map T_t^{FM} from α_0 to α_t is

$$T_t^{\mathrm{FM}} = \Sigma_0^{1/2}\Big((1-t)^2\mathrm{Id} + t^2\Sigma_0^{-1/2}\Sigma_1\Sigma_0^{-1/2}\Big)^{1/2}\Sigma_0^{-1/2}. \tag{16.12}$$

In particular,

$$T_1^{\mathrm{FM}} = \Sigma_0^{1/2}\big(\Sigma_0^{-1/2}\Sigma_1\Sigma_0^{-1/2}\big)^{1/2}\Sigma_0^{-1/2}. \tag{16.13}$$

This terminal map coincides with the quadratic optimal transport map

$$T^{\mathrm{OT}} = \Sigma_0^{-1/2}\big(\Sigma_0^{1/2}\Sigma_1\Sigma_0^{1/2}\big)^{1/2}\Sigma_0^{-1/2} \tag{16.14}$$

if and only if $\Sigma_0\Sigma_1 = \Sigma_1\Sigma_0$.

Proof. The conditional-expectation formula gives $v_t(z) = \mathbb{E}[X_1 - X_0 \mid Z_t = z]$. Since all variables are jointly Gaussian, this conditional expectation is linear and

$$v_t(z) = \mathrm{Cov}(X_1 - X_0, Z_t)\,\mathrm{Cov}(Z_t)^{-1}z = \big(t\Sigma_1 - (1-t)\Sigma_0\big)\Sigma_t^{-1}z,$$

which proves (16.11). To solve the characteristic equation, whiten the source by setting

$$C = \Sigma_0^{-1/2}\Sigma_1\Sigma_0^{-1/2}, \qquad \widetilde{Z}_t = \Sigma_0^{-1/2}Z_t.$$

In these coordinates the source covariance is Id and $\widetilde{\Sigma}_t = (1-t)^2\mathrm{Id} + t^2C$. Because Id and C commute, the affine flow map in whitened coordinates is simply $\widetilde{T}_t = \widetilde{\Sigma}_t^{1/2}$. Indeed,

$$\frac{\mathrm{d}}{\mathrm{d}t}\widetilde{\Sigma}_t^{1/2} = \big(tC - (1-t)\mathrm{Id}\big)\widetilde{\Sigma}_t^{-1/2},$$

which is exactly the equation $\dot{\widetilde{T}}_t = \widetilde{A}_t\widetilde{T}_t$ with $\widetilde{T}_0 = \mathrm{Id}$. Returning to the original coordinates gives (16.12), and $t = 1$ gives (16.13).

Both T_1^{FM} and T^{OT} push $\mathcal{N}(0,\Sigma_0)$ to $\mathcal{N}(0,\Sigma_1)$. The Brenier map between nondegenerate Gaussians is the unique symmetric positive definite linear map with this property. Hence $T_1^{\mathrm{FM}} = T^{\mathrm{OT}}$ if and only if T_1^{FM} is symmetric positive definite. The map T_1^{FM} is similar to $C^{1/2}$, so if it is symmetric then it is automatically positive definite. It remains to characterize symmetry. Since $C^{1/2}$ is symmetric positive definite,

$$(T_1^{\mathrm{FM}})^\top = \Sigma_0^{-1/2}C^{1/2}\Sigma_0^{1/2}.$$

Thus symmetry of T_1^{FM} is equivalent to $\Sigma_0C^{1/2} = C^{1/2}\Sigma_0$, hence to $\Sigma_0C = C\Sigma_0$ by functional calculus. Multiplying this identity on the left and right by $\Sigma_0^{1/2}$ gives $\Sigma_0\Sigma_1 = \Sigma_1\Sigma_0$. Conversely, if Σ_0 and Σ_1 commute, they are orthogonally co-diagonalizable, and both (16.13) and (16.14) reduce in that basis to the diagonal map with entries $\sqrt{\lambda_{1,k}/\lambda_{0,k}}$. This proves the equivalence. □

The proposition gives a compact warning about a common overinterpretation of flow matching.

> **Remark 16.9 (Changing the bridge speed does not restore optimality).** The same terminal map (16.13) is obtained for every continuously differentiable scalar schedule $Z_t = a_tX_0 + b_tX_1$ satisfying $(a_0,b_0) = (1,0)$, $(a_1,b_1) = (0,1)$, and $a_t^2\mathrm{Id} + b_t^2C \succ 0$ for all t. Indeed, after whitening, the covariance path is $a_t^2\mathrm{Id} + b_t^2C$, and the flow map is its positive square root. Thus changing the speed of a nondegenerate scalar Gaussian bridge, for instance by using an OU schedule, cannot repair the non-optimality created by non-commuting covariances.
>
> Commuting covariances reduce the terminal map to independent one-dimensional scalings, whereas non-commuting covariances create a non-symmetric affine map, hence a transport with a rotational or shearing component. More generally, mixture-like paths can create the same obstruction even when every instantaneous velocity looks natural. This distinction is closely related to counterexamples showing that flow maps associated with Fokker–Planck or diffusion-type evolutions do not in general provide optimal transport maps [324]. In particular, starting from an isotropic Gaussian does not by itself guarantee optimality once the target distribution is non-Gaussian; additional structural assumptions on the path or on the coupling are needed.

Variations on the interpolant. The geometry of the generated trajectories depends on the chosen interpolant, not only on the two endpoint laws. There is first a harmless ambiguity: a monotone reparametrization $Z_t = (1-\lambda(t))X + \lambda(t)Y$ of the linear bridge only changes the speed of the flow,

$$v_t(z) = \lambda'(t)\, v^{\mathrm{lin}}_{\lambda(t)}(z), \qquad v^{\mathrm{lin}}_r(z) = \mathbb{E}[Y - X \mid (1-r)X + rY = z].$$

It therefore leaves the spatial integral curves unchanged. Diffusion models use a genuinely different family of noising paths. If

$$Z_t = a_t X + b_t Y, \qquad Y \sim \mathcal{N}(0, \sigma^2 \mathrm{Id}),$$

then both the mixture centers and the component variances are changed. Writing p_t for the density of Z_t and $s_t = \nabla \log p_t$, Tweedie's formula gives, away from times where $a_t = 0$,

$$v_t(z) = a_t' \,\mathbb{E}[X \mid Z_t = z] + b_t' \,\mathbb{E}[Y \mid Z_t = z] = \frac{a_t'}{a_t} z + \left(\frac{a_t' b_t^2}{a_t} - b_t' b_t \right) \sigma^2 s_t(z).$$

For the linear bridge, $a_t = 1 - t$ and $b_t = t$, this recovers the formula above. For the variance-preserving Ornstein–Uhlenbeck noising used in diffusion models,

$$a_\tau = e^{-\tau}, \qquad b_\tau = \sqrt{1 - e^{-2\tau}},$$

one obtains the forward probability-flow velocity $v_\tau(z) = -z - \sigma^2 \nabla \log p_\tau(z)$. Sampling integrates this ODE backward in τ. Equivalently, after introducing reverse time s, the velocity changes sign and becomes $z + \sigma^2 \nabla \log p_\tau(z)$. The resulting trajectories are more curved than for the linear bridge because the centers and variances evolve according to the OU/Fokker–Planck scaling rather than by affine interpolation. Numerically, the integration is stopped at a small positive noising time before the Dirac endpoint, where the score becomes singular.

The finite-time coefficients $a_t = \cos(\pi t/2)$ and $b_t = \sin(\pi t/2)$ are not a new spatial interpolant: they are exactly the OU coefficients after the time change $\tau = -\log \cos(\pi t/2)$. A genuinely different scalar bridge is

$$a_t = (1-t)(1-2t), \qquad b_t = t,$$

whose data coefficient changes sign before vanishing. This overshooting bridge is mainly a diagnostic example: it keeps the same endpoints, but its intermediate mixture reflects through the origin and produces visibly different reverse trajectories.

Figure 16.4 contrasts the reverse OU probability-flow trajectories with straight quadratic-OT displacement rays toward the same target atoms.

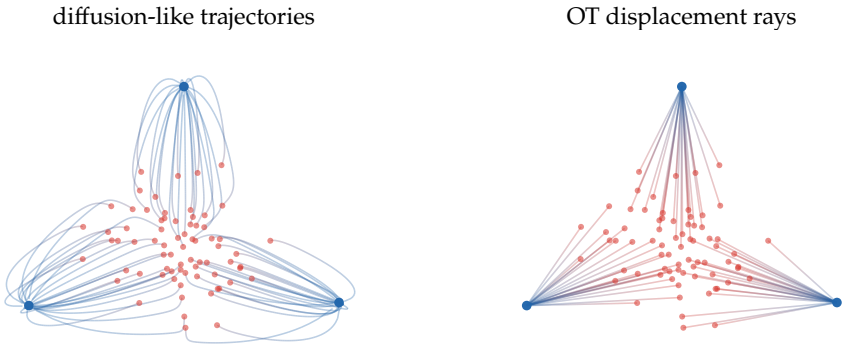


Figure 16.4: Diffusion-style sampling trajectories compared with OT rays in the three-Dirac setting of Figure 16.3. Red particles are sampled from the centered Gaussian endpoint and transported toward the three blue atoms. The left panel integrates the reverse probability-flow ODE for the variance-preserving OU noising $a_\tau = e^{-\tau}$, $b_\tau = \sqrt{1 - e^{-2\tau}}$, using the closed-form Gaussian-mixture score and stopping just before the singular Dirac endpoint. The right panel uses the straight displacement rays selected by a quadratic OT matching to the same atoms.

Figure 16.5 compares the linear, variance-preserving OU and overshooting bridges, isolating how the interpolation schedule bends the reverse trajectories.

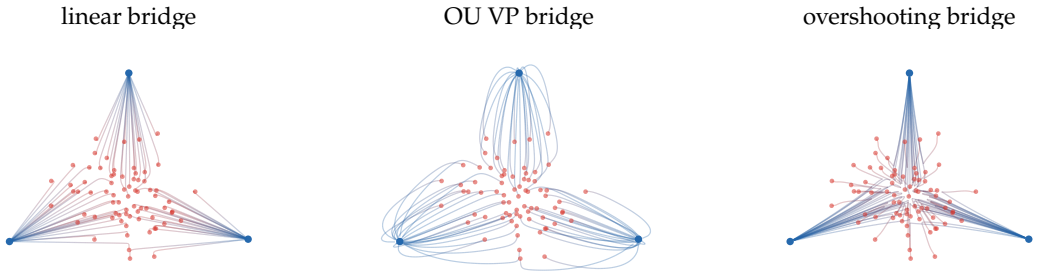


Figure 16.5: Effect of the interpolant on exact reverse-flow trajectories. All panels use the same three-Dirac target and the same Gaussian endpoint. The linear bridge $a_t = 1 - t$, $b_t = t$ produces almost radial curves. The variance-preserving OU bridge $a_\tau = e^{-\tau}$, $b_\tau = \sqrt{1 - e^{-2\tau}}$ changes the relative speed of contraction and noising. The overshooting bridge $a_t = (1 - t)(1 - 2t)$, $b_t = t$ is not a time reparameterization of either one and produces a more pronounced bending of the reverse trajectories.

16.2 One-Step Generative Models

One-step generative models try to keep the geometric training principle of flows while removing the expensive multi-step integration at sampling time. The idea is to evolve the model distribution during training, but to store the final evolution in a single generator evaluation.

One-step models via parameter-domain discrepancy flows. The most direct one-step strategy is to train the generator parameters themselves by descending a discrepancy between generated and data distributions. Let ζ be a simple latent distribution, let $f_\theta : \mathcal{Z} \to \mathcal{X}$ be a neural generator, and let β be the data distribution. The objective is to find θ such that

$$(f_\theta)_\sharp \zeta = \beta,$$

or, in practice, such that the generated law is close to β. Given a discrepancy $\mathcal{D}$ on probability measures, define

$$\mathcal{E}_\beta(\gamma) := \mathcal{D}(\gamma, \beta), \qquad H_\beta(\theta) := \mathcal{E}_\beta\big((f_\theta)_\sharp \zeta\big) = \mathcal{D}\big((f_\theta)_\sharp \zeta, \beta\big). \tag{16.15}$$

This viewpoint includes Wasserstein GANs, MMD-GANs and Sinkhorn generative models [19, 199, 239]; when $\mathcal{D}$ is written through a dual potential or discriminator, it connects directly with the adversarial formulation of Section 7.4. The resulting training dynamics is the ordinary Euclidean gradient flow in parameter space,

$$\dot\theta_t = -\nabla_\theta H_\beta(\theta_t). \tag{16.16}$$

This parameter flow induces a flow of generated measures, but this induced flow is not intrinsic in general. Set

$$\alpha_t := (f_{\theta_t})_\sharp \zeta, \qquad X_t = f_{\theta_t}(Z), \qquad Z \sim \zeta.$$

If $t \mapsto \theta_t$ is smooth and f_θ is differentiable with respect to θ, then

$$\dot X_t = \partial_\theta f_{\theta_t}(Z)\dot\theta_t.$$

For α_t-a.e. x, let $\eta_{t,x}$ be the disintegration of the latent law ζ with respect to the map f_{θ_t}. Thus $\eta_{t,x}$ is supported on the fiber $\{z \,;\, f_{\theta_t}(z) = x\}$, and for every bounded measurable ψ,

$$\int \psi(z)\,\mathrm{d}\zeta(z) = \int_{\mathcal{X}} \left(\int \psi(z)\,\mathrm{d}\eta_{t,x}(z) \right) \mathrm{d}\alpha_t(x).$$

The Eulerian velocity of the generated measure is therefore the conditional average

$$v_t(x) = \int \partial_\theta f_{\theta_t}(z)\dot\theta_t \,\mathrm{d}\eta_{t,x}(z) = -\int \partial_\theta f_{\theta_t}(z)\nabla_\theta H_\beta(\theta_t)\,\mathrm{d}\eta_{t,x}(z). \tag{16.17}$$

Indeed, for every smooth test function φ,

$$\frac{\mathrm{d}}{\mathrm{d}t}\int \varphi(x)\,\mathrm{d}\alpha_t(x) = \int \langle \nabla\varphi(x),\, v_t(x)\rangle\,\mathrm{d}\alpha_t(x),$$

which is the weak form of the continuity equation

$$\partial_t \alpha_t + \operatorname{div}(\alpha_t v_t) = 0. \tag{16.18}$$

The velocity (16.17) depends on the parameter value θ_t and on the latent disintegration, not only on the measure α_t. If the parametrization is non-identifiable, two different parameters can generate the same measure while inducing different velocities. Thus (16.16) is a Euclidean gradient flow on parameter space, and only after push-forward a model-dependent flow on measures. It coincides with an intrinsic Wasserstein gradient flow only in special cases where v_t agrees with the Wasserstein velocity

$$-\nabla \delta_\gamma \mathcal{E}_\beta(\gamma)\big|_{\gamma=\alpha_t}.$$

More generally, a projected Wasserstein flow on the model manifold would require projecting this intrinsic velocity onto the infinitesimal velocities attainable by variations of θ, using the $L^2(\alpha_t)$ metric. The ordinary Euclidean parameter gradient flow (16.16) does not perform this projection in general. In machine learning, however, β is accessed through minibatches. If

$$\hat{\beta}_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{Y_i}, \qquad Y_i \overset{\text{i.i.d.}}{\sim} \beta,$$

then the implemented update often replaces H_β by the random objective $H_{\hat{\beta}_n}$. For nonlinear discrepancies this is not, in general, an unbiased gradient estimator:

$$\mathbb{E}\left[\nabla_\theta H_{\hat{\beta}_n}(\theta)\right] \neq \nabla_\theta H_\beta(\theta).$$

The expectation is over the data minibatch. This bias is a finite-sample effect of inserting the empirical measure inside a nonlinear transport or adversarial optimization before differentiating; it is distinct from the optimization noise of stochastic gradient descent. MMD losses admit unbiased U-statistic variants, but OT and entropic OT objectives generally exhibit the bias–variance phenomena discussed in Section 10.3.

Example 16.10 (Application to perturbation-response prediction). In perturbation biology, one observes an unperturbed control law α and a perturbed law β, but not paired cells. The goal is to learn how a new control cell would respond. Neural OT parameterizes a Monge map, a conditional Monge map or a stochastic semi-coupling and trains it from unpaired samples so that the pushed or transported control population matches β. The distinction between a coupling, a deterministic map and an out-of-sample extrapolator is essential here: the learned object should act on cells not seen during training, not merely explain one empirical coupling [86, 356, 129, 308]. Conditional Wasserstein flows in Section 15.8 give a related geometric language when the conditioning variable is depth, context or treatment.

One-step model using Wasserstein flow of discrepancy. This construction keeps the discrepancy-minimization viewpoint of the previous paragraph but changes the dynamics. Instead of following the Euclidean gradient of the discrepancy in parameter space, one prescribes the Wasserstein gradient flow of the generated law and then learns a generator update realizing this motion. This distribution-space, natural-gradient-type dynamics is less tied to a particular parametrization and can have better convergence behavior, in particular by reducing parameter-space traps when the discrepancy has a favorable geometry.

Let ζ be a simple latent distribution and let $\alpha_\theta = (G_\theta)_\sharp \zeta$ be the model distribution. Assume that the target data distribution is β. The Wasserstein-flow construction chooses a discrepancy

$$\mathcal{E}_\beta(\alpha),$$

for instance a smoothed $\mathrm{KL}(\alpha|\beta)$, an MMD/IPM loss, or the debiased Sinkhorn divergence $\bar{\mathcal{L}}_c^\varepsilon(\alpha,\beta)$ introduced in Section 8.10. The associated formal descent is

$$\partial_t \alpha_t + \operatorname{div}(\alpha_t w_t) = 0, \qquad w_t(x) = -\nabla \delta_\alpha \mathcal{E}_\beta(\alpha_t)(x). \tag{16.19}$$

Instead of integrating (16.19) at inference time, one fits, at each training time t, a parametric residual field U_{η_t} along the current model distribution:

$$\min_\eta \int \|U_\eta(x) - w_t(x)\|^2 \, \mathrm{d}\alpha_t(x). \tag{16.20}$$

In a particle or generator implementation, the learned residual is then used to update the current generator by

$$\alpha_\theta^+ = (\mathrm{Id} + \tau U_{\eta_t})_\sharp \alpha_\theta, \qquad \text{or equivalently} \qquad G_\theta^+(z) = G_\theta(z) + \tau U_{\eta_t}(G_\theta(z)).$$

This is an ideal function-space update. If the residuals were stored as a growing composition, evaluation would still require one layer per training update and the model would not be one-step. A genuine one-step implementation therefore fits the updated outputs back into a fixed generator architecture, or distills the accumulated transport into one network, so that the final sampler remains a single evaluation of G_θ.

After many training updates, the fixed-architecture or distilled generator is evaluated once at test time. This is the organizing principle behind recent one-step methods based on Wasserstein gradient flows: W-Flow uses such a construction with the Sinkhorn divergence as a tractable global discrepancy [267], while drifting methods evolve the generated distribution during training through a fitted vector field and also admit one-step inference [184]. The gradient-flow interpretation of drifting models, and its relation to KL, MMD, sliced-Wasserstein and Sinkhorn-type discrepancies, is analyzed in [254, 272]. These ideas are also connected to the Sinkhorn-type normalization dynamics used to model attention in Sinkformers [468].

Sliced-Wasserstein flow. A particularly transparent instance uses the sliced objective introduced for imaging and barycenters by Rabin, Peyré, Delon and Bernot [444]. Take

$$\mathcal{E}_\beta(\alpha) = \frac{1}{2}\,\mathrm{SW}_2^2(\alpha, \beta),$$

with SW_2 defined in Section 11.2. With the continuity-equation convention $\partial_t \alpha_t + \mathrm{div}(\alpha_t w_t) = 0$, the descent velocity points from the current projected law toward the target projected law. More precisely, assume that α has a density and sufficient moments so that the projected monotone maps are uniquely defined for σ-almost every direction. If T_θ denotes the one-dimensional monotone transport from $(P_\theta)_\sharp \alpha$ to $(P_\theta)_\sharp \beta$, then the formal $\mathcal{W}_2$-gradient-flow velocity is

$$w_{\mathrm{SW}}[\alpha, \beta](x) = \int_{\mathbb{S}^{d-1}} \big(T_\theta(P_\theta x) - P_\theta x\big)\, \theta \, \mathrm{d}\sigma(\theta). \tag{16.21}$$

The sign follows from the one-dimensional identity: the first variation of $\frac{1}{2}\mathcal{W}_2^2(\alpha, \beta)$ has spatial derivative $s - T(s)$, so the descent velocity is $T(s) - s$, and composing with P_θ lifts it in the direction θ. Thus empirical implementations only require one-dimensional sorting along sampled directions, followed by averaging the lifted projected displacements. This is the sliced-Wasserstein flow studied by Cozzi and Santambrogio [161], who prove long-time convergence under their hypotheses when the target is Gaussian and also show that the limiting characteristic map is not, in general, the optimal transport map. When both the evolving law and the target are Gaussian, the averaged velocity is affine and the flow closes on means and covariances; this finite-dimensional closure is revisited in Section 16.5, where the sliced objective appears in the Gaussian closure catalogue of Proposition 16.23.

The difference between the full $\mathcal{W}_2$ descent and its sliced analogue is already visible on a simple shape example. Figure 16.6 compares the exact empirical flow of $\frac{1}{2}\mathcal{W}_2^2(\cdot, \beta)$, where one fixed optimal assignment gives straight relaxation curves, with the $\mathcal{W}_2$-gradient flow of $\frac{1}{2}\mathrm{SW}_2^2(\cdot, \beta)$, where the velocity is recomputed from projected monotone rearrangements.

Stein variational gradient descent. Stein variational gradient descent (SVGD) is another deterministic particle flow that fits naturally in this one-step viewpoint [346]. Its original motivation is Bayesian sampling: given a target probability $\beta = \rho_\beta \, \mathrm{d}x$ known through its score $\nabla \log \rho_\beta = -\nabla V$, but not necessarily through its normalizing constant, drive a particle cloud toward β without estimating the score of the current empirical law. Geometrically, this replaces the Wasserstein gradient flow of $\mathrm{KL}(\alpha|\beta)$, whose tangent norm is $L^2(\alpha)$, by the kernelized Benamou–Brenier geometry of Section 14.3, whose velocities live in a vector-valued RKHS.

For the formal density-level calculation, assume $\alpha = \rho_\alpha \, \mathrm{d}x$ and $\beta = \rho_\beta \, \mathrm{d}x$ have smooth positive densities, and let v be a smooth compactly supported vector field. For the perturbation $\alpha_\varepsilon = (\mathrm{Id} + \varepsilon v)_\sharp \alpha$, integration by parts gives the Stein first variation

$$\frac{\mathrm{d}}{\mathrm{d}\varepsilon}\,\mathrm{KL}(\alpha_\varepsilon|\beta)\Big|_{\varepsilon=0} = -\int \big(\langle \nabla \log \rho_\beta(x),\, v(x)\rangle + \mathrm{div}\, v(x)\big)\mathrm{d}\alpha(x).$$

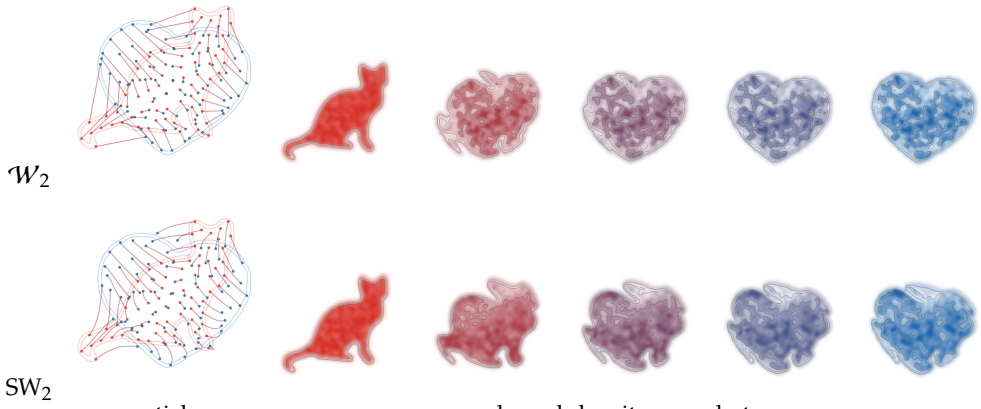


Figure 16.6: Full Wasserstein and sliced-Wasserstein flows from a cat-shaped empirical law to a heart-shaped target. The left panel in each row shows representative particle trajectories colored from red to blue. The five panels on the right render kernel-density estimates of all particles at common normalized times. The $\mathcal{W}_2$ flow follows the fixed optimal assignment and therefore straight relaxation rays, whereas the sliced flow averages one-dimensional sorted rearrangements over projection directions, producing curved trajectories and a different transient density evolution.

The bracket is the Langevin–Stein operator applied to v and averaged under α; because it only evaluates v, $\operatorname{div} v$, and the target score at sample locations, it remains meaningful when α is empirical. Optimizing this linear functional over the unit ball of $\mathcal{H}_k^d$ and using the reproducing property gives the RKHS steepest-descent direction,

$$v_\alpha^{\mathrm{SVGD}}(x) = \int \Big(k(y,x)\nabla \log \rho_\beta(y) + \nabla_y k(y,x)\Big)\mathrm{d}\alpha(y). \tag{16.22}$$

The associated mean-field equation is the transport equation

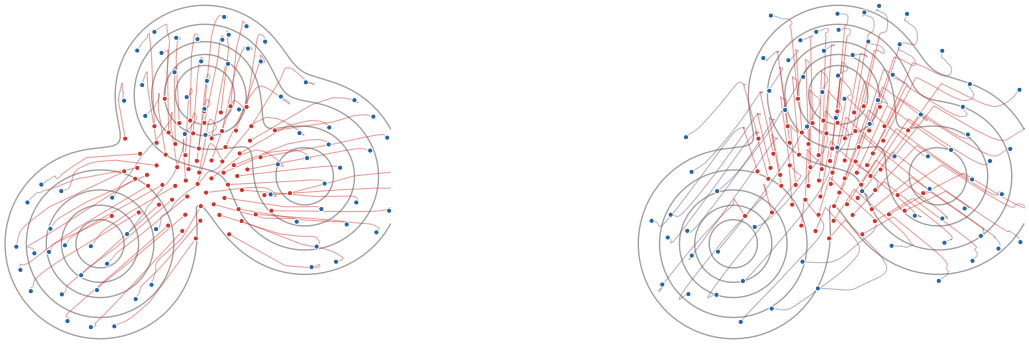
$$\partial_t \alpha_t + \operatorname{div}\big(\alpha_t v_{\alpha_t}^{\mathrm{SVGD}}\big) = 0. \tag{16.23}$$

For particles $\alpha_n^\ell = n^{-1}\sum_i \delta_{X_i^\ell}$, this becomes the explicit update

$$X_i^{\ell+1} = X_i^\ell + \tau\,\frac{1}{n}\sum_{j=1}^{n}\Big(k(X_j^\ell, X_i^\ell)\nabla \log \rho_\beta(X_j^\ell) + \nabla_{X_j} k(X_j^\ell, X_i^\ell)\Big). \tag{16.24}$$

The first term attracts particles toward increasing target log-density; the second term is a kernel repulsion that prevents immediate collapse. Liu's gradient-flow interpretation [345], the geometric analysis of Duncan, Nuesken and Szpruch [194], and the many-particle/long-time analysis of Nuesken and Renger [405] make precise in which sense this is a gradient flow in a Stein or kernelized transport geometry. In generative modeling terms, SVGD transports a simple empirical latent law by repeated smooth particle updates. It is therefore close in spirit to the drifting fields above, but its velocity is not learned by regression: it is the closed-form RKHS steepest-descent direction of the KL functional.

Figure 16.7 contrasts this RKHS flow with a particle closure of the Wasserstein gradient flow of $\mathrm{KL}(\alpha|\beta)$. The latter evaluates the current score $\nabla \log \rho_\alpha$ by a Gaussian KDE, while SVGD avoids this density estimate and uses the target score together with a kernel repulsion.



$\mathcal{W}_2$ KDE closure SVGD

Figure 16.7: Particle trajectories for two deterministic descents of relative entropy. Both descents target the same three-Gaussian density, shown by gray contours. The left panel approximates the $\mathcal{W}_2$ gradient flow of $\mathrm{KL}(\alpha|\beta)$ by the KDE velocity $\nabla \log \rho_\beta - \nabla \log \rho_{\alpha,h}$. The right panel uses the RBF-kernel SVGD velocity (16.22), where target-score attraction is coupled with RKHS repulsion. Both panels use the same initial particles and color trajectory time from red to blue.

Self-corrected drifting fields. Drifting methods need not start from an exact Wasserstein gradient. They often prescribe an attraction-minus-repulsion field and then regress this field in $L^2(\alpha_t)$. A simple continuous version uses a strictly positive kernel $K_\varepsilon(x, y)$ for which the following integrals are finite, and defines, for any probability measure ν,

$$B_\varepsilon[\nu](x) := \frac{\int (y - x) K_\varepsilon(x, y)\, \mathrm{d}\nu(y)}{\int K_\varepsilon(x, y)\, \mathrm{d}\nu(y)}. \tag{16.25}$$

For the Gaussian kernel $K_\varepsilon(x, y) = \exp(-\|x - y\|^2/(2\varepsilon))$, this normalized field is a score of a smoothed density:

$$B_\varepsilon[\nu](x) = \varepsilon \nabla \log \left(\int K_\varepsilon(x, y)\, \mathrm{d}\nu(y) \right). \tag{16.26}$$

The drifting velocity is then

$$u_t(x) = B_\varepsilon[\beta](x) - B_\varepsilon[\alpha_t](x) = \varepsilon \nabla \log \frac{\int K_\varepsilon(x, y)\, \mathrm{d}\beta(y)}{\int K_\varepsilon(x, y)\, \mathrm{d}\alpha_t(y)}. \tag{16.27}$$

The first term pulls samples toward data, while the second term corrects self-attraction and prevents all particles from collapsing onto the same high-density region. For a fixed reference measure, $B_\varepsilon[\nu]$ is precisely the Gaussian mean-shift displacement in (16.36): it moves x toward the local kernel barycenter of ν. Hence self-corrected drifting can be read as the difference between a target mean-shift field and the current model's own mean-shift field. Sinkhorn drifting replaces these one-sided kernel normalizations by two-sided entropic OT couplings, so that the cross and self terms are normalized by Sinkhorn scaling rather than by a single denominator [272].

Figure 16.8 illustrates why the self term matters: it corrects the collapse created by target attraction alone and improves coverage of separated modes.

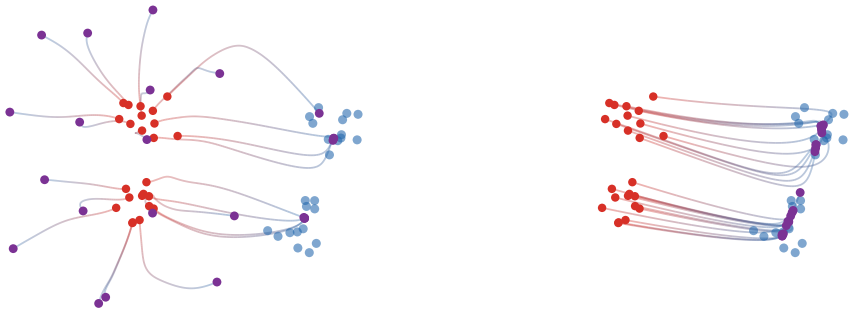
Proposition 16.11 (Instantaneous gradient representation of drifting)**.** *Let $\alpha_t = \rho_t\, \mathrm{d}x$ be a smooth curve of probability measures with positive densities, and let $u_t = \nabla \varphi_t$ be a smooth time-dependent gradient field. Define the semi-relaxed functional*

$$\mathcal{R}_t(\alpha|\alpha_t) := -\int \varphi_t(x)\, \mathrm{d}\alpha(x) + \int \varphi_t(x)\, \mathrm{d}\alpha_t(x). \tag{16.28}$$

Here α_t and φ_t are frozen when taking the first variation with respect to the first argument α. Then the continuity equation

$$\partial_t \alpha_t + \mathrm{div}(\alpha_t u_t) = 0$$

is the formal Wasserstein gradient descent of the frozen time-dependent functional $\alpha \mapsto \mathcal{R}_t(\alpha|\alpha_t)$.



raw kernel drift self-corrected drift

Figure 16.8: Drifting trajectories for a small particle generator. The raw Laplacian-kernel drift has weak long-range attraction and can leave particles away from the data modes. The self-corrected field uses the difference $B_\varepsilon[\beta] - B_\varepsilon[\alpha_t]$, so a longer integration brings particles to the blue modes while repelling them from their own current concentration.

Proof. Since α_t and φ_t are fixed in the variation with respect to α, the first variation is

$$\delta_\alpha \mathcal{R}_t(\alpha|\alpha_t)(x) = -\varphi_t(x).$$

By Proposition 15.7,

$$\mathbb{W}\mathcal{R}_t(\alpha|\alpha_t) = \nabla \delta_\alpha \mathcal{R}_t(\alpha|\alpha_t) = -\nabla\varphi_t = -u_t.$$

The Wasserstein gradient-descent velocity is the negative of this gradient, namely u_t. Substituting this velocity in the continuity equation gives the claimed flow. □

The proposition is an instantaneous representation, not a claim that $(\alpha_t)_t$ is the gradient flow of one fixed energy: both φ_t and the reference argument α_t are frozen before differentiating with respect to α.

Example 16.12 (Kernel drifting as a frozen surrogate). For the Gaussian-kernel drift (16.27), set

$$\varphi_t(x) = \varepsilon \log \frac{\int K_\varepsilon(x,y)\,\mathrm{d}\beta(y)}{\int K_\varepsilon(x,y)\,\mathrm{d}\alpha_t(y)}.$$

Then $u_t = \nabla\varphi_t$, so Proposition 16.11 shows that kernel drifting is the Wasserstein gradient descent of

$$\mathcal{R}_t^{\mathrm{drift}}(\alpha|\alpha_t) = \varepsilon \int \log \frac{\int K_\varepsilon(x,y)\,\mathrm{d}\alpha_t(y)}{\int K_\varepsilon(x,y)\,\mathrm{d}\beta(y)}\,\mathrm{d}\alpha(x) + \text{constant}.$$

The surrogate vanishes at $\alpha = \alpha_t$, but it need not be nonnegative and is therefore not a divergence. It is "semi-relaxed" because the current model α_t is used to build the potential, but it is not varied inside the denominator when computing the first variation in α.

Remark 16.13 (General fields and projection onto gradients). A general regressed field b_t is not necessarily the minimal Wasserstein tangent representative. Such representatives belong to the $L^2(\alpha_t)$ closure of gradient fields, and fields producing the same continuity-equation variation can differ by a weighted divergence-free component. The gradient component is obtained by the weighted projection

$$\nabla\varphi_t = \operatorname*{argmin}_{\nabla\varphi} \int \|\nabla\varphi(x) - b_t(x)\|^2\,\mathrm{d}\alpha_t(x).$$

One may first normalize b_t pointwise, for instance by $b_t/(\|b_t\| + \eta)$, or globally by $\|b_t\|_{L^2(\alpha_t)}$, before this projection. Proposition 16.11 then applies to the projected field. Non-gradient components can still be useful in a parametric model, but they are not descent directions of a scalar functional for the $\mathcal{W}_2$ Riemannian metric.

16.3 Moment Measures

Moment measures give another way to make a whole distribution from one convex potential. Instead of first fixing a simple source law and then learning a transport map, one asks for a convex function whose own log-concave density is pushed forward by its gradient. This couples sampling and mapping in a rigid way: the same potential defines both the source density and the Brenier map. The reward is a hidden convex structure: after a suitable optimal-transport reformulation, a nonlinear equation on convex functions becomes a convex minimization problem for probability densities. This is one of the cleanest places where optimal transport, Prékopa-type inequalities and convex geometry meet.

Definition 16.14 (Moment measure of a convex function)**.** Let $u : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ be a proper lower-semicontinuous convex function such that

$$Z_u := \int_{\mathbb{R}^d} e^{-u(x)}\,\mathrm{d}x < +\infty.$$

The normalized log-concave measure associated with u is

$$\eta_u := Z_u^{-1} e^{-u(x)}\,\mathrm{d}x.$$

Whenever ∇u is defined η_u-almost everywhere, the moment measure of u is

$$\mathfrak{M}(u) := (\nabla u)_\sharp \eta_u.$$

The normalization removes additive constants in u. Translations of the argument are another invariance: if $u_a(x) = u(x - a)$, then η_{u_a} is the translate of η_u, while $\nabla u_a(x) = \nabla u(x - a)$, hence $\mathfrak{M}(u_a) = \mathfrak{M}(u)$. A first obstruction is immediate. Formally, if u is smooth and e^{-u} decays fast enough for the boundary term to vanish, then

$$\int y\,\mathrm{d}\mathfrak{M}(u)(y) = Z_u^{-1} \int \nabla u(x) e^{-u(x)}\,\mathrm{d}x = -Z_u^{-1} \int \nabla(e^{-u(x)})\,\mathrm{d}x = 0.$$

Thus moment measures are necessarily centered. The nonsmooth theory uses essentially continuous convex functions: these are lower-semicontinuous convex functions whose set of discontinuity points has zero $\mathcal{H}^{d-1}$ measure. Since a convex function is continuous in the interior of its effective domain, this condition controls only its boundary behavior.

Figure 16.9 shows the forward construction in one dimension. The map u' is implicit in the push-forward, but the display focuses on the two visible measures: the log-concave source $\eta_u = Z_u^{-1} e^{-u}\mathrm{d}x$ and the resulting moment measure $\mathfrak{M}(u)$.

Theorem 16.15 (Cordero–Erausquin–Klartag)**.** *Let $\alpha \in \mathcal{P}_1(\mathbb{R}^d)$ be a probability measure. There exists an essentially continuous convex function u such that $\alpha = \mathfrak{M}(u)$ if and only if*

$$\int y\,\mathrm{d}\alpha(y) = 0 \qquad \textit{and} \qquad \operatorname{supp}(\alpha) \textit{ is not contained in a hyperplane.}$$

Under these conditions, u is unique up to translations of the argument and additive constants.

This theorem is due to Cordero–Erausquin and Klartag [155]. It is a functional analogue of a Minkowski-type problem: the target measure prescribes how the gradient image of a log-concave density should be distributed. The hyperplane condition is the natural non-degeneracy assumption; otherwise the prescribed gradient image lives in a lower-dimensional affine direction and no coercive full-dimensional convex potential can be recovered.

Example 16.16 (Quadratic potentials). Let $A \in \mathbb{R}^{d\times d}$ be symmetric positive definite and

$$u(x) = \frac{1}{2} x^\top A x + c.$$

Then $\eta_u = \mathcal{N}(0, A^{-1})$ and $\nabla u(x) = Ax$, so

$$\mathfrak{M}(u) = \mathcal{N}(0, A).$$

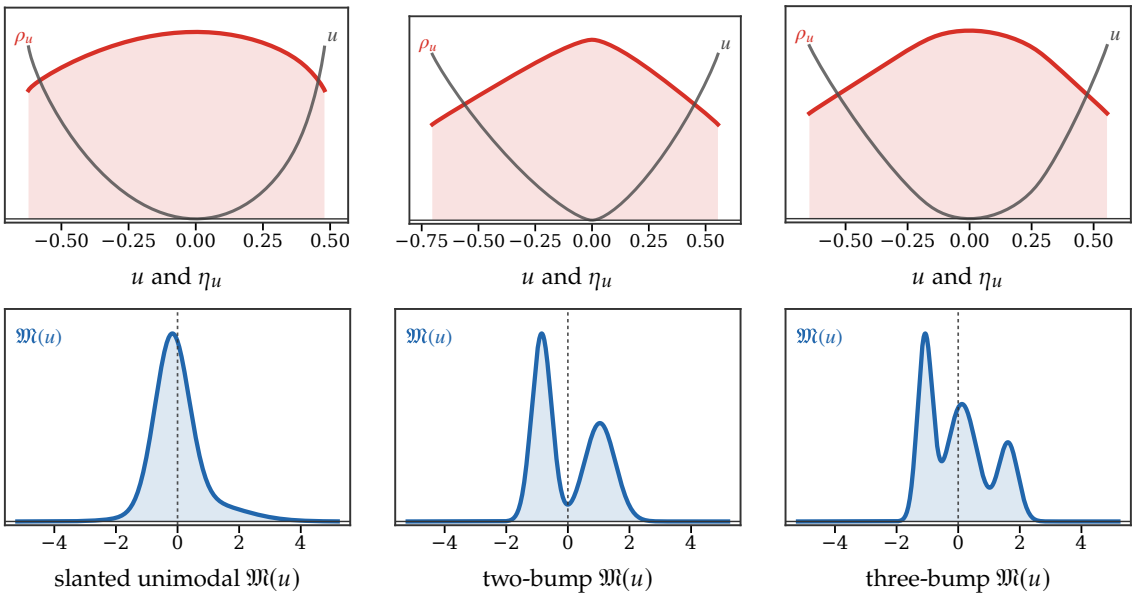


Figure 16.9: Forward moment-measure construction in one dimension. Each column shows a convex potential u chosen so that the moment measure has a prescribed shape: a skewed unimodal density, two bumps, and three bumps with different widths and heights. The top row overlays u (gray, vertically rescaled) with the density of the log-concave source $\eta_u = Z_u^{-1} e^{-u}\,\mathrm{d}x$ (red), while the bottom row shows $\mathfrak{M}(u) = (u')_\sharp \eta_u$ (blue); the dashed vertical line marks the zero barycenter.

Thus centered non-degenerate Gaussians are moment measures. This example also shows the self-dual flavor of the construction: the covariance of the log-concave source is inverted by the linear gradient map.

Optimal-transport variational formulation. Santambrogio [470] reformulates the moment-measure problem as a minimization over absolutely continuous probability measures. For a centered $\alpha \in \mathcal{P}_1(\mathbb{R}^d)$, define the maximal-correlation transport functional, with values in $\mathbb{R} \cup \{+\infty\}$,

$$C_\alpha(\eta) := \sup_{\pi \in \mathcal{U}(\eta,\alpha)} \int_{\mathbb{R}^d \times \mathbb{R}^d} x \cdot y \,\mathrm{d}\pi(x,y). \tag{16.29}$$

By Kantorovich duality for the scalar-product cost,

$$C_\alpha(\eta) = \inf_{v \text{ convex}} \left\{ \int v(x)\,\mathrm{d}\eta(x) + \int v^*(y)\,\mathrm{d}\alpha(y) \right\}, \tag{16.30}$$

where the infimum is over convex functions for which both integrals are well defined and v^* is the Legendre transform. If $\eta, \alpha \in \mathcal{P}_2(\mathbb{R}^d)$, then

$$C_\alpha(\eta) = \frac{1}{2}\int \|x\|^2\,\mathrm{d}\eta(x) + \frac{1}{2}\int \|y\|^2\,\mathrm{d}\alpha(y) - \frac{1}{2}\mathcal{W}_2^2(\eta,\alpha). \tag{16.31}$$

The variational problem attached to a centered target α is

$$\min_{\eta \in \mathcal{P}_1(\mathbb{R}^d)} \{\mathcal{H}(\eta) + C_\alpha(\eta)\}, \qquad \mathcal{H}(r\,\mathrm{d}x) := \int r(x)\log r(x)\,\mathrm{d}x, \tag{16.32}$$

with $\mathcal{H}(\eta) = +\infty$ when η is not absolutely continuous. The centering of α makes this functional invariant under translations of η, since translating η by a vector a changes $\int x \cdot y\,\mathrm{d}\pi$ by $a \cdot \int y\,\mathrm{d}\alpha(y) = 0$.

Proposition 16.17 (Variational characterization of moment measures)**.** *Let $\alpha \in \mathcal{P}_1(\mathbb{R}^d)$ be centered and not supported on a hyperplane. Then the minimization problem (16.32) admits a solution, unique up to translations. Every minimizer is a probability measure with log-concave density of the form*

$$\eta = Z_u^{-1} e^{-u}\,\mathrm{d}x,$$

where u is convex, essentially continuous, and satisfies the moment-measure equation

$$\alpha = (\nabla u)_\sharp \eta.$$

Conversely, any essentially continuous convex u satisfying this equation yields a global minimizer. If $\alpha \in \mathcal{P}_2(\mathbb{R}^d)$, the objective $\eta \mapsto \mathcal{H}(\eta) + C_\alpha(\eta)$ is displacement convex along $\mathcal{W}_2$ geodesics of absolutely continuous measures in $\mathcal{P}_2(\mathbb{R}^d)$ whenever the terms are finite. When merely $\alpha \in \mathcal{P}_1(\mathbb{R}^d)$, the maximal-correlation term remains convex along such finite-second-moment geodesics by approximation; existence and uniqueness on the full $\mathcal{P}_1$ domain follow from the lower-semicontinuity argument of Santambrogio.

Proof. The existence proof has two ingredients. First, since α is centered, translating η does not change either $\mathcal{H}(\eta)$ or $C_\alpha(\eta)$, so one can center a minimizing sequence. Second, the assumption that α is not supported on a hyperplane gives a coercive lower bound of the form $C_\alpha(\eta) \geqslant c_\alpha \int \|x\| \, \mathrm{d}\eta(x)$ for centered absolutely continuous η, with $c_\alpha > 0$. Together with the lower-semicontinuity estimates for entropy and maximal correlation, this yields a minimizer [470].

Let $\eta = r \, \mathrm{d}x$ be a minimizer and let u be a convex optimizer in the dual formula (16.30). Keeping u fixed and varying η in (16.32) gives the Euler equation

$$\log r(x) + 1 + u(x) = \text{constant} \qquad \text{on } \{r > 0\},$$

so $\eta = Z_u^{-1} e^{-u} \mathrm{d}x$. The optimality condition for the scalar-product transport problem says that an optimal coupling is supported on $\{(x, y) : y \in \partial u(x)\}$. Since η is absolutely continuous and u is convex, $\partial u(x) = \{\nabla u(x)\}$ for η-almost every x, hence $\alpha = (\nabla u)_\sharp \eta$.

Conversely, assume $\eta = Z_u^{-1} e^{-u} \mathrm{d}x$ and $\alpha = (\nabla u)_\sharp \eta$. Let ν be a smooth compactly supported competitor, and let T be the Brenier map from η to ν. Along the geodesic $\eta_t = ((1-t)\mathrm{Id} + tT)_\sharp \eta$, the right derivative of the entropy at $t = 0$ is

$$\frac{\mathrm{d}}{\mathrm{d}t} \mathcal{H}(\eta_t)\Big|_{t=0^+} = -\int \langle T(x) - x, \, \nabla u(x) \rangle \, \mathrm{d}\eta(x),$$

where the identity follows by differentiating the Jacobian formula and integrating by parts. The dual optimizer u gives the upper directional bound

$$\frac{\mathrm{d}}{\mathrm{d}t} C_\alpha(\eta_t)\Big|_{t=0^+} \leqslant \int \langle T(x) - x, \, \nabla u(x) \rangle \, \mathrm{d}\eta(x).$$

Conversely, Santambrogio's derivative estimate gives

$$\frac{\mathrm{d}}{\mathrm{d}t} C_\alpha(\eta_t)\Big|_{t=0^+} \geqslant \int \langle T(x) - x, \, \nabla u(x) \rangle \, \mathrm{d}\eta(x),$$

because $(\mathrm{Id}, \nabla u)_\sharp \eta$ is optimal for the scalar-product problem. The two bounds coincide, so the first-order terms cancel. Hence the one-sided derivative of $\mathcal{H} + C_\alpha$ at η in every such direction is zero. Displacement convexity implies global minimality, and approximation removes the smooth compact-support restriction. Strict displacement convexity of entropy gives uniqueness, except for translations; translations do not change C_α because α is centered.

It remains to justify the convexity assertion. The entropy term is displacement convex by McCann's theorem, recalled in Theorem 15.19. If α has finite second moment, identity (16.31) writes C_α as the sum of the 1-convex moment term $\eta \mapsto \frac{1}{2} \int \|x\|^2 \, \mathrm{d}\eta$ and the (-1)-convex term $\eta \mapsto -\frac{1}{2} \mathcal{W}_2^2(\eta, \alpha)$, hence C_α is displacement convex. For a target with only a finite first moment, Santambrogio obtains the same convexity along $\mathcal{P}_2$ geodesics by approximation and proves the full variational characterization by lower semicontinuity. □

Remark 16.18 (Where the convexity is hidden). If one eliminates η first, the problem becomes the convex-potential functional

$$u \mapsto \int u^*(y) \, \mathrm{d}\alpha(y) - \log\left(\int e^{-u(x)} \, \mathrm{d}x \right).$$

This is a Toland-type duality: the functional is not visibly convex as a function of u, because the first term is convex while the logarithmic partition term is concave in this parametrization. Cordero–

Erausquin and Klartag make the hidden convexity visible by changing variables to the dual potential $\varphi = u^*$. Since $u = \varphi^*$ for closed convex potentials, the same functional becomes

$$\varphi \mapsto \int \varphi(y)\,\mathrm{d}\alpha(y) - \log\left(\int e^{-\varphi^*(x)}\,\mathrm{d}x\right).$$

The first term is now affine in φ. The core Prékopa–Leindler input is that $\varphi \mapsto \log \int e^{-\varphi^*}$ is concave along convex combinations of convex functions; equivalently, the negative log-partition in the display is convex. Santambrogio's formulation reveals the same mechanism in transport language: the difficult convexity becomes the displacement convexity of an entropy-plus-maximal-correlation functional in the measure variable η.

Conjugate moment measures for generation. The moment-measure factorization suggests a generative recipe: sample X from the log-concave law $Z_u^{-1}e^{-u}$ and output $\nabla u(X)$. This ties sampling and mapping through the same convex potential. Vesseron, Béthune and Cuturi [536] argue that this direct factorization can be poorly adapted to practical generative modeling, and propose instead the conjugate factorization

$$\beta = (\nabla w^*)_\sharp \left(Z_w^{-1} e^{-w(z)}\,\mathrm{d}z\right). \tag{16.33}$$

Here ∇w^* is the Brenier map from the learned log-concave source to the target distribution β. This keeps the single-convex-potential philosophy, but places the transport map on the conjugate side; it can be parameterized by input-convex neural networks and trained using OT solvers. From the viewpoint of this chapter, moment measures are therefore a rigorous convex-analytic prototype for one-step generators based on gradients of convex potentials.

16.4 Evolution in Depth of Transformers

Deep residual architectures can be read as time discretizations of ODEs or PDEs. For transformers, the transported objects are token measures and the velocity is induced by attention.

Discrete self-attention, its permutation equivariance and its mean-field extension were introduced in the paragraph "Mean-field attention" of Section 4.5; see Definition 4.36. Here the number of tokens is fixed in the discrete model and the limit concerns depth: the number T of residual attention layers grows while each layer makes an update of size $1/T$, as in continuous-depth neural networks [127]. This is distinct from the many-token limit discussed around Definition 4.36. Since the attention weights are normalized using the whole current token law, the resulting velocity is nonlinear and context dependent. Related continuous-depth and Sinkhorn-normalized dynamics are studied in [468, 117].

We focus on the non-causal single-head mechanism of (4.14), ignoring MLP layers, layer normalization and masking. This stripped-down framework isolates the transport structure of a residual attention block.

Attention as a context-dependent velocity. We retain the notation $\theta = (Q, K, V)$ and Γ_θ of Definition 4.36. Whereas (4.14) replaces a token by its attention output, a residual block adds this output through a skip connection. With residual scale $1/T$, it acts as

$$x_i \longmapsto x_i + \frac{1}{T}\,\mathrm{Att}_\theta^{(n)}(x_1, \dots, x_n)_i = x_i + \frac{1}{T}\sum_{j=1}^{n} \frac{e^{\langle Qx_i, Kx_j\rangle} V x_j}{\sum_{\ell=1}^{n} e^{\langle Qx_i, Kx_\ell\rangle}}. \tag{16.34}$$

The attention output is therefore interpreted as a context-dependent velocity, and the skip connection performs one explicit Euler step along this field.

Token measure evolution. Let $x_i^{(\ell)}$ be token i after layer ℓ, and let $\alpha^{(\ell)} = n^{-1}\sum_i \delta_{x_i^{(\ell)}}$ be the corresponding empirical law. With $\tau = 1/T$, equation (16.34) is exactly

$$\alpha^{(\ell+1)} = \left(\mathrm{Id} + \tau\Gamma_{\theta^{(\ell)}}[\alpha^{(\ell)}]\right)_\sharp \alpha^{(\ell)}, \qquad \ell = 0, \dots, T-1. \tag{16.35}$$

Unlike the non-residual map Att_θ of Definition 4.36, this update keeps each current token and adds a small attention displacement. Set $t_\ell = \ell/T$ and suppose that $\theta^{(\ell)}$ samples a sufficiently regular depth-dependent path $t \mapsto \theta_t$. As $T \to +\infty$, a time interpolation of (16.35) converges formally to the conservation equation

$$\partial_t \alpha_t + \mathrm{div}(\alpha_t \Gamma_{\theta_t}[\alpha_t]) = 0.$$

L^2 attention and mean shift. A particularly geometric variant replaces the dot-product score $\langle Qx, Ky\rangle$ by a negative squared Euclidean score $s_\varepsilon(x,y) = -\|x-y\|^2/(2\varepsilon)$. Take, for simplicity, the same token space for queries, keys and values, and set

$$K_\varepsilon(x,y) = \exp(-\|x-y\|^2/(2\varepsilon)), \qquad \rho_\varepsilon[\alpha](x) = \int K_\varepsilon(x,y)\mathrm{d}\alpha(y), \qquad m_\varepsilon[\alpha](x) = \frac{\int yK_\varepsilon(x,y)\mathrm{d}\alpha(y)}{\rho_\varepsilon[\alpha](x)}.$$

The map $x \mapsto m_\varepsilon[\alpha](x)$ is exactly Gaussian-kernel attention, i.e. normalized kernel regression over tokens; such L^2 or Gaussian-kernel scores are used explicitly in transformer variants motivated by Lipschitz control and projection-free attention [303, 320]. Classical mean shift, however, uses the displacement from the current point to this local barycenter. This gives

$$M_\varepsilon[\alpha](x) \coloneqq m_\varepsilon[\alpha](x) - x = \frac{\int (y-x)K_\varepsilon(x,y)\mathrm{d}\alpha(y)}{\rho_\varepsilon[\alpha](x)} = \varepsilon\nabla \log \rho_\varepsilon[\alpha](x) \tag{16.36}$$

and, when α is empirical, $\rho_\varepsilon[\alpha]$ is a Gaussian kernel density estimate up to normalization. Thus $M_\varepsilon[\alpha]$ is the classical Gaussian mean-shift vector [224, 133, 152]. If the data measure α is frozen, the update $x \leftarrow m_\varepsilon[\alpha](x)$ is the usual mode-seeking mean-shift iteration. If instead all support points move and the empirical measure is recomputed after every step, one obtains the self-consistent, or *blurring*, mean-shift process [128]. Its damped update is

$$x_i^{(\ell+1)} = (1-\tau)x_i^{(\ell)} + \tau m_\varepsilon[\alpha^{(\ell)}](x_i^{(\ell)}) = x_i^{(\ell)} + \tau M_\varepsilon[\alpha^{(\ell)}](x_i^{(\ell)})$$

which is an explicit Euler step of the continuous-time mean-shift equation

$$\partial_t \alpha_t + \mathrm{div}\big(\alpha_t M_\varepsilon[\alpha_t]\big) = 0.$$

For $\tau = 1$ this is discrete blurring mean shift; for small τ it approximates a transport PDE that moves each token uphill along the log of the smoothed token density. This distinction between the raw barycentric attention output m_ε and the velocity $M_\varepsilon = m_\varepsilon - \mathrm{Id}$ is important: adding m_ε directly as a residual would produce a different drift. The mean-shift form isolates a purely metric attention mechanism from the learned bilinear geometry of $\langle Qx, Ky\rangle$.

Consensus and Markov averaging. When the averaging measure evolves together with the particles, mean shift becomes a consensus model. The Hegselmann–Krause model updates each opinion by averaging the opinions inside a confidence neighborhood [273]; its finite-agent convergence theory was developed further in [66], and an Eulerian formulation for probability measures appears in [101]. The row-normalized environmental average below is also characteristic of Motsch–Tadmor dynamics [391].

To isolate this mechanism, replace K_ε by a positive kernel K and write

$$M_K[\alpha](x) \coloneqq \frac{\int (y-x)K(x,y)\mathrm{d}\alpha(y)}{\int K(x,y)\mathrm{d}\alpha(y)}. \tag{16.37}$$

For an empirical law $\alpha = \sum_{i=1}^n a_i\delta_{x_i}$, where $a_i > 0$ and $\sum_i a_i = 1$, let $X \in \mathbb{R}^{n\times d}$ have rows $x_i^\top$, set $\mathrm{a} = (a_i)_i$, and define

$$(\mathrm{K}_X)_{i,j} = K(x_i,x_j), \qquad \mathsf{P}_X \coloneqq \mathrm{diag}(\mathrm{K}_X\mathrm{a})^{-1}\mathrm{K}_X\,\mathrm{diag}(\mathrm{a}). \tag{16.38}$$

The matrix P_X is positive and row-stochastic, $\mathsf{P}_X\mathbb{1}_n = \mathbb{1}_n$. Its ith row is precisely the probability law used to average the cloud from the viewpoint of x_i. Figure 9.1 visualizes the corresponding loss of memory for a fixed Markov matrix; mean shift applies the same averaging mechanism to every spatial coordinate, but with a matrix that changes with the cloud. The particle system (16.37) is therefore

$$\dot{x}_i = \frac{\sum_j a_jK(x_i,x_j)(x_j-x_i)}{\sum_j a_jK(x_i,x_j)}, \qquad \text{or equivalently} \qquad \dot{X} = (\mathsf{P}_X - I_n)X. \tag{16.39}$$

Thus blurring mean shift is a state-dependent Markov averaging process.

Dobrushin contraction from Hilbert geometry. The Sinkhorn analysis already contains the two geometries needed here. Section 9.2 introduced the variation seminorm of Definition 9.6 on functions modulo additive constants, while Section 9.4 introduced Hilbert's metric (9.5) on positive vectors modulo multiplication. Logarithmic coordinates identify these quotients exactly:

$$d_{\mathcal{H}}(e^z, e^{z'}) = \|z - z'\|_V.$$

A row-stochastic matrix P is order preserving and satisfies $\mathsf{P}(z + s\mathbb{1}_n) = \mathsf{P}z + s\mathbb{1}_n$. It is therefore a linear topical map, so Proposition 9.8 gives nonexpansiveness on the additive quotient. Strict positivity yields more: Birkhoff contraction can be linearized at the constant ray to obtain a strict variation contraction.

Proposition 16.19 (Dobrushin contraction is the tangent form of Birkhoff contraction)**.** *Let $\mathsf{P} \in \mathbb{R}_+^{n\times n}$ be row-stochastic. Its Dobrushin coefficient is*

$$\delta(\mathsf{P}) := \frac{1}{2}\max_{i,\ell}\sum_j \left|\mathsf{P}_{i,j} - \mathsf{P}_{\ell,j}\right| = 1 - \min_{i,\ell}\sum_j \min\{\mathsf{P}_{i,j}, \mathsf{P}_{\ell,j}\}. \tag{16.40}$$

It is the exact operator norm induced by $\|\cdot\|_V$ on $\mathbb{R}^n/\mathrm{Span}(\mathbb{1}_n)$:

$$\delta(\mathsf{P}) = \sup_{z\notin \mathrm{Span}(\mathbb{1}_n)} \frac{\|\mathsf{P}z\|_V}{\|z\|_V}. \tag{16.41}$$

Consequently,

$$\|\mathsf{P}z\|_V \leqslant \delta(\mathsf{P})\|z\|_V \qquad (z \in \mathbb{R}^n). \tag{16.42}$$

If $\mathsf{P} > 0$, then this exact tangent factor is bounded by the Birkhoff coefficient of Theorem 9.16:

$$\delta(\mathsf{P}) \leqslant \lambda(\mathsf{P}) < 1. \tag{16.43}$$

Proof. Fix two rows $p = (\mathsf{P}_{i,j})_j$ and $p' = (\mathsf{P}_{\ell,j})_j$, and remove their common mass $r_j = \min\{p_j, p'_j\}$. The two residual nonnegative measures have the same mass $1 - \sum_j r_j = \frac{1}{2}\sum_j \left|p_j - p'_j\right|$. Their averages of z can differ by at most this mass times $\|z\|_V$. Maximizing over the two output rows proves the upper bound in (16.41). Conversely, for a pair of rows attaining the maximum, choose $z_j = 1$ when $p_j \geqslant p'_j$ and $z_j = 0$ otherwise. Then $\|z\|_V = 1$ and $(\mathsf{P}z)_i - (\mathsf{P}z)_\ell = \frac{1}{2}\sum_j \left|p_j - p'_j\right|$, proving equality and hence (16.42).

Assume now that $\mathsf{P} > 0$. Apply Theorem 9.16 to $u_s = e^{sz}$ and $\mathbb{1}_n$. Since $\mathsf{P}\mathbb{1}_n = \mathbb{1}_n$,

$$d_{\mathcal{H}}(\mathsf{P}e^{sz}, \mathbb{1}_n) \leqslant \lambda(\mathsf{P}) d_{\mathcal{H}}(e^{sz}, \mathbb{1}_n) = \lambda(\mathsf{P})|s|\|z\|_V.$$

Moreover, $\mathsf{P}e^{sz} = \mathbb{1}_n + s\mathsf{P}z + O(s^2)$, hence $\log(\mathsf{P}e^{sz}) = s\mathsf{P}z + O(s^2)$. Divide by $|s|$ and let $s \to 0$ to obtain $\|\mathsf{P}z\|_V \leqslant \lambda(\mathsf{P})\|z\|_V$. Taking the exact quotient norm (16.41) yields (16.43). This Hilbert–Hopf–Dobrushin relation extends to Markov operators on general cones [235]. □

By duality, the same coefficient is the exact contraction factor of the adjoint Markov evolution on probability vectors: for $\mathrm{p}, \mathrm{p}' \in \Sigma_n$,

$$\|\mathsf{P}^\top \mathrm{p} - \mathsf{P}^\top \mathrm{p}'\|_{\ell^1} \leqslant \delta(\mathsf{P})\|\mathrm{p} - \mathrm{p}'\|_{\ell^1}.$$

Thus (16.42) contracts observables modulo constants, while the adjoint inequality contracts probability laws in total variation. The proposition also separates two useful constants. The Dobrushin coefficient is sharp but depends on the normalized rows. The Birkhoff coefficient can be looser, but its cross-ratio formula is explicit and invariant under positive diagonal scalings, precisely the invariance exploited in the Sinkhorn proof of Theorem 9.17.

The same coefficient applies to the measure-valued dynamics. For $\alpha \in \mathcal{P}(C)$, where $C \subset \mathbb{R}^d$ is compact, define the Markov operator

$$(\mathsf{P}_\alpha h)(x) := \int_C h(y) p_{\alpha,x}(y)\mathrm{d}\alpha(y), \qquad p_{\alpha,x}(y) := \frac{K(x,y)}{Z_\alpha(x)}, \qquad Z_\alpha(x) := \int_C K(x,z)\mathrm{d}\alpha(z). \tag{16.44}$$

Its Dobrushin coefficient is

$$\delta(\mathsf{P}_\alpha) := \frac{1}{2}\sup_{x,x'\in \mathrm{supp}\,\alpha}\int_C |p_{\alpha,x}(y) - p_{\alpha,x'}(y)|\,\mathrm{d}\alpha(y). \tag{16.45}$$

Writing $\|h\|_{V,S} = \sup_S h - \inf_S h$ for a compact set S, the common-mass proof of Proposition 16.19 gives

$$\|\mathsf{P}_\alpha h\|_{V,\operatorname{supp}\alpha} \leqslant \delta(\mathsf{P}_\alpha)\|h\|_{V,\operatorname{supp}\alpha}.$$

To obtain a uniform computable bound, set

$$\eta_K(C) \coloneqq \sup_{x,x',y,y'\in C} \frac{K(x,y)K(x',y')}{K(x',y)K(x,y')}, \qquad \lambda_K(C) \coloneqq \frac{\sqrt{\eta_K(C)}-1}{\sqrt{\eta_K(C)}+1}. \tag{16.46}$$

Row normalization does not change projective cross-ratios. In the discrete case, for instance,

$$\frac{(\mathsf{P}_X)_{i,j}(\mathsf{P}_X)_{\ell,r}}{(\mathsf{P}_X)_{i,r}(\mathsf{P}_X)_{\ell,j}} = \frac{K(x_i,x_j)K(x_\ell,x_r)}{K(x_i,x_r)K(x_\ell,x_j)}. \tag{16.47}$$

Consequently, for every weighted configuration supported in C,

$$\delta(\mathsf{P}_X) \leqslant \lambda(\mathsf{P}_X) = \lambda(\mathrm{K}_X) \leqslant \lambda_K(C). \tag{16.48}$$

The equality in the middle is the same diagonal-scaling invariance used by Sinkhorn: the normalization $\operatorname{diag}(\mathrm{K}_X\mathrm{a})^{-1}$ and the weight matrix $\operatorname{diag}(\mathrm{a})$ disappear from every projective cross-ratio. The integral-operator form of the Birkhoff–Hopf theorem and the same linearization as above therefore give

$$\bar\delta_K(C) \coloneqq \sup_{\alpha\in\mathcal{P}(C)} \delta(\mathsf{P}_\alpha) \leqslant \lambda_K(C) < 1. \tag{16.49}$$

In particular, if $0 < k_- \leqslant K \leqslant k_+$ on $C\times C$, then

$$\lambda_K(C) \leqslant \frac{k_+ - k_-}{k_+ + k_-}. \tag{16.50}$$

Theorem 16.20 (Dobrushin consensus for positive mean shift). *Let $\alpha_0 \in \mathcal{P}(\mathbb{R}^d)$ have compact support, and set*

$$C_0 = \operatorname{conv}(\operatorname{supp}\alpha_0), \qquad D_0 = \operatorname{diam}(\operatorname{supp}\alpha_0), \qquad \bar\delta = \bar\delta_K(C_0) < 1.$$

Assume that $K : C_0\times C_0 \to (0,+\infty)$ is Lipschitz.

For an empirical initial measure $\alpha_0 = \sum_i a_i\delta_{x_i^{(0)}}$, let $X^{(\ell)}$ follow the damped blurring iteration

$$X^{(\ell+1)} = \big((1-\tau)I_n + \tau\mathsf{P}_{X^{(\ell)}}\big)X^{(\ell)}, \qquad 0<\tau\leqslant 1,$$

define

$$\delta^{(\ell)} = \delta(\mathsf{P}_{X^{(\ell)}}), \qquad q^{(\ell)} = 1-\tau(1-\delta^{(\ell)}), \qquad q_\tau = 1-\tau(1-\bar\delta) < 1.$$

Then

$$D(X^{(\ell)}) \coloneqq \max_{i,j}\|x_i^{(\ell)} - x_j^{(\ell)}\| \leqslant D_0\prod_{r=0}^{\ell-1} q^{(r)} \leqslant q_\tau^\ell D_0. \tag{16.51}$$

All particles converge to a common $x_\infty^{\mathrm{d}} \in C_0$, and

$$\max_i \|x_i^{(\ell)} - x_\infty^{\mathrm{d}}\| \leqslant \frac{D_0}{1-\bar\delta}q_\tau^\ell.$$

The characteristic solution of

$$\partial_t\alpha_t + \operatorname{div}\big(\alpha_t M_K[\alpha_t]\big) = 0, \qquad \alpha_{t=0} = \alpha_0, \tag{16.52}$$

is globally defined and has nested convex hulls,

$$\operatorname{conv}(\operatorname{supp}\alpha_t) \subset \operatorname{conv}(\operatorname{supp}\alpha_s) \qquad (t\geqslant s\geqslant 0).$$

Writing $D(t) = \operatorname{diam}(\operatorname{supp}\alpha_t)$, one has the adaptive estimate

$$D(t) \leqslant D(s)\exp\!\left(-\int_s^t \big[1-\delta(\mathsf{P}_{\alpha_r})\big]\mathrm{d}r\right) \leqslant e^{-(1-\bar\delta)(t-s)}D(s) \qquad (t\geqslant s\geqslant 0). \tag{16.53}$$

Consequently there exists $x_\infty^{\mathrm{c}} \in C_0$ such that

$$\mathcal{W}_\infty(\alpha_t,\delta_{x_\infty^{\mathrm{c}}}) \leqslant \frac{D_0}{1-\bar\delta}e^{-(1-\bar\delta)t}. \tag{16.54}$$

Proof. Because K is positive and Lipschitz on the compact set $C_0 \times C_0$, its minimum is positive. Hence the denominator in (16.37) is uniformly bounded away from zero, and the characteristic vector field is uniformly Lipschitz in x and Lipschitz in α for $\mathcal{W}_1$. Moreover, $x + M_K[\alpha](x) = \mathsf{P}_\alpha \mathrm{Id}(x)$ lies in $\mathrm{conv}(\mathrm{supp}\,\alpha)$. No characteristic can cross an outward supporting hyperplane of the current convex hull, so the convex hulls are nested and the dynamics is globally well posed.

For any compact set $S \subset \mathbb{R}^d$, the support-function identity

$$\mathrm{diam}(S) = \sup_{\theta \in \mathbb{S}^{d-1}} \left(\sup_{x \in S} \langle \theta,\, x \rangle - \inf_{x \in S} \langle \theta,\, x \rangle \right) \tag{16.55}$$

reduces Euclidean diameter contraction to scalar variation contraction. For a finite configuration, set $z_\theta = X\theta \in \mathbb{R}^n$. Proposition 16.19 gives

$$\begin{aligned} \|z_\theta^{(\ell+1)}\|_V &\leqslant (1-\tau)\|z_\theta^{(\ell)}\|_V + \tau \|\mathsf{P}_{X^{(\ell)}} z_\theta^{(\ell)}\|_V \\ &\leqslant \big[1 - \tau(1 - \delta(\mathsf{P}_{X^{(\ell)}}))\big] \|z_\theta^{(\ell)}\|_V \leqslant q_\tau \|z_\theta^{(\ell)}\|_V. \end{aligned}$$

Taking the supremum over θ gives the one-step factor $q^{(\ell)}$; multiplying these factors and using $q^{(\ell)} \leqslant q_\tau$ proves (16.51). Since $\|x_i^{(\ell+1)} - x_i^{(\ell)}\| \leqslant \tau D(X^{(\ell)})$, every particle path is Cauchy; the vanishing diameter forces a common limit, and summing the uniform geometric tail yields the displayed pointwise estimate.

For the measure-valued flow, let $f_\theta(x) = \langle \theta,\, x \rangle$ and denote by $w_\theta(t) = \|f_\theta\|_{V, \mathrm{supp}\,\alpha_t}$ the width of the support in direction θ. The upper Dini derivative of the extrema along characteristics satisfies

$$D^+ w_\theta(t) \leqslant \|\mathsf{P}_{\alpha_t} f_\theta\|_{V, \mathrm{supp}\,\alpha_t} - w_\theta(t) \leqslant -\big[1 - \delta(\mathsf{P}_{\alpha_t})\big] w_\theta(t).$$

Gronwall's lemma, followed by the support-function identity (16.55), proves (16.53). Finally, every characteristic satisfies $\|\dot{X}_t\| \leqslant D(t)$. It is therefore Cauchy, all characteristic limits coincide because $D(t) \to 0$, and integrating the uniform exponential bound from t to $+\infty$ gives (16.54). □

For the Gaussian kernel $K_\varepsilon(x, y) = e^{-\|x-y\|^2/(2\varepsilon)}$, the cross-ratio has the closed form

$$\log \frac{K_\varepsilon(x, y) K_\varepsilon(x', y')}{K_\varepsilon(x', y) K_\varepsilon(x, y')} = \frac{\langle x - x',\, y - y' \rangle}{\varepsilon}.$$

Hence, for a compact convex set C of diameter D,

$$\eta_{K_\varepsilon}(C) = e^{D^2/\varepsilon}, \qquad \lambda_{K_\varepsilon}(C) = \tanh\!\left(\frac{D^2}{4\varepsilon}\right). \tag{16.56}$$

Combining this identity with (16.48) gives the projectively certified adaptive estimate for the full discrete update,

$$D(X^{(\ell+1)}) \leqslant \tanh\!\left(\frac{D(X^{(\ell)})^2}{4\varepsilon}\right) D(X^{(\ell)}) \leqslant \frac{D(X^{(\ell)})^3}{4\varepsilon},$$

while the continuous flow obeys

$$D(t) \leqslant D(s) \exp\!\left(- \int_s^t \left[1 - \tanh\!\left(\frac{D(r)^2}{4\varepsilon} \right) \right] \mathrm{d}r \right). \tag{16.57}$$

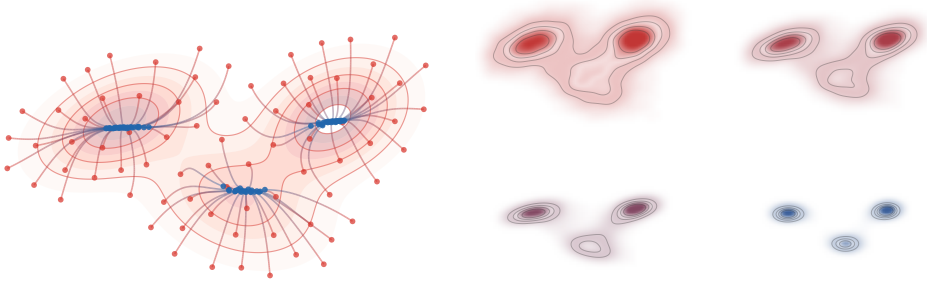
Thus strict positivity gives a global exponential rate, and the certified contraction becomes much stronger once the cloud is narrow relative to the bandwidth. These are Birkhoff bounds; the exact Dobrushin factors can be smaller.

> **Remark 16.21 (Sinkhorn and mean shift share one projective mechanism).** The link with Sinkhorn is exact at the level of positive operators, but the state variables differ. In Theorem 9.17, entrywise multiplication and inversion are Hilbert isometries, while the two multiplications by K and $\mathrm{K}^\top$ each contribute one Birkhoff factor; hence a full scaling cycle contracts rays by $\lambda(\mathrm{K})^2$ as in (9.6). Mean shift performs one row-normalized multiplication by the state-dependent matrix K_X. Diagonal normalization preserves its Birkhoff factor, and row stochasticity fixes the constant ray; the tangent dynamics on $\mathbb{R}^n/\mathrm{Span}(\mathbb{1}_n)$ therefore contracts by the exact factor $\delta(\mathsf{P}_X) \leqslant \lambda(\mathrm{K}_X)$.

> The analogy has precise limits. Sinkhorn compares two scaling iterates for a fixed kernel in the multiplicative projective metric. Mean shift contracts coordinate widths inside one evolving cloud in the additive tangent norm; because $X \mapsto \mathsf{P}_X X$ is state dependent, the theorem does not assert contraction between two different clouds. Nevertheless, both slow regimes have the same origin: when the positive kernel has a large projective diameter, its Birkhoff factor approaches one. This occurs for small entropic temperature in Sinkhorn and for a narrow Gaussian bandwidth relative to the cloud diameter in mean shift.

Scope of the consensus result. The projective argument also clarifies what is not conserved and when clustering replaces consensus. Because row normalization makes the pairwise coefficients asymmetric, the consensus point need not equal the barycenter of α_0; that barycenter is conserved for the constant kernel, but not in general. The self-consistent flow should not be confused with classical mode seeking, where the data measure is frozen and different initial points may converge to different modes. Finally, one-point consensus can fail for the hard confidence kernel $K_R(x,y) = \mathbf{1}_{\{\|x-y\| \leqslant R\}}$: once the interaction graph disconnects, its Dobrushin and Birkhoff coefficients can equal one. The discrete Hegselmann–Krause dynamics then converges to opinion clusters [66]; under the hypotheses of the Eulerian theory, the limiting measure is a finite convex combination of Dirac masses separated by at least R [101]. For the Gaussian kernel, global positivity ultimately forces a single Dirac, although $1 - \tanh(D_0^2/(4\varepsilon))$ can be very small and the multimodal transient can be long.

Figure 16.10 visualizes this transient contraction: particles climb the current smoothed density while the density itself sharpens, well before the eventual consensus guaranteed by Theorem 16.20.



trajectories on the initial density later density snapshots

Figure 16.10: Continuous-time mean shift for a densely sampled three-Gaussian mixture. Left: initial density level sets, in red, and representative particle paths of $\dot{x} = M_\varepsilon[\alpha_t](x)$, colored from red to blue. Right: four later kernel-density renderings of α_t at increasing times, with the same red-to-blue time palette; the initial density is omitted because it is shown on the left. The snapshots show the long multimodal transient before the globally positive Gaussian interactions drive the cloud toward the one-point consensus of Theorem 16.20.

Gradient structure and limitations. When the token space has dimension d and the query/key space has dimension r, take $Q, K \in \mathbb{R}^{r\times d}$ and $V \in \mathbb{R}^{d\times d}$. If $V = Q^\top K$, the field $\Gamma_\theta[\alpha]$ is a gradient vector field in the token variable. Indeed, define the log-partition potential

$$\Phi_\alpha(x) = \int \exp(\langle Qx,\, Ky\rangle)\mathrm{d}\alpha(y), \qquad U_\alpha(x) = \log \Phi_\alpha(x).$$

Then

$$\nabla_x U_\alpha(x) = \frac{\int Q^\top K y \, \exp(\langle Qx,\, Ky\rangle)\mathrm{d}\alpha(y)}{\int \exp(\langle Qx,\, Kz\rangle)\mathrm{d}\alpha(z)} = \Gamma_\theta[\alpha](x).$$

This proves only that the velocity is an instantaneous gradient in x; it does not by itself identify a Wasserstein energy. Indeed, the natural scalar candidate

$$f_{\mathrm{att}}(\alpha) = \int U_\alpha(x)\,\mathrm{d}\alpha(x)$$

has first variation

$$\delta f_{\mathrm{att}}(\alpha)(z) = U_\alpha(z) + \int \frac{\exp(\langle Qx,\, Kz\rangle)}{\Phi_\alpha(x)}\,\mathrm{d}\alpha(x),$$

up to an additive constant. The second term is the response of every query normalization to a perturbation of the key distribution, and its spatial gradient is absent from $\Gamma_\theta[\alpha]$. Thus, without additional symmetry or integrability conditions, the displayed attention PDE is a transportation dynamics rather than the Wasserstein gradient flow of this fixed scalar functional. Special variants recover additional structure: Sinkhorn attention can be interpreted through doubly stochastic normalization and Wasserstein-type gradient flows [468, 117], while layer normalization leads naturally to dynamics on the sphere and to modified metrics. The key open difficulty for the present viewpoint is training: after the architecture has been rewritten as a controlled transport equation, learning corresponds to optimizing the time-dependent parameters $(\theta_t)_t$ rather than merely analyzing the forward PDE for fixed parameters.

16.5 Flows over the Gaussian Manifold

Gaussian measures provide a useful testing ground for the preceding dynamics. They are not invariant under a general Wasserstein gradient flow: a nonlinear velocity usually creates non-Gaussian densities immediately. The useful substitute is to either identify affine velocities, which exactly preserve Gaussianity, or to project the dynamics onto the Gaussian manifold. In both cases the measure PDE reduces to matrix ODEs for the mean and covariance. This viewpoint is emphasized in the survey [426] and is useful for comparing diffusion paths, Wasserstein gradient flows, drifting fields and transformer-type dynamics.

For constrained gradient flows on this family, the covariance equation is the finite-dimensional Bures–Wasserstein gradient flow on positive definite matrices. Thus Gaussian closure is not just a computational shortcut: it is the restriction of Wasserstein geometry to the Gaussian submanifold, where affine gradient fields encode tangent vectors. Figure 16.11 first compares three bridge-type Gaussian closures from a source α_0 to a target γ; the exact gradient-flow closures for specified energies $f(\alpha)$ are catalogued afterwards in Proposition 16.23.

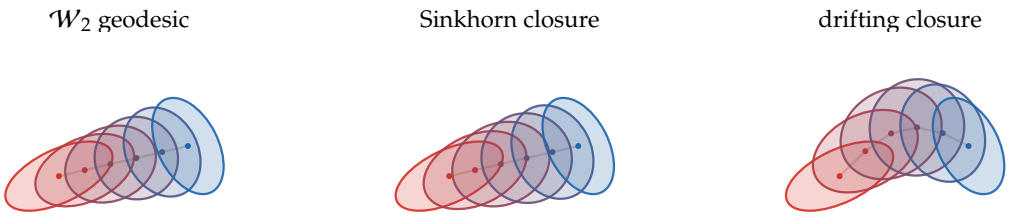


Figure 16.11: Gaussian closures from a red source α_0 to a blue target $\gamma = \mathcal{N}(\bar{m}, \bar{\Sigma})$. The left panel is the constant-speed $\mathcal{W}_2$ geodesic, equivalently the displacement interpolation minimizing the Benamou–Brenier action between α_0 and γ. The middle panel is an entropic Sinkhorn/Schrödinger bridge-style closure for the quadratic cost $|x-y|^2$ and regularization strength $\varepsilon > 0$; it is a bridge toward γ, not the gradient flow of a fixed energy $f(\alpha)$, and the entropic noise inflates intermediate covariances. The right panel is a prescribed non-variational drifting flow, governed by a continuity equation with an affine Gaussian-preserving velocity, chosen so that the mean follows a curved path while the covariance is moment-matched to the same endpoint γ.

Gaussianity preservation. The first question is invariance: one wants a simple criterion ensuring that the continuity equation does not leave the finite-dimensional Gaussian family.

Proposition 16.22 (Affine velocities preserve Gaussianity)**.** *Let $\alpha_0 = \mathcal{N}(m_0, \Sigma_0)$, with $\Sigma_0 \succ 0$. Let $b_t \in \mathbb{R}^d$ and $A_t \in \mathbb{R}^{d\times d}$ be locally integrable on a time interval, and let (m_t, Σ_t) solve*

$$\dot{m}_t = b_t, \qquad \dot{\Sigma}_t = A_t\Sigma_t + \Sigma_t A_t^\top, \qquad (m_{t=0}, \Sigma_{t=0}) = (m_0, \Sigma_0).$$

Then, as long as this matrix ODE is defined, $\Sigma_t \succ 0$ and

$$\alpha_t = \mathcal{N}(m_t, \Sigma_t)$$

is the solution of the continuity equation

$$\partial_t \alpha_t + \operatorname{div}(\alpha_t v_t) = 0, \qquad v_t(x) = b_t + A_t(x - m_t).$$

In particular, if a smooth functional f has a Wasserstein gradient on Gaussian measures of the affine form

$$\mathbb{W}f(\mathcal{N}(m,\Sigma))(x) = b_f(m,\Sigma) + A_f(m,\Sigma)(x-m),$$

with $A_f(m,\Sigma)$ symmetric, then the Wasserstein gradient flow of f, initialized from any non-degenerate Gaussian, stays Gaussian and satisfies

$$\dot{m}_t = h_f(m_t,\Sigma_t), \qquad \dot{\Sigma}_t = H_f(m_t,\Sigma_t),$$

where

$$h_f(m,\Sigma) = -b_f(m,\Sigma), \qquad H_f(m,\Sigma) = -\big(A_f(m,\Sigma)\Sigma + \Sigma A_f(m,\Sigma)\big).$$

Conversely, fix a non-degenerate Gaussian $\mathcal{N}(m,\Sigma)$. Suppose that a finite-energy Wasserstein tangent field $V_{m,\Sigma}$ is represented by a gradient in $L^2(\mathcal{N}(m,\Sigma))$ and is tangent to the Gaussian manifold at this Gaussian. Then there exist $b(m,\Sigma)$ and a symmetric matrix $A(m,\Sigma)$ such that

$$V_{m,\Sigma}(x) = b(m,\Sigma) + A(m,\Sigma)(x-m) \qquad \mathcal{N}(m,\Sigma)\text{-a.e.}$$

Consequently, under the same smoothness assumptions, if the Wasserstein gradient flow of a functional preserves the Gaussian family for all non-degenerate Gaussian initial data, then $\mathbb{W}f(\mathcal{N}(m,\Sigma))$ is affine on each Gaussian. Without the Wasserstein gradient representative, this converse is false because one may add velocity fields with zero $\mathcal{N}(m,\Sigma)$-weighted divergence.

Finally, any smooth Gaussian curve with positive definite covariance can be generated by an affine velocity. If one wants the velocity to be a Wasserstein tangent gradient, one chooses the unique symmetric solution of the Lyapunov equation

$$A_t\Sigma_t + \Sigma_t A_t = \dot{\Sigma}_t.$$

Proof. Let X_t follow the characteristic ODE $\dot{X}_t = b_t + A_t(X_t - m_t)$ with $X_0 \sim \mathcal{N}(m_0,\Sigma_0)$. Since $\dot{m}_t = b_t$, the centered variable $\tilde{X}_t = X_t - m_t$ solves the homogeneous linear ODE $\dot{\tilde{X}}_t = A_t\tilde{X}_t$. If Φ_t is the fundamental matrix $\dot{\Phi}_t = A_t\Phi_t$, $\Phi_{t=0} = \mathrm{Id}$, then

$$X_t = m_t + \Phi_t(X_0 - m_0), \qquad \Sigma_t = \Phi_t\Sigma_0\Phi_t^\top.$$

Hence X_t is Gaussian and $\Sigma_t \succ 0$, and

$$\dot{\Sigma}_t = \frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}\big(\tilde{X}_t\tilde{X}_t^\top\big) = A_t\Sigma_t + \Sigma_t A_t^\top.$$

This proves Gaussian preservation and the moment ODE. The Wasserstein gradient-flow statement follows by inserting the descent velocity $v_t = -\mathbb{W}f(\alpha_t)$.

For the converse, fix $\alpha = \mathcal{N}(m,\Sigma)$ and denote its density by ρ. Tangency to the Gaussian manifold means that the density variation $-\operatorname{div}(\rho V)$ is generated by some moment variation $(\dot{m},\dot{\Sigma})$, with $\dot{\Sigma}$ symmetric. Set $b = \dot{m}$, and let $A = A^\top$ be the unique solution of $A\Sigma + \Sigma A = \dot{\Sigma}$. By the first part of the proposition, the affine gradient field $V_0(x) = b + A(x-m)$ generates exactly the same infinitesimal Gaussian variation. Hence $\operatorname{div}\big(\rho(V - V_0)\big) = 0$ in the distributional sense. Both V and V_0 belong to the $L^2(\alpha)$ closure of gradient fields. The weighted Helmholtz decomposition therefore makes $V - V_0$ simultaneously a tangent gradient and orthogonal to every tangent gradient, so $V = V_0$ in $L^2(\alpha)$. Equivalently, for a smooth representative $V - V_0 = \nabla\psi$, integration by parts gives $\int \|\nabla\psi\|^2 \mathrm{d}\alpha = 0$. This proves that the Wasserstein tangent representative is affine. The qualification is essential: without selecting the gradient representative, one can add a nonzero field with zero α-weighted divergence without changing the Gaussian curve.

For a prescribed smooth Gaussian curve, set $b_t = \dot{m}_t$ and choose any matrix A_t satisfying $A_t\Sigma_t + \Sigma_t A_t^\top = \dot{\Sigma}_t$. Since Σ_t is positive definite, the Lyapunov map $A \mapsto A\Sigma_t + \Sigma_t A$ is invertible on symmetric matrices, which gives the unique symmetric choice when a gradient velocity is required. In that case v_t is the gradient of the quadratic potential $x \mapsto \langle b_t,\, x\rangle + \langle A_t(x-m_t),\, x-m_t\rangle/2$. □

Gaussian-preserving gradient flows. We now instantiate the affine-gradient viewpoint by tracking functionals whose full Wasserstein gradient is affine on Gaussian inputs. The catalogue below contains exact Gaussian-preserving Wasserstein flows, not projected or constrained flows; the separate constrained construction is discussed only afterwards.

Proposition 16.23 (Gaussian closure catalogue). *Let $\gamma = \mathcal{N}(\bar{m}, \bar{\Sigma})$ on $\mathbb{R}^d$, with $\bar{\Sigma} \succ 0$, and let the initial condition be $\alpha_0 = \mathcal{N}(m_0, \Sigma_0)$, with $\Sigma_0 \succ 0$. For each functional displayed below, let α_t be the usual Wasserstein gradient flow on $\mathcal{P}_2(\mathbb{R}^d)$, initialized at α_0. Then the Gaussian family is invariant for these dynamics: as long as the solution exists and $\Sigma_t \succ 0$,*

$$\alpha_t = \mathcal{N}(m_t, \Sigma_t),$$

and the mean and covariance satisfy

$$\dot{m}_t = h(m_t, \Sigma_t), \qquad \dot{\Sigma}_t = H(m_t, \Sigma_t).$$

Write $\delta_m = m - \bar{m}$, $A = \bar{\Sigma}^{-1}$, and

$$M_{\Sigma,\bar{\Sigma}} := \Sigma^{-1/2}\big(\Sigma^{1/2}\bar{\Sigma}\Sigma^{1/2}\big)^{1/2}\Sigma^{-1/2}.$$

With the normalizations displayed in the first column, the mean vector field h and covariance vector field H are listed in the following table. Gradients with respect to Σ are symmetric Frobenius gradients on the cone of covariance matrices.

Gaussian-preserving Wasserstein gradient flows.

Functional $f(\alpha)$	$h(m, \Sigma)$	$H(m, \Sigma)$
$g(m_\alpha, \Sigma_\alpha)$	$-\nabla_m g$	$-2(\Sigma\nabla_\Sigma g + \nabla_\Sigma g\,\Sigma)$
$\int \Big(\frac{1}{2}x^\top Bx + \langle \ell,\, x\rangle\Big)\mathrm{d}\alpha(x)$, $B = B^\top$	$-(Bm + \ell)$	$-(\Sigma B + B\Sigma)$
$\frac{1}{4}\iint (x-y)^\top G(x-y)\mathrm{d}\alpha(x)\mathrm{d}\alpha(y)$, $G = G^\top$	0	$-(\Sigma G + G\Sigma)$
$\mathrm{KL}(\alpha\|\gamma)$	$-A\delta_m$	$2\mathrm{Id} - \Sigma A - A\Sigma$
$\mathcal{I}(\alpha\|\gamma)$	$-2A^2\delta_m$	$4\Sigma^{-1} - 2\Sigma A^2 - 2A^2\Sigma$
$\mathcal{W}_2^2(\alpha, \gamma)$	$-2\delta_m$	$2(M_{\Sigma,\bar{\Sigma}}\Sigma + \Sigma M_{\Sigma,\bar{\Sigma}} - 2\Sigma)$
$\mathrm{MMD}_k^2(\alpha, \gamma)$, $k(x,y) = \langle x,\, y\rangle^2$	$-4Rm$	$-4(\Sigma R + R\Sigma)$
$\bar{\mathcal{L}}^\varepsilon_{\|\cdot-\cdot\|^2}(\alpha, \gamma)$	$-2\delta_m$	$-2(\Sigma G_\varepsilon + G_\varepsilon\Sigma)$
$\mathrm{SW}_2^2(\alpha, \gamma)$	$-\frac{2}{d}\delta_m$	$-2(\Sigma G_{\mathrm{sw}} + G_{\mathrm{sw}}\Sigma)$

Here, in the MMD row,

$$R = \Sigma + m\,m^\top - \bar{\Sigma} - \bar{m}\,\bar{m}^\top.$$

The debiased Sinkhorn row uses the notation of Corollary 9.25: for Gaussian inputs,

$$\bar{\mathcal{L}}^\varepsilon_{\|\cdot-\cdot\|^2}(\alpha, \gamma) = \|\delta_m\|^2 + \mathcal{B}_\varepsilon(\Sigma, \bar{\Sigma})^2,$$

with the closed-form covariance gradient

$$G_\varepsilon(\Sigma, \bar{\Sigma}) = \tau_\varepsilon(\Sigma) - \bar{\Sigma}^{1/2}\tau_\varepsilon\big(B^{1/2}_{\Sigma,\bar{\Sigma}}\big)B^{-1/2}_{\Sigma,\bar{\Sigma}}\bar{\Sigma}^{1/2}, \qquad B_{\Sigma,\bar{\Sigma}} = \bar{\Sigma}^{1/2}\Sigma\bar{\Sigma}^{1/2}.$$

Here τ_ε is the scalar function

$$\tau_\varepsilon(r) := \frac{\sqrt{\varepsilon^2 + 16r^2} - \varepsilon}{4r}, \qquad r > 0,$$

applied to positive matrices by spectral calculus. Equivalently, for $M \succ 0$,

$$\tau_\varepsilon(M) = \big(\sqrt{\varepsilon^2\mathrm{Id} + 16M^2} - \varepsilon\mathrm{Id}\big)(4M)^{-1}.$$

With this convention, $G_\varepsilon = \nabla_\Sigma\mathcal{B}_\varepsilon(\Sigma, \bar{\Sigma})^2$. In the sliced row, σ is the normalized spherical measure on $\mathbb{S}^{d-1}$, and

$$G_{\mathrm{sw}}(\Sigma, \bar{\Sigma}) = \int_{\mathbb{S}^{d-1}} \left(1 - \sqrt{\frac{\theta^\top\bar{\Sigma}\theta}{\theta^\top\Sigma\theta}}\right)\theta\theta^\top\,\mathrm{d}\sigma(\theta).$$

Here

$$\mathcal{I}(\alpha|\gamma) = \int \left|\nabla\log\frac{\rho(x)}{\rho_\gamma(x)}\right|^2\rho(x)\,\mathrm{d}x = \int \big|\nabla\log\rho(x) + A(x-\bar{m})\big|^2\rho(x)\,\mathrm{d}x \qquad (\alpha = \rho\,\mathrm{d}x),$$

where ρ_γ is the density of γ.

Proof. For the moment-functional row, the formula is exact on the full Wasserstein space. Indeed, write

$$m_\alpha = \int x \, \mathrm{d}\alpha(x), \qquad \Sigma_\alpha = \int (x - m_\alpha)(x - m_\alpha)^\top \mathrm{d}\alpha(x).$$

If η is a signed perturbation with zero total mass, then

$$\mathrm{d}m_\alpha[\eta] = \int x \, \mathrm{d}\eta(x), \qquad \mathrm{d}\Sigma_\alpha[\eta] = \int (x - m_\alpha)(x - m_\alpha)^\top \mathrm{d}\eta(x).$$

Thus, up to an irrelevant additive constant, the first variation is

$$\delta f(\alpha)(x) = \langle \nabla_m g, \, x \rangle + \langle \nabla_\Sigma g, \, (x - m_\alpha)(x - m_\alpha)^\top \rangle,$$

and its spatial gradient is the affine field

$$\nabla_x \delta f(\alpha)(x) = \nabla_m g + 2 \nabla_\Sigma g \, (x - m_\alpha).$$

Consequently the full Wasserstein descent velocity is affine, and Proposition 16.22 gives the displayed moment equations.

For the quadratic potential row, the ambient first variation is

$$\delta f(\alpha)(x) = \frac{1}{2} x^\top B x + \langle \ell, \, x \rangle, \qquad \nabla_x \delta f(\alpha)(x) = Bm + \ell + B(x - m),$$

which gives the displayed affine velocity. For the quadratic interaction row, the ambient first variation is

$$\delta f(\alpha)(x) = \frac{1}{2} \int (x - y)^\top G (x - y) \mathrm{d}\alpha(y), \qquad \nabla_x \delta f(\alpha)(x) = G(x - m_\alpha).$$

These first variations are quadratic in x, hence the affine Gaussian flow coincides with the full Wasserstein gradient flow.

For the KL row, write $\alpha = \rho \, \mathrm{d}x$ and let ρ_γ denote the density of γ. The ambient first variation is $\log(\rho/\rho_\gamma)$ up to an additive constant. If $\alpha = \mathcal{N}(m, \Sigma)$, then

$$\nabla \log(\rho/\rho_\gamma)(x) = -\Sigma^{-1}(x - m) + A(x - \bar{m}),$$

so the descent velocity is $v(x) = -A\delta_m + (\Sigma^{-1} - A)(x - m)$. Proposition 16.22 gives $h(m, \Sigma) = -A\delta_m$ and $H(m, \Sigma) = 2\mathrm{Id} - \Sigma A - A\Sigma$. For the Fisher row, writing $u = \log(\rho/\rho_\gamma)$, the ambient first variation of $\frac{1}{2}\mathcal{I}(\alpha|\gamma)$ is, up to constants,

$$-\Delta u - \frac{1}{2}|\nabla u|^2 - \langle \nabla u, \, \nabla \log \rho_\gamma \rangle.$$

For Gaussian ρ and Gaussian γ, this is quadratic in x. Multiplying by two for $\mathcal{I}(\alpha|\gamma)$ gives the descent velocity

$$v(x) = -2A^2\delta_m + 2(\Sigma^{-2} - A^2)(x - m).$$

Hence the Wasserstein velocity is affine and the Gaussian family is closed, with the displayed h and H. Although the ambient Fisher flow is a fourth-order PDE for a generic density, it is therefore an exact finite-dimensional closure in this quadratic-reference case.

For $\mathcal{W}_2^2(\alpha, \gamma)$, the Brenier map from $\mathcal{N}(m, \Sigma)$ to γ is $T(x) = \bar{m} + M_{\Sigma,\bar{\Sigma}}(x - m)$. The descent velocity for the unhalved squared distance is $2(T - \mathrm{Id})$, which gives the mean and covariance equations through Proposition 16.22. For the MMD row, $k(x, y) = \langle x, \, y \rangle^2$ identifies the kernel mean embedding with the raw second moment. Thus

$$\mathrm{MMD}_k^2(\alpha, \gamma) = \|\Sigma + m \, m^\top - \bar{\Sigma} - \bar{m} \, \bar{m}^\top\|_{\mathrm{F}}^2 = \|R\|_{\mathrm{F}}^2,$$

and the ambient first variation is $2x^\top R x$, whose descent velocity is $-4Rx$. This gives $h(m, \Sigma) = -4Rm$ and $H(m, \Sigma) = -4(\Sigma R + R\Sigma)$.

For the Sinkhorn row, Corollary 9.25 gives the closed Gaussian formula as a squared mean displacement plus the smoothed Bures term $\mathcal{B}_\varepsilon^2$. This is not only a formula on the Gaussian submanifold. The first variation of the entropic value with respect to the first marginal is a Sinkhorn dual potential,

and for Gaussian marginals the cross and self dual potentials are quadratic. Their debiased combination is therefore quadratic, so its spatial gradient is affine. Since the restriction of the functional to Gaussians is

$$(m, \Sigma) \mapsto \|m - \bar{m}\|^2 + \mathcal{B}_\varepsilon(\Sigma, \bar{\Sigma})^2,$$

the quadratic coefficient is $G_\varepsilon = \nabla_\Sigma \mathcal{B}_\varepsilon(\Sigma, \bar{\Sigma})^2$, and the moment-functional computation gives the displayed h and H. To compute this gradient, set $B_{\Sigma,\bar{\Sigma}} = \bar{\Sigma}^{1/2}\Sigma\bar{\Sigma}^{1/2}$. Since $\psi_\varepsilon'(r) = -2\tau_\varepsilon(r)$, differentiating $\operatorname{tr}\psi_\varepsilon(B_{\Sigma,\bar{\Sigma}}^{1/2})$ gives

$$-\bar{\Sigma}^{1/2}\tau_\varepsilon\big(B_{\Sigma,\bar{\Sigma}}^{1/2}\big)B_{\Sigma,\bar{\Sigma}}^{-1/2}\bar{\Sigma}^{1/2},$$

whereas differentiating the self term $-\frac{1}{2}\operatorname{tr}\psi_\varepsilon(\Sigma)$ gives $\tau_\varepsilon(\Sigma)$. This proves the displayed closed form for G_ε. When $\bar{\Sigma} = \mathrm{Id}$, this covariance contribution is a spectral function of Σ. If λ is an eigenvalue of Σ, the corresponding scalar velocity is

$$4\sqrt{\lambda + \varepsilon^2/16} - 4\sqrt{\lambda^2 + \varepsilon^2/16},$$

which gives the displayed covariance-eigenvalue equation by functional calculus.

For sliced Wasserstein, the exact ambient Wasserstein gradient is obtained by averaging the one-dimensional gradients. Each projection is Gaussian:

$$(P_\theta)_\sharp\alpha = \mathcal{N}(\langle\theta,\, m\rangle, \theta^\top\Sigma\theta), \qquad (P_\theta)_\sharp\gamma = \mathcal{N}(\langle\theta,\, \bar{m}\rangle, \theta^\top\bar{\Sigma}\theta).$$

Hence

$$\mathrm{SW}_2^2(\alpha, \gamma) = \int_{\mathbb{S}^{d-1}} \left[\langle\theta,\, \delta_m\rangle^2 + \left(\sqrt{\theta^\top\Sigma\theta} - \sqrt{\theta^\top\bar{\Sigma}\theta}\right)^2\right] \mathrm{d}\sigma(\theta).$$

If T_θ is the monotone transport from $(P_\theta)_\sharp\alpha$ to $(P_\theta)_\sharp\gamma$, then the descent velocity for the unhalved sliced objective is

$$2\int_{\mathbb{S}^{d-1}} \big(T_\theta(P_\theta x) - P_\theta x\big)\theta\, \mathrm{d}\sigma(\theta).$$

For Gaussian marginals, T_θ is affine. Thus this velocity is affine in x, and Proposition 16.22 applies. The spherical identity $\int \theta\theta^\top \mathrm{d}\sigma(\theta) = \mathrm{Id}/d$ gives the mean equation, and differentiating the covariance term gives G_{sw}. □

Example 16.24 (Linear mean-field networks as cross-covariance flows). Consider the two-layer mean-field model of Section 15.4, and take the linear activation $\sigma(s) = s$, so that

$$\psi((u, v), z) = v\,\langle u,\, z\rangle.$$

We restrict this example to centered neuron laws,

$$\int (u, v)\mathrm{d}\alpha(u, v) = 0, \qquad \Sigma_\alpha = \begin{pmatrix} \Sigma_{u,u}(\alpha) & \Sigma_{u,v}(\alpha) \\ \Sigma_{v,u}(\alpha) & \Sigma_{v,v}(\alpha) \end{pmatrix},$$

and use the lower-left cross-covariance block

$$\Sigma_{v,u}(\alpha) = \int vu^\top \mathrm{d}\alpha(u, v) \in \mathbb{R}^{d'\times d}.$$

The predictor is therefore the linear map

$$G_{\alpha_t}(z) = \Sigma_{v,u}(\alpha_t)z.$$

For the squared Euclidean loss, set

$$S = \int zz^\top \mathrm{d}\rho(z, y), \qquad R = \int yz^\top \mathrm{d}\rho(z, y).$$

The learning energy of (15.29) is then the covariance functional

$$f(\alpha) = g(\Sigma_\alpha), \qquad g(\Sigma) = \frac{1}{2}\operatorname{tr}\big(\Sigma_{v,u}S\Sigma_{u,v}\big) - \operatorname{tr}\big(R\Sigma_{u,v}\big) + \frac{1}{2}\int \|y\|^2 \mathrm{d}\rho(z, y).$$

This puts the model exactly in the centered moment-functional row of Proposition 16.23. To see that it recovers the usual particle equation, write

$$E_\alpha := \Sigma_{v,u}(\alpha)S - R.$$

The first variation is

$$\delta f(\alpha)(u,v) = \langle E_\alpha,\, vu^\top \rangle = v^\top E_\alpha u.$$

Hence the particle velocity in parameter space is linear:

$$-\nabla_{(u,v)}\delta f(\alpha)(u,v) = -\begin{pmatrix} 0 & E_\alpha^\top \\ E_\alpha & 0 \end{pmatrix}\begin{pmatrix} u \\ v \end{pmatrix}.$$

Equivalently, at the level of g,

$$\nabla_\Sigma g = \frac{1}{2}\begin{pmatrix} 0 & E_\alpha^\top \\ E_\alpha & 0 \end{pmatrix}.$$

The factor $1/2$ in the covariance gradient comes from the symmetry of Σ: the upper-right block is the transpose of the lower-left block. Substituting this gradient in Proposition 16.23 gives

$$\dot m_t = 0, \qquad \dot\Sigma_t = -(\Sigma_t L_t + L_t \Sigma_t), \qquad L_t = \begin{pmatrix} 0 & E_t^\top \\ E_t & 0 \end{pmatrix}, \qquad E_t = \Sigma_{v,u}(\alpha_t)S - R.$$

Thus a centered Gaussian law of neurons remains centered Gaussian, and the dynamics is driven by the cross-covariance block alone. This exact closure is special to the linear activation; for nonlinear activations, Gaussian closures are usually projections rather than invariant families.

One-dimensional case. Fix a Gaussian target $\gamma = \mathcal{N}(\bar m, \bar\sigma^2)$, with $\bar\sigma > 0$, and write $\alpha = \mathcal{N}(m, \sigma^2)$, with $\sigma > 0$. By Proposition 2.54, the parametrization $(m,\sigma) \mapsto \mathcal{N}(m,\sigma^2)$ identifies the one-dimensional Gaussian family with the open Euclidean half-plane. Hence every Gaussian-preserving functional in Proposition 16.23 is represented by the scalar restriction

$$g_f(m,\sigma) := f\big(\mathcal{N}(m,\sigma^2)\big).$$

Set $\delta = m - \bar m$. For the Sinkhorn row below, with $\varepsilon > 0$, abbreviate

$$\psi_\varepsilon(r) := -\frac{\sqrt{\varepsilon^2 + 16r^2} - \varepsilon}{2} - \frac{\varepsilon}{2}\log\left(\frac{2\varepsilon}{\sqrt{\varepsilon^2 + 16r^2} + \varepsilon}\right), \qquad r > 0.$$

The four restrictions used in Figure 16.12 are then

Functional $f(\alpha)$	Scalar restriction $g_f(m,\sigma)$
$\mathrm{KL}(\alpha \mid \gamma)$	$\log\frac{\bar\sigma}{\sigma} + \frac{\sigma^2+\delta^2}{2\bar\sigma^2} - \frac{1}{2}$
$\mathcal{W}_2^2(\alpha,\gamma)$	$\delta^2 + (\sigma - \bar\sigma)^2$
$\bar{\mathcal{L}}^\varepsilon_{\|\cdot-\cdot\|^2}(\alpha,\gamma)$	$\delta^2 + \psi_\varepsilon(\sigma\bar\sigma) - \frac{1}{2}\psi_\varepsilon(\sigma^2) - \frac{1}{2}\psi_\varepsilon(\bar\sigma^2)$
$\mathcal{I}(\alpha\|\gamma)$	$\frac{\delta^2}{\bar\sigma^4} + \frac{(\sigma^2-\bar\sigma^2)^2}{\sigma^2\bar\sigma^4}$

The Sinkhorn formula is the one-dimensional specialization of Corollary 9.25; since $\psi_\varepsilon(r) \to -2r$ as $\varepsilon \downarrow 0$, it recovers the squared-Wasserstein row. Figure 16.12 compares the four landscapes and their Euclidean gradient-flow trajectories. By Proposition 16.23, these are exact ambient Wasserstein flows rather than the constrained Gaussian approximations introduced next.

Constrained evolution on the Gaussian manifold. The preceding affine-gradient examples have a limited scope: most Wasserstein gradient flows of a functional $f(\alpha)$ are not closed on the Gaussian manifold. A nonlinear ambient velocity creates higher-order moments immediately, so the exact Gaussian closure usually fails. The constrained viewpoint deliberately replaces the full evolution by its projection onto $\mathcal{G}$, forcing the curve to remain Gaussian while keeping the Wasserstein tangent geometry.

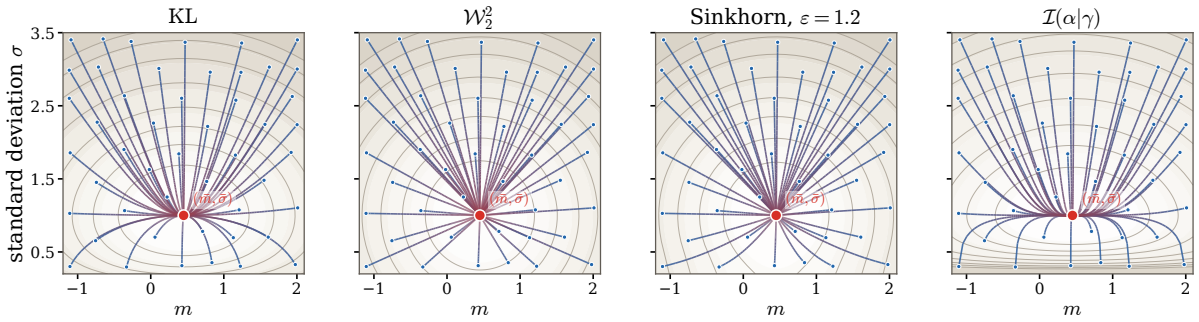


Figure 16.12: One-dimensional Gaussian Wasserstein gradient flows in the Euclidean parameter half-plane (m, σ). From left to right, the panels show relative entropy, squared Wasserstein distance, the debiased Sinkhorn divergence with $\varepsilon = 1.2$, and relative Fisher information. The same 48 initial parameters are marked in blue; the trajectories progress through violet toward the common target $(\bar{m}, \bar{\sigma})$ in red. Each background uses its own level scale to reveal the geometry of the corresponding energy. The $\mathcal{W}_2^2$ trajectories are straight rays, whereas the other functionals bend the evolution of the standard deviation.

Let

$$\mathcal{G} = \{\mathcal{N}(m, \Sigma) : m \in \mathbb{R}^d,\ \Sigma \succ 0\}$$

be the Gaussian submanifold of $\mathcal{P}_2(\mathbb{R}^d)$. The Wasserstein gradient of a functional constrained to a smooth submanifold $\mathcal{M} \subset \mathcal{P}_2$ is defined as the Riesz representative of the differential restricted to tangent velocities of $\mathcal{M}$. Equivalently, it is the small-step limit of the constrained JKO scheme

$$\alpha^{(\ell+1)} \in \operatorname*{argmin}_{\alpha \in \mathcal{M}} \frac{1}{2\tau} \mathcal{W}_2^2(\alpha, \alpha^{(\ell)}) + f(\alpha).$$

For $\mathcal{M} = \mathcal{G}$, tangent velocities are affine gradient fields $v(x) = b + A(x - m)$ with $A = A^\top$. The constrained gradient is therefore the $L^2(\mathcal{N}(m, \Sigma))$ projection of the ambient Wasserstein gradient onto this finite-dimensional affine space, whenever the ambient gradient exists.

Proposition 16.25 (Gaussian-constrained Wasserstein gradients)**.** *Let f be a smooth functional and assume that its restriction to nondegenerate Gaussian measures can be written as*

$$f(\mathcal{N}(m, \Sigma)) = F(m, \Sigma).$$

Then the Wasserstein gradient constrained to the Gaussian family is the affine vector field

$$v_F(x) = \nabla_m F(m, \Sigma) + 2\nabla_\Sigma F(m, \Sigma)(x - m),$$

where $\nabla_\Sigma F$ denotes the symmetric matrix derivative. Equivalently, v_F is the $L^2(\mathcal{N}(m, \Sigma))$ projection of the ambient Wasserstein gradient onto affine gradient fields, whenever the ambient gradient exists. Hence the gradient descent flow constrained to Gaussian measures satisfies

$$\dot{m}_t = -\nabla_m F(m_t, \Sigma_t), \qquad \dot{\Sigma}_t = -2\big(\Sigma_t \nabla_\Sigma F(m_t, \Sigma_t) + \nabla_\Sigma F(m_t, \Sigma_t)\Sigma_t\big), \tag{16.58}$$

and the descent velocity is affine.

Proof. Test the functional along a Gaussian tangent vector, represented by an affine gradient field $v(x) = b + A(x - m)$ with A symmetric. The induced first-order variations are $\dot{m} = b$ and $\dot{\Sigma} = A\Sigma + \Sigma A$. Therefore

$$\mathrm{d}F(m, \Sigma)[b, A\Sigma + \Sigma A] = \langle \nabla_m F,\ b \rangle + \operatorname{tr}(\nabla_\Sigma F(A\Sigma + \Sigma A)).$$

Since A, Σ and $\nabla_\Sigma F$ are symmetric, the second term equals

$$2\operatorname{tr}(\nabla_\Sigma F\, A\Sigma) = \int \langle 2\nabla_\Sigma F(x - m),\ A(x - m) \rangle \mathrm{d}\mathcal{N}(m, \Sigma)(x).$$

Together with the mean term, this gives

$$\mathrm{d}F(m, \Sigma)[\dot{m}, \dot{\Sigma}] = \int \langle v_F(x),\ v(x) \rangle \mathrm{d}\mathcal{N}(m, \Sigma)(x)$$

for all affine gradient fields v. This identifies the constrained Wasserstein gradient in the induced $L^2(\alpha)$ metric, or equivalently the projection of the ambient gradient when it exists. Substituting the descent velocity $-v_F$ in Proposition 16.22 gives (16.58). □

This proposition is the organizing rule for constrained Gaussian closures: once the scalar energy has been reduced to a function of (m, Σ), its constrained Wasserstein gradient is automatically affine and the covariance follows the Bures-type ODE (16.58). When the first variation of f is quadratic, this constrained gradient coincides with the full Wasserstein gradient.

Non-variational Gaussian-preserving flows. The last examples are not ordinary gradient flows of a fixed scalar energy on the full Wasserstein space. They preserve Gaussianity because the prescribed velocity field is affine when evaluated on Gaussian measures.

Example 16.26 (Flow matching and diffusion paths between Gaussians). Consider a prescribed Gaussian interpolation $\alpha_t = \mathcal{N}(m_t, \Sigma_t)$. Proposition 16.22 shows that an exact flow-matching velocity can be taken affine:

$$v_t(x) = \dot{m}_t + A_t(x - m_t), \qquad A_t\Sigma_t + \Sigma_t A_t = \dot{\Sigma}_t.$$

In the isotropic case $\Sigma_t = s_t^2\mathrm{Id}$, this reduces to the transparent formula

$$v_t(x) = \dot{m}_t + \frac{\dot{s}_t}{s_t}(x - m_t).$$

For instance, the diffusion noising path

$$X_t = a_t X_0 + \sigma_t Z, \qquad Z \sim \mathcal{N}(0, \mathrm{Id}),$$

has $m_t = a_t m_0$ and $\Sigma_t = a_t^2\Sigma_0 + \sigma_t^2\mathrm{Id}$. Thus, in the Gaussian case, diffusion paths and flow-matching paths reduce to the same mean-covariance bookkeeping, although the corresponding training objectives are different.

Example 16.27 (Gaussian kernel drifting). Let the target be $\gamma = \mathcal{N}(\bar{m}, \bar{\Sigma})$ and assume $\alpha_t = \mathcal{N}(m_t, \Sigma_t)$. For the Gaussian kernel

$$K_\varepsilon(x, y) = \exp(-\|x - y\|^2/(2\varepsilon)),$$

the normalized field (16.25) satisfies

$$B_\varepsilon[\alpha_t](x) = -\varepsilon(\Sigma_t + \varepsilon\mathrm{Id})^{-1}(x - m_t).$$

Indeed the smoothed density $x \mapsto \int K_\varepsilon(x, y)\mathrm{d}\alpha_t(y)$ is proportional to the Gaussian density with mean m_t and covariance $\Sigma_t + \varepsilon\mathrm{Id}$. Thus $B_\varepsilon[\alpha_t]$ is the mean-shift vector of a Gaussian density: it points linearly toward the Gaussian mode, with strength set by the bandwidth. The drifting velocity (16.27) is therefore the difference of two affine mean-shift fields; it is affine and preserves Gaussianity. With

$$A_t = (\Sigma_t + \varepsilon\mathrm{Id})^{-1}, \qquad \bar{A} = (\bar{\Sigma} + \varepsilon\mathrm{Id})^{-1},$$

the ODE is

$$\dot{m}_t = \varepsilon\bar{A}(\bar{m} - m_t), \qquad \dot{\Sigma}_t = \varepsilon\big((A_t - \bar{A})\Sigma_t + \Sigma_t(A_t - \bar{A})\big).$$

This finite-dimensional model explains the stabilizing role of the self-normalized repulsion term in drifting: without it, the covariance equation loses the $A_t\Sigma_t + \Sigma_t A_t$ contribution.

Example 16.28 (Gaussian closure of attention dynamics). For the transformer PDE, assume $\alpha = \mathcal{N}(m, \Sigma)$. Since exponential tilting preserves Gaussianity,

$$\frac{\int e^{\langle Qx, Ky\rangle}\, y\, \mathrm{d}\alpha(y)}{\int e^{\langle Qx, Kz\rangle}\, \mathrm{d}\alpha(z)} = m + \Sigma K^\top Q x.$$

Therefore

$$\Gamma_\theta[\alpha](x) = Vm + V\Sigma K^\top Q x$$

is affine. The Gaussian token law is preserved and satisfies

$$\dot{m}_t = (V_t + V_t\Sigma_t K_t^\top Q_t)m_t, \qquad \dot{\Sigma}_t = B_t\Sigma_t + \Sigma_t B_t^\top, \qquad B_t = V_t\Sigma_t K_t^\top Q_t.$$

When $V_t = Q_t^\top K_t$, the matrix $B_t = Q_t^\top K_t \Sigma_t K_t^\top Q_t$ is symmetric positive semidefinite, matching the gradient-field case mentioned above. This closure is not a convergence theorem for trained transformers. It is instead a tractable model of how attention can shear, amplify or contract a cloud of tokens through its covariance.

Contractive Gaussian projection. The preceding examples show when Gaussianity is preserved or imposed by projection. Gelbrich's inequality [236] gives a useful variational explanation: replacing a measure by the Gaussian with the same first two moments cannot increase its Wasserstein distance to another similarly projected measure.

Definition 16.29 (Gaussian projection)**.** For $\alpha \in \mathcal{P}_2(\mathbb{R}^d)$, its Gaussian projection is

$$\mathrm{Gauss}(\alpha) := \mathcal{N}(m_\alpha, \Sigma_\alpha),$$

the Gaussian with the same mean and covariance as α.

Theorem 16.30 (Gelbrich theorem)**.** *For every $\alpha, \beta \in \mathcal{P}_2(\mathbb{R}^d)$,*

$$\mathcal{W}_2^2\big(\mathrm{Gauss}(\alpha), \mathrm{Gauss}(\beta)\big) = \|m_\alpha - m_\beta\|^2 + \mathcal{B}^2(\Sigma_\alpha, \Sigma_\beta) \leqslant \mathcal{W}_2^2(\alpha, \beta).$$

Proof. Take any coupling (X, Y) of α and β, center the variables, and write $C = \mathbb{E}[(X - m_\alpha)(Y - m_\beta)^\top]$. In the positive definite case, positivity of the block covariance matrix implies the factorization $C = \Sigma_\alpha^{1/2} K \Sigma_\beta^{1/2}$ with $\|K\|_{\mathrm{op}} \leqslant 1$, and therefore, by operator/nuclear norm duality,

$$\operatorname{tr} C \leqslant \operatorname{tr}\left((\Sigma_\alpha^{1/2} \Sigma_\beta \Sigma_\alpha^{1/2})^{1/2}\right).$$

The semidefinite case follows by adding $\eta\mathrm{Id}$ to both covariance matrices and letting $\eta \downarrow 0$. Expanding $\mathbb{E}\|X - Y\|^2$ gives the lower bound

$$\mathbb{E}\|X - Y\|^2 \geqslant \|m_\alpha - m_\beta\|^2 + \mathcal{B}^2(\Sigma_\alpha, \Sigma_\beta).$$

Taking the infimum over couplings proves the inequality, while equality for Gaussian laws is Proposition 2.54. □

Proposition 12.8 is an earlier application of this contraction: because every Gaussian input is fixed by Gauss, projecting any competitor cannot increase the quadratic Wasserstein barycenter objective.

The following preservation criterion is a direct consequence of Gelbrich's theorem and was explained to us by Hugo Lavenant. It says that a functional which does not increase under moment-matched Gaussian projection admits Gaussian minimizing movements from Gaussian initial data.

Theorem 16.31 (Hugo Lavenant Gaussian-preservation criterion)**.** *Let $f : \mathcal{P}_2(\mathbb{R}^d) \to (-\infty, +\infty]$ satisfy*

$$f\big(\mathrm{Gauss}(\alpha)\big) \leqslant f(\alpha) \qquad \forall \alpha \in \mathcal{P}_2(\mathbb{R}^d),$$

with Gauss *defined in Theorem 16.30. If γ is Gaussian and $\alpha^\star$ minimizes the JKO step*

$$\eta \mapsto f(\eta) + \frac{1}{2\tau}\mathcal{W}_2^2(\gamma, \eta),$$

then $\mathrm{Gauss}(\alpha^\star)$ *is also a minimizer. If this JKO minimizer is unique, it is Gaussian. Consequently, if every step from Gaussian data has a unique minimizer and the resulting minimizing movements converge in $\mathcal{W}_2$, each discrete iterate is Gaussian and every limit curve is Gaussian as well, possibly with a singular limiting covariance.*

Proof. For the JKO claim, $\mathrm{Gauss}(\gamma) = \gamma$ because γ is Gaussian. Hence, for any competitor η,

$$f\big(\mathrm{Gauss}(\eta)\big) + \frac{1}{2\tau}\mathcal{W}_2^2\big(\gamma, \mathrm{Gauss}(\eta)\big) \leqslant f(\eta) + \frac{1}{2\tau}\mathcal{W}_2^2(\gamma, \eta).$$

Applying this to a minimizer $\eta = \alpha^\star$ shows that $\mathrm{Gauss}(\alpha^\star)$ is again a minimizer. Uniqueness forces $\alpha^\star = \mathrm{Gauss}(\alpha^\star)$. □

Moment closure beyond Gaussianity. Gaussian preservation concerns the full law, whereas moment closure asks only whether selected statistics obey an autonomous system. The first row of Proposition 16.23 in fact gives a distribution-free closure. If $f(\alpha) = g(m_\alpha, \Sigma_\alpha)$ and $G = \nabla_\Sigma g(m_\alpha, \Sigma_\alpha)$ denotes the symmetric Frobenius gradient, then the Wasserstein descent velocity is $v_\alpha(x) = -\nabla_m g(m_\alpha, \Sigma_\alpha) - 2G(x - m_\alpha)$. It is affine in x for every α, not only for Gaussian laws. Consequently, every sufficiently regular flow starting from an arbitrary $\alpha_0 \in \mathcal{P}_2(\mathbb{R}^d)$ satisfies

$$\dot{m}_t = -\nabla_m g(m_t, \Sigma_t), \qquad \dot{\Sigma}_t = -2\big(\Sigma_t \nabla_\Sigma g(m_t, \Sigma_t) + \nabla_\Sigma g(m_t, \Sigma_t)\Sigma_t\big). \tag{16.59}$$

For Gaussian initial data, this affine velocity also preserves Gaussianity; for general initial data, the law need not become Gaussian, but its mean and covariance still follow the same closed vector field (h, H) listed in that proposition.

This suggests asking which linear statistics $\alpha \mapsto \int \varphi \, \mathrm{d}\alpha$ admit an exact closure. For scalar statistics, the answer is completely characterized by an eikonal equation.

Proposition 16.32 (Scalar moment closure and eikonal characterization)**.** *Let $\Omega \subset \mathbb{R}^d$ be connected and open, let $\varphi \in C^2(\Omega)$, and set $m_\varphi(\alpha) := \int_\Omega \varphi(x)\mathrm{d}\alpha(x)$ for probability measures such that φ and $\|\nabla\varphi\|^2$ are integrable. For every smooth $g : \mathbb{R} \to \mathbb{R}$, the Wasserstein gradient flow of $f_g(\alpha) = g(m_\varphi(\alpha))$ satisfies, whenever the differentiation is justified,*

$$\dot{m}_t = -g'(m_t) \int_\Omega \|\nabla\varphi(x)\|^2 \mathrm{d}\alpha_t(x), \qquad m_t = m_\varphi(\alpha_t).$$

This equation is autonomous for every g, meaning that its right-hand side depends on α_t only through m_t, if and only if there are constants $a, b \in \mathbb{R}$ such that

$$\|\nabla\varphi(x)\|^2 = a + b\varphi(x) \qquad (x \in \Omega). \tag{16.60}$$

In that case, $\dot{m}_t = -(a + bm_t)g'(m_t)$. Moreover, on every connected open set $U \subset \{x : \nabla\varphi(x) \neq 0\}$, the restriction of (16.60) to U is equivalent to the existence of an eikonal coordinate $r \in C^2(U)$ and constants $c, \kappa, \lambda \in \mathbb{R}$ such that

$$\|\nabla r\| = 1, \qquad \varphi = c + \kappa r + \frac{\lambda}{2} r^2. \tag{16.61}$$

Under (16.60), one may take $\lambda = b/2$.

Proof. The first variation is $\delta f_g(\alpha)(x) = g'(m_\varphi(\alpha))\varphi(x)$, so the continuity equation gives the displayed evolution of m_t. Condition (16.60) then proves sufficiency.

For necessity, it is enough to take $g(s) = s$ and write $\psi = \|\nabla\varphi\|^2$. Autonomy for Dirac laws first shows that $\psi = q \circ \varphi$ for some function q on $\varphi(\Omega)$. Because Ω is connected, $\varphi(\Omega)$ is an interval. Comparing a Dirac law with any two-point mixture having the same φ-moment shows that

$$q\big((1-\zeta)s_0 + \zeta s_1\big) = (1-\zeta)q(s_0) + \zeta q(s_1) \qquad (0 \leqslant \zeta \leqslant 1).$$

Hence q is affine, which is precisely (16.60).

On a regular set U, the right-hand side of (16.60) is positive. Fix $s_0 \in \varphi(U)$ and define

$$r(x) = \int_{s_0}^{\varphi(x)} \frac{\mathrm{d}s}{\sqrt{a + bs}}.$$

Then $\|\nabla r\| = 1$. Inverting this change of variables gives (16.61) with $c = s_0$, $\kappa = \sqrt{a + bs_0}$, and $\lambda = b/2$. Conversely, (16.61) gives $\|\nabla\varphi\|^2 = \kappa^2 - 2\lambda c + 2\lambda\varphi$. □

Example 16.33 (A closed signed-distance moment). Let $\mathcal{S}$ be a C^2 hypersurface and let $r_\mathcal{S}$ be its signed distance on a sufficiently small tubular neighborhood U, where $\|\nabla r_\mathcal{S}\| = 1$. For $\varphi(x) = r_\mathcal{S}(x) + \lambda r_\mathcal{S}(x)^2/2$, one has $\|\nabla\varphi\|^2 = 1 + 2\lambda\varphi$. Consequently, for any solution whose support remains in U, the scalar moment obeys $\dot{m}_t = -(1 + 2\lambda m_t)g'(m_t)$ under the objective $f_g(\alpha) = g(\int \varphi \, \mathrm{d}\alpha)$. For $\lambda = 0$ and a genuinely curved $\mathcal{S}$, the observable is generally non-polynomial and the descent velocity is proportional to the non-affine normal field $\nabla r_\mathcal{S}$. This closure therefore does not arise from the mean–covariance mechanism above.

Conclusion

Optimal transport is useful because it keeps several viewpoints active at once. It is a linear program over couplings, a duality theory for potentials, a geometry on probability measures, a source of PDEs and gradient flows, and a computational toolbox built around linear programming, Sinkhorn scaling and low-dimensional projections. These viewpoints reinforce each other: Brenier maps explain geodesics, geodesic convexity explains convergence of flows, entropic regularization turns transport into scalable differentiable losses, and dual norms or sliced variants reveal what is gained and lost when OT is simplified.

For machine learning, this interplay is especially stimulating. Modern generative modeling, diffusion and flow matching, inverse problems, robust optimization, and even continuous-depth views of transformers all ask for ways to move, compare or learn distributions in high dimension. These applications do not merely consume existing OT theory; they create difficult mathematical questions because empirical measures are singular, dimensions are large, models are parametrized and non-convex, and computational approximations are inseparable from statistical error. The strength of OT is precisely that it gives a common language for these tensions, while still leaving enough open structure for new mathematics to be needed.

Bibliography

[1] Martial Agueh and Guillaume Carlier. Barycenters in the Wasserstein space. *SIAM Journal on Mathematical Analysis*, 43(2):904–924, 2011.

[2] Martial Agueh and Guillaume Carlier. Vers un théorème de la limite centrale dans l'espace de Wasserstein? *Comptes Rendus Mathematique*, 355(7):812–818, 2017.

[3] Michael S. Albergo, Nicholas M. Boffi, and Eric Vanden-Eijnden. Stochastic interpolants: A unifying framework for flows and diffusions. *Journal of Machine Learning Research*, 26(209):1–80, 2025.

[4] Syed Mumtaz Ali and Samuel D Silvey. A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society. Series B (Methodological)*, 28(1):131–142, 1966.

[5] Jason Altschuler, Francis Bach, Alessandro Rudi, and Jonathan Niles-Weed. Massively scalable Sinkhorn distances via the Nyström method. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

[6] Jason Altschuler, Francis Bach, Alessandro Rudi, and Jonathan Weed. Approximating the quadratic transportation metric in near-linear time. *arXiv preprint arXiv:1810.10046*, 2018.

[7] Jason Altschuler, Jonathan Weed, and Philippe Rigollet. Near-linear time approximation algorithms for optimal transport via Sinkhorn iteration. In *Advances in Neural Information Processing Systems*, volume 30, pages 1964–1974, 2017.

[8] Jason M. Altschuler and Enric Boix-Adsera. Polynomial-time algorithms for multimarginal optimal transport problems with structure. *Mathematical Programming*, 199(1–2):1107–1178, 2023.

[9] Pedro C Álvarez-Esteban, E del Barrio, JA Cuesta-Albertos, and C Matrán. A fixed-point approach to barycenters in Wasserstein space. *Journal of Mathematical Analysis and Applications*, 441(2):744–762, 2016.

[10] David Alvarez-Melis, Stefanie Jegelka, and Tommi S. Jaakkola. Towards optimal transport with global invariances. In Kamalika Chaudhuri and Masashi Sugiyama, editors, *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 1870–1879. PMLR, 16–18 Apr 2019.

[11] Luigi Ambrosio and Nicola Gigli. A user's guide to optimal transport. In *Modelling and Optimisation of Flows on Networks*, volume 2062 of *Lecture Notes in Mathematics*, pages 1–155. Springer, 2013.

[12] Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient Flows in Metric Spaces and in the Space of Probability Measures*. Springer, 2006.

[13] Brandon Amos, Lei Xu, and J. Zico Kolter. Input convex neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 146–155. PMLR, 2017.

[14] Ethan Anderes, Steffen Borgwardt, and Jacob Miller. Discrete Wasserstein barycenters: optimal transport for discrete data. *Mathematical Methods of Operations Research*, 84(2):389–409, 2016.

[15] Alexandr Andoni, Piotr Indyk, and Robert Krauthgamer. Earth mover distance over high-dimensional spaces. In *Proceedings of the nineteenth annual ACM-SIAM Symposium on Discrete Algorithms*, pages 343–352. Society for Industrial and Applied Mathematics, 2008.

[16] Francisco Andrade, Gabriel Peyré, and Clarice Poon. Sparsistency for inverse optimal transport. In *International Conference on Learning Representations*, 2024.

[17] Francisco Andrade, Gabriel Peyré, and Clarice Poon. Learning from samples: Inverse problems over measures via sharpened Fenchel–Young losses. *arXiv preprint arXiv:2505.07124*, 2025.

[18] David Applebaum. *Lévy Processes and Stochastic Calculus*. Cambridge University Press, 2 edition, 2009.

[19] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 214–223. PMLR, 2017.

[20] David Arthur and Sergei Vassilvitskii. *k*-means++: The advantages of careful seeding. In *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1027–1035, 2007.

[21] Hédy Attouch, Jérôme Bolte, Patrick Redont, and Antoine Soubeyran. Proximal alternating minimization and projection methods for nonconvex problems: an approach based on the Kurdyka-Łojasiewicz inequality. *Math. Oper. Res.*, 35(2):438–457, May 2010.

[22] Hédy Attouch, Juan Peypouquet, and Patrick Redont. Fast convex optimization via inertial dynamics with hessian driven damping, 2016.

[23] Franz Aurenhammer. Power diagrams: properties, algorithms and applications. *SIAM Journal on Computing*, 16(1):78–96, 1987.

[24] Franz Aurenhammer, Friedrich Hoffmann, and Boris Aronov. Minkowski-type theorems and least-squares clustering. *Algorithmica*, 20(1):61–76, 1998.

[25] Haim Avron, Michael Kapralov, Cameron Musco, Christopher Musco, Ameya Velingker, and Amir Zandieh. Random Fourier features for kernel ridge regression: Approximation bounds and statistical guarantees. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 253–262. PMLR, 2017.

[26] Julio Backhoff Veraguas, Mathias Beiglböck, and Gudmund Pammer. Existence, duality, and cyclical monotonicity for weak transport costs. *Calculus of Variations and Partial Differential Equations*, 58(6):203, 2019.

[27] Dominique Bakry and Michel Émery. Diffusions hypercontractives. In *Séminaire de Probabilités XIX, 1983/84*, volume 1123 of *Lecture Notes in Mathematics*, pages 177–206. Springer, 1985.

[28] Dominique Bakry, Ivan Gentil, and Michel Ledoux. *Analysis and Geometry of Markov Diffusion Operators*, volume 348 of *Grundlehren der mathematischen Wissenschaften*. Springer, 2014.

[29] Raphaël Barboni, Gabriel Peyré, and Franccois-Xavier Vialard. Understanding the training of infinitely deep and wide ResNets with conditional optimal transport. *arXiv preprint arXiv:2403.12887*, 2024.

[30] Heinz H Bauschke and Adrian S Lewis. Dykstra's algorithm with Bregman projections: a convergence proof. *Optimization*, 48(4):409–427, 2000.

[31] Martin Beckmann. A continuous model of transportation. *Econometrica*, 20:643–660, 1952.

[32] Mathias Beiglböck, Pierre Henry-Labordère, and Friedrich Penkner. Model-independent bounds for option prices: a mass transport approach. *Finance and Stochastics*, 17(3):477–501, 2013.

[33] Riccardo Bellazzi, Andrea Codegoni, Stefano Gualandi, Giovanna Nicora, and Eleonora Vercesi. The gene mover's distance: Single-cell similarity via optimal transport. *arXiv preprint arXiv:2102.01218*, 2021.

[34] Aurélien Bellet, Amaury Habrard, and Marc Sebban. Metric learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 9(1):1–151, 2015.

[35] Jean-David Benamou and Yann Brenier. A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem. *Numerische Mathematik*, 84(3):375–393, 2000.

[36] Jean-David Benamou and Guillaume Carlier. Augmented lagrangian methods for transport optimization, mean field games and degenerate elliptic equations. *Journal of Optimization Theory and Applications*, 167(1):1–26, 2015.

[37] Jean-David Benamou, Guillaume Carlier, Marco Cuturi, Luca Nenna, and Gabriel Peyré. Iterative Bregman projections for regularized transportation problems. *SIAM Journal on Scientific Computing*, 37(2):A1111–A1138, 2015.

[38] Jean-David Benamou, Guillaume Carlier, Quentin Mérigot, and Edouard Oudet. Discretization of functionals involving the Monge–Ampère operator. *Numerische Mathematik*, 134(3):611–636, 2016.

[39] Jean-David Benamou, Guillaume Carlier, and Luca Nenna. Generalized incompressible flows, multi-marginal transport and Sinkhorn algorithm. *Numerische Mathematik*, 142(1):33–54, 2019.

[40] Jean-David Benamou, Guillaume Carlier, and Filippo Santambrogio. Variational mean field games. In Nicola Bellomo, Pierre Degond, and Eitan Tadmor, editors, *Active Particles, Volume 1: Advances in Theory, Models, and Applications*, Modeling and Simulation in Science, Engineering and Technology, pages 141–171. Birkhäuser/Springer, Cham, 2017.

[41] Jean-David Benamou, Francis Collino, and Jean-Marie Mirebeau. Monotone and consistent discretization of the Monge-Ampere operator. *Mathematics of Computation*, 85(302):2743–2775, 2016.

[42] Jean-David Benamou, Brittany D Froese, and Adam M Oberman. Numerical solution of the optimal transportation problem using the Monge–Ampere equation. *Journal of Computational Physics*, 260:107–126, 2014.

[43] Christian Berg, Jens Peter Reus Christensen, and Paul Ressel. *Harmonic Analysis on Semigroups*. Number 100 in Graduate Texts in Mathematics. Springer Verlag, 1984.

[44] Alain Berlinet and Christine Thomas-Agnan. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Kluwer Academic Publishers, 2003.

[45] Abraham Berman and Naomi Shaked-Monderer. *Completely Positive Matrices*. World Scientific, 2003.

[46] Robert J. Berman. The Sinkhorn algorithm, parabolic optimal transport and geometric Monge–Ampère equations. *Numerische Mathematik*, 145(4):771–836, 2020.

[47] Donald J. Berndt and James Clifford. Using dynamic time warping to find patterns in time series. In Usama M. Fayyad and Ramasamy Uthurusamy, editors, *Knowledge Discovery in Databases: Papers from the 1994 AAAI Workshop*, number WS-94-03 in AAAI Technical Report, pages 359–370. AAAI Press, 1994.

[48] Andrew C. Berry. The accuracy of the Gaussian approximation to the sum of independent variates. *Transactions of the American Mathematical Society*, 49(1):122–136, 1941.

[49] Dimitri P Bertsekas. A new algorithm for the assignment problem. *Mathematical Programming*, 21(1):152–171, 1981.

[50] Dimitri P Bertsekas. Auction algorithms for network flow problems: a tutorial introduction. *Computational Optimization and Applications*, 1(1):7–66, 1992.

[51] Dimitri P Bertsekas and Jonathan Eckstein. Dual coordinate step methods for linear network flow problems. *Mathematical Programming*, 42(1):203–243, 1988.

[52] Dimitris Bertsimas and John N Tsitsiklis. *Introduction to Linear Optimization*. Athena Scientific, 1997.

[53] Paul J. Besl and Neil D. McKay. A method for registration of 3-D shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(2):239–256, 1992.

[54] Rajendra Bhatia, Tanvi Jain, and Yongdo Lim. On the Bures–Wasserstein distance between positive definite matrices. *Expositiones Mathematicae*, 37(2):165–191, 2019.

[55] Rabi N. Bhattacharya and R. Ranga Rao. *Normal Approximation and Asymptotic Expansions*, volume 64 of *Classics in Applied Mathematics*. Society for Industrial and Applied Mathematics, Philadelphia, 2010.

[56] Jérémie Bigot, Elsa Cazelles, and Nicolas Papadakis. Central limit theorems for entropy-regularized optimal transport on finite spaces and statistical applications. *Electronic Journal of Statistics*, 13(2):5120–5150, 2019.

[57] Jérémie Bigot, Raúl Gouet, Thierry Klein, and Alfredo López. Geodesic PCA in the Wasserstein space by convex pca. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 53(1):1–26, 2017.

[58] Jérémie Bigot and Thierry Klein. Characterization of barycenters in the Wasserstein space by averaging optimal transport maps. *ESAIM: Probability and Statistics*, 22:35–57, 2018.

[59] Ugo Bindini, Luigi De Pascale, and Anna Kausamo. On deterministic solutions for multi-marginal optimal transport with Coulomb cost. *arXiv preprint arXiv:2011.05063*, 2020.

[60] Mikołaj Bińkowski, Danica J. Sutherland, Michael Arbel, and Arthur Gretton. Demystifying MMD GANs. In *International Conference on Learning Representations*, 2018.

[61] Garrett Birkhoff. Tres observaciones sobre el algebra lineal. *Universidad Nacional de Tucumán Revista Series A*, 5:147–151, 1946.

[62] Garrett Birkhoff. Extensions of jentzsch's theorem. *Transactions of the American Mathematical Society*, 85(1):219–227, 1957.

[63] Jose Blanchet and Karthyek Murthy. Quantifying distributional model risk via optimal transport. *Mathematics of Operations Research*, 44(2):565–600, 2019.

[64] Mathieu Blondel, Arthur Mensch, and Jean-Philippe Vert. Differentiable divergences between time series. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 3853–3861. PMLR, 2021.

[65] Mathieu Blondel, Vivien Seguy, and Antoine Rolet. Smooth and sparse optimal transport. In *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pages 880–889. PMLR, 2018.

[66] Vincent D. Blondel, Julien M. Hendrickx, and John N. Tsitsiklis. On Krause's multi-agent consensus model with state-dependent connectivity. *IEEE Transactions on Automatic Control*, 54(11):2586–2597, 2009.

[67] Sergey G. Bobkov. Berry–Esseen bounds and Edgeworth expansions in the central limit theorem for transport distances. *Probability Theory and Related Fields*, 170:229–262, 2018.

[68] Sergey G. Bobkov and Michel Ledoux. One-dimensional empirical measures, order statistics, and Kantorovich transport distances. *Memoirs of the American Mathematical Society*, 261(1259), 2019.

[69] Vladimir I. Bogachev. *Measure Theory*. Springer, Berlin, 2007.

[70] Léa Bohbot, Cyril Letrouit, Gabriel Peyré, and Franccois-Xavier Vialard. Token sample complexity of attention. *arXiv preprint arXiv:2512.10656*, 2025.

[71] Emmanuel Boissard, Thibaut Le Gouic, and Jean-Michel Loubes. Distribution's template estimate with Wasserstein metrics. *Bernoulli*, 21(2):740–759, 2015.

[72] Franccois Bolley, Arnaud Guillin, and Cédric Villani. Quantitative concentration inequalities for empirical measures on non-compact spaces. *Probability Theory and Related Fields*, 137(3):541–593, 2007.

[73] Jérôme Bolte and Adrien Blanchet. A family of functional inequalities: Łojasiewicz inequalities and displacement convex functions. *arXiv preprint arXiv:1612.02619*, 2016.

[74] Jérôme Bolte, Aris Daniilidis, Olivier Ley, and Laurent Mazet. Characterizations of Łojasiewicz inequalities: subgradient flows, talweg, convexity. *Transactions of the American Mathematical Society*, 362(6):3319–3363, 2010.

[75] Nicolas Bonneel and Julie Digne. A survey of optimal transport for computer graphics and computer vision. *Computer Graphics Forum*, 42(2):439–460, 2023.

[76] Nicolas Bonneel, Julien Rabin, Gabriel Peyré, and Hanspeter Pfister. Sliced and Radon Wasserstein barycenters of measures. *Journal of Mathematical Imaging and Vision*, 51(1):22–45, 2015.

[77] Nicolas Bonnotte. *Unidimensional and Evolution Methods for Optimal Transportation*. PhD thesis, Université Paris-Sud and Scuola Normale Superiore, 2013. NNT: 2013PA112297. HAL: https://theses.hal.science/tel-00946781v1.

[78] Walter Braun and Klaus Hepp. The Vlasov dynamics and its fluctuations in the $1/N$ limit of interacting classical particles. *Communications in Mathematical Physics*, 56(2):101–113, 1977.

[79] Lev M Bregman. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR Computational Mathematics and Mathematical Physics*, 7(3):200–217, 1967.

[80] Lev M. Bregman, Yair Censor, and Simeon Reich. Dykstra's algorithm as the nonlinear extension of bregman's optimization method. *Journal of Convex Analysis*, 6(2):319–333, 1999.

[81] Yann Brenier. Décomposition polaire et réarrangement monotone des champs de vecteurs. *C. R. Acad. Sci. Paris Sér. I Math.*, 305(19):805–808, 1987.

[82] Yann Brenier. Polar factorization and monotone rearrangement of vector-valued functions. *Communications on Pure and Applied Mathematics*, 44(4):375–417, 1991.

[83] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901, 2020.

[84] Richard A Brualdi. *Combinatorial Matrix Classes*, volume 108. Cambridge University Press, 2006.

[85] Kevin Buchin, André Nusser, and Sampson Wong. Computing continuous dynamic time warping of time series in polynomial time. In Xavier Goaoc and Michael Kerber, editors, *38th International Symposium on Computational Geometry (SoCG 2022)*, volume 224 of *Leibniz International Proceedings in Informatics*, pages 22:1–22:16, Dagstuhl, Germany, 2022. Schloss Dagstuhl – Leibniz-Zentrum für Informatik.

[86] Charlotte Bunne, Andreas Krause, and Marco Cuturi. Supervised training of conditional Monge maps. *arXiv preprint arXiv:2206.14262*, 2022.

[87] Dmitri Burago, Yuri Burago, and Sergei Ivanov. *A Course in Metric Geometry*, volume 33. American Mathematical Society Providence, RI, 2001.

[88] Samuel Burer, Kurt M. Anstreicher, and Mirjam Dür. The difference between 5×5 doubly nonnegative and completely positive matrices. *Linear Algebra and its Applications*, 431(9):1539–1552, 2009.

[89] Donald Bures. An extension of Kakutani's theorem on infinite product measures to the tensor product of semifinite w^*-algebras. *Transactions of the American Mathematical Society*, 135:199–212, 1969.

[90] Rainer Burkard, Mauro Dell'Amico, and Silvano Martello. *Assignment problems*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2009.

[91] Giuseppe Buttazzo, Luigi De Pascale, and Paola Gori-Giorgi. Optimal-transport formulation of electronic density-functional theory. *Physical Review A*, 85(6):062502, 2012.

[92] Maarten Buyl and Tijl De Bie. Optimal transport of classifiers to fairness. *arXiv preprint arXiv:2202.03814*, 2022.

[93] Luis Caffarelli. The Monge-Ampere equation and optimal transportation, an elementary review. *Lecture Notes in Mathematics, Springer-Verlag*, pages 1–10, 2003.

[94] Luis Caffarelli, Mikhail Feldman, and Robert McCann. Constructing optimal maps for Monge's transport problem as a limit of strictly convex costs. *Journal of the American Mathematical Society*, 15(1):1–26, 2002.

[95] Luis A Caffarelli, Sergey A Kochengin, and Vladimir I Oliker. Problem of reflector design with given far-field scattering data. In *Monge Ampère equation: applications to geometry and optimization*, volume 226, page 13, 1999.

[96] Luis A Caffarelli and Robert J McCann. Free boundaries in optimal transport and Monge-Ampère obstacle problems. *Annals of Mathematics*, 171(2):673–730, 2010.

[97] Emanuele Caglioti, Franccois Golse, and Mikaela Iacobelli. A gradient flow approach to quantization of measures. *Mathematical Models and Methods in Applied Sciences*, 25(10):1845–1885, 2015.

[98] Emanuele Caglioti, Franccois Golse, and Mikaela Iacobelli. Quantization of probability distributions and gradient flows in space dimension 2. *Annales de l'Institut Henri Poincaré C, Analyse Non Linéaire*, 35(6):1531–1555, 2018.

[99] Emanuele Caglioti, Franccois Golse, and Thierry Paul. Quantum optimal transport is cheaper. *Journal of Statistical Physics*, 181(1):149–162, 2020.

[100] Tianji Cai, Junyi Cheng, Nathaniel Craig, and Katy Craig. Linearized optimal transport for collider events. *Physical Review D*, 102(11):116019, 2020.

[101] Claudio Canuto, Fabio Fagnani, and Paolo Tilli. An eulerian approach to the analysis of Krause's consensus models. *SIAM Journal on Control and Optimization*, 50(1):243–265, 2012.

[102] Pierre Cardaliaguet and P. Jameson Graber. Mean field games systems of first order. *ESAIM: Control, Optimisation and Calculus of Variations*, 21(3):690–722, 2015.

[103] Eric A Carlen and Jan Maas. An analog of the 2-Wasserstein metric in non-commutative probability under which the fermionic Fokker–Planck equation is gradient flow for the entropy. *Communications in Mathematical Physics*, 331(3):887–926, 2014.

[104] Guillaume Carlier. Remarks on Toland's duality, convexity constraint and optimal transport. *Pacific Journal of Optimization*, 4(3):423–432, 2008.

[105] Guillaume Carlier, Victor Chernozhukov, and Alfred Galichon. Vector quantile regression: An optimal transport approach. *The Annals of Statistics*, 44(3):1165–1192, 2016.

[106] Guillaume Carlier, Victor Chernozhukov, and Alfred Galichon. Vector quantile regression beyond the specified case. *Journal of Multivariate Analysis*, 161:96–102, 2017.

[107] Guillaume Carlier, Lénaïc Chizat, and Maxime Laborde. Displacement smoothness of entropic optimal transport. *ESAIM: Control, Optimisation and Calculus of Variations*, 30:25, 2024.

[108] Guillaume Carlier, Vincent Duval, Gabriel Peyré, and Bernhard Schmitzer. Convergence of entropic schemes for optimal transport and gradient flows. *SIAM Journal on Mathematical Analysis*, 49(2):1385–1418, 2017.

[109] Guillaume Carlier, Katharina Eichinger, and Alexey Kroshnin. Entropic-Wasserstein barycenters: PDE characterization, regularity and CLT. *SIAM Journal on Mathematical Analysis*, 53(5):5880–5914, 2021.

[110] Guillaume Carlier and Ivar Ekeland. Matching for teams. *Economic Theory*, 42(2):397–418, 2010.

[111] Guillaume Carlier, Alessio Figalli, Quentin Mérigot, and Yi Wang. Sharp comparisons between sliced and standard 1-Wasserstein distances. *arXiv preprint arXiv:2510.16465*, 2025.

[112] Guillaume Carlier, Alfred Galichon, and Filippo Santambrogio. From knothe's transport to Brenier's map and a continuation method for optimal transport. *SIAM Journal on Mathematical Analysis*, 41(6):2554–2576, 2010.

[113] Guillaume Carlier, Chloé Jimenez, and Filippo Santambrogio. Optimal transportation with traffic congestion and Wardrop equilibria. *SIAM Journal on Control and Optimization*, 47(3):1330–1350, 2008.

[114] Guillaume Carlier, Paul Pegon, and Luca Tamanini. Convergence rate of general entropic optimal transport costs. *Calculus of Variations and Partial Differential Equations*, 62(4):116, 2023.

[115] René Carmona and Franccois Delarue. *Probabilistic Theory of Mean Field Games with Applications I: Mean Field FBSDEs, Control, and Games*, volume 83 of *Probability Theory and Stochastic Modelling*. Springer, Cham, 2018.

[116] José A Carrillo, Alina Chertock, and Yanghong Huang. A finite-volume method for nonlinear nonlocal equations with a gradient flow structure. *Communications in Computational Physics*, 17:233–258, 1 2015.

[117] Valérie Castin, Pierre Ablin, José Antonio Carrillo, and Gabriel Peyré. A unified perspective on the dynamics of deep transformers. *arXiv preprint arXiv:2501.18322*, 2025.

[118] Valérie Castin, Pierre Ablin, and Gabriel Peyré. How smooth is attention? In *International Conference on Machine Learning*, 2024.

[119] Yair Censor and Simeon Reich. The Dykstra algorithm with Bregman projections. *Communications in Applied Analysis*, 2:407–419, 1998.

[120] Carlo Cercignani, Reinhard Illner, and Mario Pulvirenti. *The Mathematical Theory of Dilute Gases*. Springer, New York, 1994.

[121] Shouvanik Chakrabarti, Yiming Huang, Tongyang Li, Soheil Feizi, and Xiaodi Wu. Quantum Wasserstein generative adversarial networks. *Advances in Neural Information Processing Systems*, 32, 2019.

[122] Timothy M Chan. Optimal output-sensitive convex hull algorithms in two and three dimensions. *Discrete & Computational Geometry*, 16(4):361–368, 1996.

[123] Laetitia Chapel, Mokhtar Z. Alaya, and Gilles Gasso. Partial optimal transport with applications on positive-unlabeled learning. *arXiv preprint arXiv:2002.08276*, 2020.

[124] Jean-Franccois Chassagneux, Łukasz Szpruch, and Alvin Tse. Weak quantitative propagation of chaos via differential calculus on the space of measures. *The Annals of Applied Probability*, 32(3):1929–1969, 2022.

[125] Jannis Chemseddine, Paul Hagemann, Gabriele Steidl, and Christian Wald. Conditional wasserstein distances with applications in bayesian OT flow matching. *Journal of Machine Learning Research*, 26(141):1–47, 2025.

[126] Louis H. Y. Chen, Larry Goldstein, and Qi-Man Shao. *Normal Approximation by Stein's Method*. Springer, 2011.

[127] Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David Duvenaud. Neural ordinary differential equations. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

[128] Ting-Li Chen. On the convergence and consistency of the blurring mean-shift process. *Annals of the Institute of Statistical Mathematics*, 67(1):157–176, 2015.

[129] Yanshuo Chen, Zhengmian Hu, Wei Chen, and Heng Huang. Fast and scalable Wasserstein-1 neural optimal transport solver for single-cell perturbation prediction. *arXiv preprint arXiv:2411.00614*, 2024.

[130] Yongxin Chen, Wilfrid Gangbo, Tryphon T Georgiou, and Allen Tannenbaum. On the matrix Monge-Kantorovich problem. *European Journal of Applied Mathematics*, 31(4):574–600, 2020.

[131] Yongxin Chen, Tryphon T Georgiou, and Michele Pavon. On the relation between optimal transport and Schrödinger bridges: A stochastic control viewpoint. *Journal of Optimization Theory and Applications*, 169(2):671–691, 2016.

[132] Yongxin Chen, Tryphon T Georgiou, and Allen Tannenbaum. Matrix optimal mass transport: a quantum mechanical approach. *arXiv preprint arXiv:1610.03041*, 2016.

[133] Yizong Cheng. Mean shift, mode seeking, and clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(8):790–799, 1995.

[134] Victor Chernozhukov, Alfred Galichon, Marc Hallin, and Marc Henry. Monge–Kantorovich depth, quantiles, ranks and signs. *The Annals of Statistics*, 45(1):223–256, 2017.

[135] Sinho Chewi, Jonathan Niles-Weed, and Philippe Rigollet. *Statistical Optimal Transport*, volume 2364 of *Lecture Notes in Mathematics*. Springer, Cham, 2025.

[136] Lénaïc Chizat. Convergence rates of gradient methods for convex optimization in the space of measures. *Open Journal of Mathematical Optimization*, 3:1–19, 2022.

[137] Lénaïc Chizat. Sparse optimization on measures with over-parameterized gradient descent. *Mathematical Programming*, 194:487–532, 2022.

[138] Lénaïc Chizat and Francis Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

[139] Lénaïc Chizat, Alex Delalande, and Tomas Vaškevičius. Sharper exponential convergence rates for Sinkhorn's algorithm in continuous settings. *Mathematical Programming*, 215(1):809–858, 2026.

[140] Lenaic Chizat, Gabriel Peyré, Bernhard Schmitzer, and Franccois-Xavier Vialard. Scaling algorithms for unbalanced transport problems. *Mathematics of Computation*, 87:2563–2609, 2018.

[141] Lénaïc Chizat, Gabriel Peyré, Bernhard Schmitzer, and Franccois-Xavier Vialard. Unbalanced optimal transport: dynamic and Kantorovich formulation. *Journal of Functional Analysis*, 274(11):3090–3123, 2018.

[142] Lénaïc Chizat, Pierre Roussillon, Flavien Léger, Franccois-Xavier Vialard, and Gabriel Peyré. Faster Wasserstein distance estimation with the Sinkhorn divergence. In *Advances in Neural Information Processing Systems*, volume 33, pages 2257–2269, 2020.

[143] Lénaïc Chizat, Bernhard Schmitzer, Gabriel Peyré, and Franccois-Xavier Vialard. An interpolating distance between optimal transport and Fisher–Rao metrics. *Foundations of Computational Mathematics*, 18(1):1–44, 2018.

[144] Krzysztof Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter Hawkins, Jared Davis, Afroz Mohiuddin, Lukasz Kaiser, David Belanger, Lucy Colwell, and Adrian Weller. Rethinking attention with performers. In *International Conference on Learning Representations*, 2021.

[145] Shui-Nee Chow, Wen Huang, Yao Li, and Haomin Zhou. Fokker-Planck equations for a free energy functional or Markov process on a graph. *Archive for Rational Mechanics and Analysis*, 203(3):969–1008, 2012.

[146] Evgenii Chzhen, Christophe Denis, Mohamed Hebiri, Luca Oneto, and Massimiliano Pontil. Fair regression with Wasserstein barycenters. *arXiv preprint arXiv:2006.07286*, 2020.

[147] Imre Ciszár. Information-type measures of difference of probability distributions and indirect observations. *Studia Scientiarum Mathematicarum Hungarica*, 2:299–318, 1967.

[148] Alexander Cloninger, Keaton Hamm, Varun Khurana, and Caroline Moosmüller. Linearized Wasserstein dimensionality reduction with approximation guarantees. *Applied and Computational Harmonic Analysis*, 74:101718, 2025.

[149] Michael B Cohen, Aleksander Madry, Dimitris Tsipras, and Adrian Vladu. Matrix scaling and balancing via box constrained Newton's method and interior point methods. In *58th IEEE Annual Symposium on Foundations of Computer Science*, pages 902–913. IEEE Computer Society, 2017.

[150] Samuel Cohen, Brandon Amos, and Yaron Lipman. Riemannian convex potential maps. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 2028–2038. PMLR, 2021.

[151] Maria Colombo and Federico Stra. Counterexamples in multimarginal optimal transport with Coulomb cost and spherically symmetric data. *arXiv preprint arXiv:1507.08522*, 2015.

[152] Dorin Comaniciu and Peter Meer. Mean shift: A robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(5):603–619, 2002.

[153] Giovanni Conforti and Luca Tamanini. A formula for the time derivative of the entropic cost and applications. *Journal of Functional Analysis*, 280(11):108964, 2021.

[154] Dario Cordero-Erausquin. Some applications of mass transport to Gaussian-type inequalities. *Archive for Rational Mechanics and Analysis*, 161(3):257–269, 2002.

[155] Dario Cordero-Erausquin and Bo'az Klartag. Moment measures. *Journal of Functional Analysis*, 268(12):3834–3866, 2015.

[156] Dario Cordero-Erausquin, Robert J. McCann, and Michael Schmuckenschläger. A Riemannian interpolation inequality à la Borell, Brascamp and Lieb. *Inventiones Mathematicae*, 146(2):219–257, 2001.

[157] Sueli IR Costa, Sandra A Santos, and João E Strapasson. Fisher information distance: a geometrical reading. *Discrete Applied Mathematics*, 197:59–69, 2015.

[158] Codina Cotar, Gero Friesecke, and Claudia Klüppelberg. Density functional theory and optimal transportation with Coulomb cost. *Communications on Pure and Applied Mathematics*, 66(4):548–599, 2013.

[159] Nicolas Courty, Rémi Flamary, Amaury Habrard, and Alain Rakotomamonjy. Joint distribution optimal transportation for domain adaptation. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 3730–3739, 2017.

[160] Nicolas Courty, Rémi Flamary, Devis Tuia, and Alain Rakotomamonjy. Optimal transport for domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(9):1853–1865, 2017.

[161] Giacomo Cozzi and Filippo Santambrogio. Long-time asymptotics of the sliced-Wasserstein flow. *arXiv preprint arXiv:2405.06313*, 2024.

[162] Harald Cramér and Herman Wold. Some theorems on distribution functions. *Journal of the London Mathematical Society*, s1-11(4):290–294, oct 1936.

[163] Keenan Crane, Clarisse Weischedel, and Max Wardetzky. Geodesics in heat: a new approach to computing distance based on heat flow. *ACM Transactions on Graphics*, 32(5):152:1–152:11, October 2013.

[164] Ashok Cutkosky and Harsh Mehta. Momentum improves normalized SGD. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 2260–2268, 2020.

[165] Marco Cuturi. Fast global alignment kernels. In *Proceedings of the 28th International Conference on Machine Learning*, pages 929–936, 2011.

[166] Marco Cuturi. Sinkhorn distances: lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems 26*, pages 2292–2300, 2013.

[167] Marco Cuturi and David Avis. Ground metric learning. *Journal of Machine Learning Research*, 15:533–564, 2014.

[168] Marco Cuturi and Mathieu Blondel. Soft-DTW: A differentiable loss function for time-series. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 894–903. PMLR, 2017.

[169] Marco Cuturi and Arnaud Doucet. Fast computation of Wasserstein barycenters. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 685–693. PMLR, 2014.

[170] Marco Cuturi and Gabriel Peyré. A smoothed dual approach for variational Wasserstein problems. *SIAM Journal on Imaging Sciences*, 9(1):320–343, 2016.

[171] Marco Cuturi and Gabriel Peyré. Semidual regularized optimal transport. *SIAM Review*, 60(4):941–965, 2018.

[172] Marco Cuturi, Jean-Philippe Vert, Øystein Birkenes, and Tomoko Matsui. A kernel for time series based on global alignments. In *2007 IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 2, pages II–413–II–416. IEEE, 2007.

[173] Bernard Dacorogna and Jürgen Moser. On a partial differential equation involving the Jacobian determinant. *Annales de l'Institut Henri Poincaré C, Analyse non linéaire*, 7(1):1–26, 1990.

[174] Robert Dadashi, Léonard Hussenot, Matthieu Geist, and Olivier Pietquin. Primal Wasserstein imitation learning. *arXiv preprint arXiv:2006.04678*, 2020.

[175] Sanjit Dandapanthula, Aleksandr Podkopaev, Shiva Prasad Kasiviswanathan, Aaditya Ramdas, and Ziv Goldfeld. Optimal transportation and alignment between gaussian measures, 2025.

[176] George B Dantzig. Application of the simplex method to a transportation problem. *Activity Analysis of Production and Allocation*, 13:359–373, 1951.

[177] Nabarun Deb, Promit Ghosal, and Bodhisattva Sen. Rates of estimation of optimal transport maps using plug-in estimators via barycentric projections. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

[178] Evan DeCorte, Fernando Mário de Oliveira Filho, and Frank Vallentin. Complete positivity and distance-avoiding sets. *Mathematical Programming*, 191:487–558, 2022.

[179] Eustasio del Barrio, Alberto González-Sanz, and Jean-Michel Loubes. Central limit theorems for general transportation costs. *arXiv preprint arXiv:2102.06379*, 2021.

[180] Eustasio del Barrio and Jean-Michel Loubes. Central limit theorems for empirical transportation cost in general dimension. *The Annals of Probability*, 47(2):926–951, 2019.

[181] Lorenzo Dello Schiavo, Jan Maas, and Francesco Pedrotti. Local conditions for global convergence of gradient flows and proximal point sequences in metric spaces. *Transactions of the American Mathematical Society*, 377(6):3779–3804, 2024.

[182] Julie Delon, Julien Salomon, and Andrei Sobolevski. Fast transport optimization for Monge costs on the circle. *SIAM Journal on Applied Mathematics*, 70(7):2239–2258, 2010.

[183] Julie Delon, Julien Salomon, and Andrei Sobolevski. Local matching indicators for transport problems with concave costs. *SIAM Journal on Discrete Mathematics*, 26(2):801–827, 2012.

[184] Mingyang Deng, He Li, Tianhong Li, Yilun Du, and Kaiming He. Generative modeling via drifting. *arXiv preprint arXiv:2602.04770*, 2026.

[185] Steffen Dereich, Michael Scheutzow, and Reik Schottstedt. Constructive quantization: Approximation by empirical measures. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 49(4):1183–1203, 2013.

[186] Ishan Deshpande, Yuan-Ting Hu, Ruoyu Sun, Ayis Pyrros, Nasir Siddiqui, Sanmi Koyejo, Zhizhen Zhao, David A. Forsyth, and Alexander G. Schwing. Max-sliced Wasserstein distance and its use for GANs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10648–10656, 2019.

[187] Simone Di Marino, Augusto Gerolin, and Luca Nenna. Optimal transportation theory with repulsive costs. *arXiv preprint arXiv:1506.04565*, 2015.

[188] Peter J. C. Dickinson and Luuk Gijben. On the computational complexity of membership problems for the completely positive cone and its dual. *Computational Optimization and Applications*, 57(2):403–415, 2014.

[189] Jean Dolbeault, Bruno Nazaret, and Giuseppe Savaré. A new class of transport distances between measures. *Calculus of Variations and Partial Differential Equations*, 34(2):193–231, 2009.

[190] Yan Dolinsky and H Mete Soner. Martingale optimal transport and robust hedging in continuous time. *Probability Theory and Related Fields*, 160(1-2):391–427, 2014.

[191] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

[192] Richard M. Dudley. The speed of mean Glivenko-Cantelli convergence. *Annals of Mathematical Statistics*, 40(1):40–50, 1969.

[193] Théo Dumont, Théo Lacombe, and Franccois-Xavier Vialard. On the existence of monge maps for the gromov–wasserstein problem. *Foundations of Computational Mathematics*, 25(2):463–510, 2025.

[194] Andrew B. Duncan, Nikolas Nüsken, and Lukasz Szpruch. On the geometry of stein variational gradient descent. *Journal of Machine Learning Research*, 24(56):1–39, 2023.

[195] Richard Duong, Nicolaj Rux, Viktor Stein, and Gabriele Steidl. Wasserstein gradient flows of MMD functionals with distance kernels under Sobolev regularization. *arXiv preprint arXiv:2411.09848*, 2024.

[196] Arnaud Dupuy, Alfred Galichon, and Yifei Sun. Estimating matching affinity matrices under low-rank constraints. *Information and Inference: A Journal of the IMA*, 8(4):677–689, 2019.

[197] Pavel Dvurechensky, Alexander Gasnikov, and Alexey Kroshnin. Computational optimal transport: Complexity by accelerated gradient descent is better than by sinkhorn's algorithm. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1367–1376. PMLR, 10–15 Jul 2018.

[198] Richard L Dykstra. An iterative procedure for obtaining I-projections onto the intersection of convex sets. *Annals of Probability*, 13(3):975–984, 1985.

[199] Gintare Karolina Dziugaite, Daniel M Roy, and Zoubin Ghahramani. Training generative neural networks via maximum mean discrepancy optimization. In *Uncertainty in Artificial Intelligence-Proceedings of the 31st Conference, UAI 2015*, pages 258–267, 2015.

[200] Weinan E, Jiequn Han, and Qianxiao Li. A mean-field optimal control formulation of deep learning. *Research in the Mathematical Sciences*, 6(1):10, 2019.

[201] Jonathan Eckstein and Dimitri P Bertsekas. On the Douglas-Rachford splitting method and the proximal point algorithm for maximal monotone operators. *Mathematical Programming*, 55:293–318, 1992.

[202] Bradley Efron. Tweedie's formula and selection bias. *Journal of the American Statistical Association*, 106(496):1602–1614, 2011.

[203] Dominik Maria Endres and Johannes E Schindelin. A new metric for probability distributions. *IEEE Transactions on Information Theory*, 49(7):1858–1860, 2003.

[204] Matthias Erbar. The heat equation on manifolds as a gradient flow in the Wasserstein space. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 46(1):1–23, 2010.

[205] Matthias Erbar. Gradient flows of the entropy for jump processes. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 50(3):920–945, 2014.

[206] Matthias Erbar and Jan Maas. Gradient flow structures for discrete porous medium equations. *Discrete and Continuous Dynamical Systems*, 34(4):1355–1374, 2014.

[207] Matthias Erbar, Martin Rumpf, Bernhard Schmitzer, and Stefan Simon. Computation of optimal transport on discrete metric measure spaces. *Numerische Mathematik*, 144(1):157–200, 2020.

[208] Carl-Gustav Esseen. On the Liapunoff limit of error in the theory of probability. *Arkiv for Matematik, Astronomi och Fysik*, 28A(9):1–19, 1942.

[209] Montacer Essid and Michele Pavon. Traversing the Schrödinger bridge strait: Robert Fortet's marvelous proof redux. *Journal of Optimization Theory and Applications*, 181:23–60, 2019.

[210] Lawrence C. Evans. *Partial Differential Equations*, volume 19 of *Graduate Studies in Mathematics*. American Mathematical Society, 2 edition, 2010.

[211] Jiaojiao Fan, Isabel Haasler, Johan Karlsson, and Yongxin Chen. On the complexity of the optimal transport problem with graph-structured cost. *arXiv preprint arXiv:2110.00627*, 2021.

[212] Pedro F. Felzenszwalb and Daniel P. Huttenlocher. Distance transforms of sampled functions. *Theory of Computing*, 8(19):415–428, 2012.

[213] Jean Feydy, Thibault Séjourné, Franccois-Xavier Vialard, Shun-Ichi Amari, Alain Trouvé, and Gabriel Peyré. Interpolating between optimal transport and MMD using Sinkhorn divergences. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 2681–2690. PMLR, 2019.

[214] Alessio Figalli. The optimal partial transport problem. *Archive for Rational Mechanics and Analysis*, 195(2):533–560, 2010.

[215] Alessio Figalli, Young-Heon Kim, and Robert J. McCann. When is multidimensional screening a convex program? *Journal of Economic Theory*, 146(2):454–478, 2011.

[216] Alessio Figalli, Young-Heon Kim, and Robert J. McCann. Regularity of optimal transport maps on multiple products of spheres. *Journal of the European Mathematical Society*, 15(4):1131–1166, 2013.

[217] Rémi Flamary, Nicolas Courty, Alexandre Gramfort, Mokhtar Z. Alaya, Aurélie Boisbunon, Stanislas Chambon, Laetitia Chapel, Adrien Corenflos, Kilian Fatras, Nemo Fournier, Léo Gautheron, Nathalie T. H. Gayraud, Hicham Janati, Alain Rakotomamonjy, Ievgen Redko, Antoine Rolet, Antony Schutz, Vivien Seguy, Danica J. Sutherland, Romain Tavenard, Alexander Tong, and Titouan Vayer. POT: Python optimal transport. *Journal of Machine Learning Research*, 22(78):1–8, 2021.

[218] Aden Forrow, Jan-Christian Hütter, Mor Nitzan, Philippe Rigollet, Geoffrey Schiebinger, and Jonathan Weed. Statistical optimal transport via factored couplings. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 2454–2465. PMLR, 2019.

[219] Robert Fortet. Résolution d'un système d'équations de M. Schrödinger. *Journal de Mathématiques Pures et Appliquées*, 19(1–4):83–105, 1940.

[220] Nicolas Fournier and Arnaud Guillin. On the rate of convergence in Wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, 162(3-4):707–738, 2015.

[221] Joel Franklin and Jens Lorenz. On the scaling of multidimensional matrices. *Linear Algebra and its Applications*, 114:717–735, 1989.

[222] Brittany D Froese and Adam M Oberman. Convergent finite difference solvers for viscosity solutions of the elliptic monge–ampère equation in dimensions two and higher. *SIAM Journal on Numerical Analysis*, 49(4):1692–1714, 2011.

[223] Charlie Frogner, Chiyuan Zhang, Hossein Mobahi, Mauricio Araya, and Tomaso A Poggio. Learning with a Wasserstein loss. In *Advances in Neural Information Processing Systems*, pages 2053–2061, 2015.

[224] Keinosuke Fukunaga and Larry D. Hostetler. The estimation of the gradient of a density function, with applications in pattern recognition. *IEEE Transactions on Information Theory*, 21(1):32–40, 1975.

[225] Alfred Galichon. *Optimal Transport Methods in Economics*. Princeton University Press, 2016.

[226] Alfred Galichon, Pierre Henry-Labordère, and Nizar Touzi. A stochastic control approach to no-arbitrage bounds given marginals, with an application to lookback options. *Annals of Applied Probability*, 24(1):312–336, 2014.

[227] Alfred Galichon and Antoine Jacquet. Substitutability, equilibrium transport, and matching models, 2024.

[228] Alfred Galichon and Bernard Salanié. Matching with trade-offs: revealed preferences over competing characteristics. Technical report, Preprint SSRN-1487307, 2009.

[229] Alfred Galichon, Larry Samuelson, and Lucas Vernet. Monotone comparative statics for equilibrium problems, 2022.

[230] Thomas O Gallouët and Leonard Monsaingeon. A JKO splitting scheme for Kantorovich–Fisher–Rao gradient flows. *SIAM Journal on Mathematical Analysis*, 49(2):1100–1130, 2017.

[231] Wilfrid Gangbo and Robert J McCann. The geometry of optimal transportation. *Acta Mathematica*, 177(2):113–161, 1996.

[232] Wilfrid Gangbo and Andrzej Swiech. Optimal maps for the multidimensional Monge-Kantorovich problem. *Communications on Pure and Applied Mathematics*, 51(1):23–45, 1998.

[233] Rui Gao and Anton J. Kleywegt. Distributionally robust stochastic optimization with Wasserstein distance. *Mathematics of Operations Research*, 48(2):603–655, 2023.

[234] Ankit Garg and Rafael Oliveira. Recent progress on scaling algorithms and applications. *arXiv preprint arXiv:1808.09669*, 2018.

[235] Stéphane Gaubert and Zheng Qu. Dobrushin's ergodicity coefficient for Markov operators on cones. *Integral Equations and Operator Theory*, 81(1):127–150, 2015.

[236] Matthias Gelbrich. On a formula for the l^2 wasserstein metric between measures on euclidean and hilbert spaces. *Mathematische Nachrichten*, 147(1):185–203, 1990.

[237] Aude Genevay, Lénaïc Chizat, Francis Bach, Marco Cuturi, and Gabriel Peyré. Sample complexity of Sinkhorn divergences. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 1574–1583. PMLR, 2019.

[238] Aude Genevay, Marco Cuturi, Gabriel Peyré, and Francis Bach. Stochastic optimization for large-scale optimal transport. In *Advances in Neural Information Processing Systems*, pages 3440–3448, 2016.

[239] Aude Genevay, Gabriel Peyré, and Marco Cuturi. Learning generative models with Sinkhorn divergences. In *Proceedings of the 21st International Conference on Artificial Intelligence and Statistics*, pages 1608–1617, 2018.

[240] Tryphon T Georgiou and Michele Pavon. Positive contraction mappings for classical and quantum Schrödinger systems. *Journal of Mathematical Physics*, 56(3):033301, 2015.

[241] Borjan Geshkovski, Cyril Letrouit, Yury Polyanskiy, and Philippe Rigollet. A mathematical perspective on transformers. *Bulletin of the American Mathematical Society*, 62(3):427–479, 2025.

[242] Ugo Gianazza, Giuseppe Savaré, and Giuseppe Toscani. The Wasserstein gradient flow of the Fisher information and the quantum drift-diffusion equation. *Archive for Rational Mechanics and Analysis*, 194(1):133–220, 2009.

[243] Ziv Goldfeld, Kengo Kato, Gabriel Rioux, and Ritwik Sadhu. Limit theorems for entropic optimal transport maps and sinkhorn divergence. *Electronic Journal of Statistics*, 18(1):980–1041, 2024.

[244] Gene H. Golub and Victor Pereyra. The differentiation of pseudo-inverses and nonlinear least squares problems whose variables separate. *SIAM Journal on Numerical Analysis*, 10(2):413–432, 1973.

[245] Gene H. Golub and Victor Pereyra. Separable nonlinear least squares: The variable projection method and its applications. *Inverse Problems*, 19(2):R1–R26, 2003.

[246] Alberto González-Sanz and Shayan Hundrieser. Weak limits for empirical entropic optimal transport: Beyond smooth costs. *arXiv preprint arXiv:2305.09745*, 2023.

[247] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2672–2680, 2014.

[248] Paula Gordaliza, Eustasio del Barrio, Fabrice Gamboa, and Jean-Michel Loubes. Obtaining fairness using optimal transport theory. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2357–2365. PMLR, 2019.

[249] Paola Gori-Giorgi, Michael Seidl, and Giovanni Vignale. Density-functional theory for strongly interacting electrons. *Physical Review Letters*, 103(16):166402, 2009.

[250] Nathael Gozlan, Cyril Roberto, Paul-Marie Samson, and Prasad Tetali. Kantorovich duality for general transport costs and applications. *Journal of Functional Analysis*, 273(11):3327–3405, 2017.

[251] Siegfried Graf and Harald Luschgy. *Foundations of Quantization for Probability Distributions*, volume 1730 of *Lecture Notes in Mathematics*. Springer, 2000.

[252] Edouard Grave, Armand Joulin, and Quentin Berthet. Unsupervised alignment of embeddings with Wasserstein procrustes. In Kamalika Chaudhuri and Masashi Sugiyama, editors, *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 1880–1890. PMLR, 16–18 Apr 2019.

[253] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(Mar):723–773, 2012.

[254] Arthur Gretton, Kevin Wenliang Li, Alexandre Galashov, James Thornton, Valentin De Bortoli, and Arnaud Doucet. On the wasserstein gradient flow interpretation of drifting models. *arXiv preprint arXiv:2605.05118*, 2026.

[255] Mikhail Gromov. *Metric Structures for Riemannian and Non-Riemannian Spaces*. Progress in Mathematics. Birkhäuser, 2001.

[256] Leonard Gross. Logarithmic Sobolev inequalities. *American Journal of Mathematics*, 97(4):1061–1083, 1975.

[257] Gaoyue Guo and Jan Obłój. Computational methods for martingale optimal transport problems. *The Annals of Applied Probability*, 29(6):3311–3347, 2019.

[258] Vineet Gupta, Tomer Koren, and Yoram Singer. Shampoo: Preconditioned stochastic tensor optimization. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1842–1850, 2018.

[259] Mert Gürbüzbalaban, Umut cSimcsekli, and Lingjiong Zhu. The heavy-tail phenomenon in SGD. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*. PMLR, 2021.

[260] Leonid Gurvits. Classical deterministic complexity of Edmonds' problem and quantum entanglement. In *Proceedings of the Thirty-Fifth Annual ACM Symposium on Theory of Computing*, pages 10–19. ACM, 2003.

[261] Leonid Gurvits. Classical complexity and quantum entanglement. *Journal of Computer and System Sciences*, 69(3):448–484, 2004.

[262] Isabel Haasler, Axel Ringh, Yongxin Chen, and Johan Karlsson. Multi-marginal optimal transport with a tree-structured cost and the Schrödinger bridge problem. *SIAM Journal on Control and Optimization*, 59(4):2428–2453, 2021.

[263] Isabel Haasler, Rahul Singh, Qinsheng Zhang, Johan Karlsson, and Yongxin Chen. Multi-marginal optimal transport and probabilistic graphical models. *arXiv preprint arXiv:2006.14113*, 2020.

[264] Eldad Haber and Lars Ruthotto. Stable architectures for deep neural networks. *Inverse Problems*, 34(1):014004, 2017.

[265] Philip Hall. On representatives of subsets. *Journal of the London Mathematical Society*, s1-10(1):26–30, 1935.

[266] Marc Hallin, Eustasio del Barrio, Juan Cuesta-Albertos, and Carlos Matrán. Distribution and quantile functions, ranks and signs in dimension d: A measure transportation approach. *The Annals of Statistics*, 49(2):1139–1165, 2021.

[267] Jiaqi Han, Puheng Li, Qiushan Guo, Renyuan Xu, Stefano Ermon, and Emmanuel J. Candès. One-step generative modeling via wasserstein gradient flows. *arXiv preprint arXiv:2605.11755*, 2026.

[268] Boris Hanin. Universal function approximation by deep neural nets with bounded width and ReLU activations. *Mathematics*, 7(10):992, 2019.

[269] Leonid G Hanin. Kantorovich-Rubinstein norm and its application in the theory of Lipschitz spaces. *Proceedings of the American Mathematical Society*, 115(2):345–352, 1992.

[270] Daniel Hauer and José Mazón. Kurdyka–łojasiewicz–Simon inequality for gradient flows in metric spaces. *Transactions of the American Mathematical Society*, 372(7):4917–4976, 2019.

[271] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[272] Ping He, Om Khangaonkar, Hamed Pirsiavash, Yikun Bai, and Soheil Kolouri. Sinkhorn-drifting generative models. *arXiv preprint arXiv:2603.12366*, 2026.

[273] Rainer Hegselmann and Ulrich Krause. Opinion dynamics and bounded confidence: Models, analysis and simulation. *Journal of Artificial Societies and Social Simulation*, 5(3), 2002.

[274] Gabor Herman. *Image Reconstruction from Projections: the Fundamentals of Computerized Tomography*. Academic Press, 1980.

[275] Johannes Hertrich, Antonin Chambolle, and Julie Delon. On the relation between rectified flows and optimal transport. In *Advances in Neural Information Processing Systems*, 2025. arXiv:2505.19712.

[276] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

[277] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851, 2020.

[278] Thomas Hofmann, Bernhard Schölkopf, and Alexander J Smola. Kernel methods in machine learning. *Annals of Statistics*, 36(3):1171–1220, 2008.

[279] Bamdad Hosseini, Alexander W. Hsu, and Amirhossein Taghvaei. Conditional optimal transport on function spaces. *SIAM/ASA Journal on Uncertainty Quantification*, 13(1):304–338, 2025.

[280] Franccois Hu, Philipp Ratz, and Arthur Charpentier. Fairness in multi-task learning via Wasserstein barycenters. In *Machine Learning and Knowledge Discovery in Databases: Research Track*, volume 14170 of *Lecture Notes in Computer Science*, pages 295–312. Springer, 2023.

[281] Minyi Huang, Roland P. Malhamé, and Peter E. Caines. Large population stochastic dynamic games: Closed-loop McKean–Vlasov systems and the Nash certainty equivalence principle. *Communications in Information and Systems*, 6(3):221–252, 2006.

[282] Geert-Jan Huizing, Laura Cantini, and Gabriel Peyré. Unsupervised ground metric learning using Wasserstein singular vectors. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 9429–9443. PMLR, 2022.

[283] Shayan Hundrieser, Marcel Klatt, and Axel Munk. Limit distributions and sensitivity analysis for empirical entropic optimal transport on countable spaces. *arXiv preprint arXiv:2105.00049*, 2021.

[284] Jan-Christian Hütter and Philippe Rigollet. Minimax estimation of smooth optimal transport maps. *The Annals of Statistics*, 49(2):1166–1194, 2021.

[285] Aapo Hyvärinen. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6:695–709, 2005.

[286] Mikaela Iacobelli. Asymptotic analysis for a very fast diffusion equation arising from the 1d quantization problem. *Discrete and Continuous Dynamical Systems*, 39(9):4929–4943, 2019.

[287] Mikaela Iacobelli, Francesco Patacchini, and Filippo Santambrogio. Weighted ultrafast diffusion equations: From well-posedness to long-time behaviour. *Archive for Rational Mechanics and Analysis*, 232(3):1165–1206, 2019.

[288] Martin Idel. A review of matrix scaling and Sinkhorn's normal form for matrices and positive maps. *arXiv preprint arXiv:1609.06349*, 2016.

[289] Piotr Indyk and Nitin Thaper. Fast image retrieval via embeddings. In *3rd International Workshop on Statistical and Computational Theories of Vision*, 2003.

[290] Hicham Janati, Marco Cuturi, and Alexandre Gramfort. Spatio-temporal alignments: Optimal transport through space and time. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 1695–1704. PMLR, 2020.

[291] Hicham Janati, Boris Muzellec, Gabriel Peyré, and Marco Cuturi. Entropic optimal transport between unbalanced Gaussian measures has a closed form. In *Advances in Neural Information Processing Systems*, volume 33, pages 10468–10479. Curran Associates, Inc., 2020.

[292] Xianhua Jiang, Lipeng Ning, and Tryphon T Georgiou. Distances and Riemannian metrics for multivariate spectral densities. *IEEE Transactions on Automatic Control*, 57(7):1723–1735, 2012.

[293] Keller Jordan, Yuchen Jin, Vlado Boza, Jiacheng You, Franz Cesista, Laker Newhouse, and Jeremy Bernstein. Muon: An optimizer for hidden layers in neural networks, 2024. Blog post.

[294] Leonid Kantorovich. On the transfer of masses (in russian). *Doklady Akademii Nauk*, 37(2):227–229, 1942.

[295] LV Kantorovich and G.S. Rubinstein. On a space of totally additive functions. *Vestn Leningrad Universitet*, 13:52–59, 1958.

[296] Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the polyak–łojasiewicz condition, 2016.

[297] Samuel Karlin. *Total Positivity, Volume I*. Stanford University Press, 1968.

[298] Johan Karlsson and Axel Ringh. Generalized Sinkhorn iterations for regularizing inverse problems using optimal mass transport. *SIAM Journal on Imaging Sciences*, 10(4):1935–1962, 2017.

[299] Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and Franccois Fleuret. Transformers are RNNs: Fast autoregressive transformers with linear attention. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 5156–5165. PMLR, 2020.

[300] Gavin Kerrigan, Giosuè Migliorini, and Padhraic Smyth. Dynamic conditional optimal transport through simulation-free flows. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

[301] Abdelwahed Khamis, Russell Tsuchida, Mohamed Tarek, Vivien Rolland, and Lars Petersson. Scalable optimal transport methods in machine learning: A contemporary survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[302] Cheol-Taek Kim and Ju-Jang Lee. Training two-layered feedforward networks with variable projection method. *IEEE Transactions on Neural Networks*, 19(2):371–375, 2008.

[303] Hyunjik Kim, George Papamakarios, and Andriy Mnih. The lipschitz constant of self-attention. *arXiv preprint arXiv:2006.04710*, 2020.

[304] Young-Heon Kim and Robert J. McCann. Continuity, curvature, and the general covariance of optimal transportation. *Journal of the European Mathematical Society*, 12(4):1009–1040, 2010.

[305] Young-Heon Kim and Robert J. McCann. Towards the smoothness of optimal maps on riemannian submersions and riemannian products (of round spheres in particular). *Journal für die reine und angewandte Mathematik*, 664:1–27, 2012.

[306] Young-Heon Kim and Brendan Pass. Wasserstein barycenters over Riemannian manifolds. *Advances in Mathematics*, 307:640–683, 2017.

[307] Marcel Klatt, Carla Tameling, and Axel Munk. Empirical regularized optimal transport: Statistical theory and applications. *SIAM Journal on Mathematics of Data Science*, 2(2):419–443, 2020.

[308] Dominik Klein, Théo Uscidda, Fabian Theis, and Marco Cuturi. GENOT: Entropic (Gromov) Wasserstein flow matching with applications to single-cell genomics. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

[309] Philip A Knight. The Sinkhorn–Knopp algorithm: convergence and applications. *SIAM Journal on Matrix Analysis and Applications*, 30(1):261–275, 2008.

[310] Herbert Knothe. Contributions to the theory of convex bodies. *Michigan Mathematical Journal*, 4(1):39–52, 1957.

[311] John E. Kolassa and Peter McCullagh. Edgeworth series for lattice distributions. *The Annals of Statistics*, 18(2):981–985, 1990.

[312] Soheil Kolouri, Se Rim Park, Matthew Thorpe, Dejan Slepcev, and Gustavo K Rohde. Optimal mass transport: signal processing and machine-learning applications. *IEEE Signal Processing Magazine*, 34(4):43–59, 2017.

[313] Soheil Kolouri, Akif B. Tosun, John A. Ozolek, and Gustavo K. Rohde. A continuous linear optimal transport approach for pattern analysis in image datasets. *Pattern Recognition*, 51:453–462, 2016.

[314] Hiroshi Konno. Maximization of a convex quadratic function under linear constraints. *Mathematical Programming*, 11(1):117–127, 1976.

[315] Jonathan Korman and Robert McCann. Optimal transportation with capacity constraints. *Transactions of the American Mathematical Society*, 367(3):1501–1521, 2015.

[316] Bernhard Korte and Jens Vygen. *Combinatorial Optimization*. Springer, 2012.

[317] J. Kruithof. Telefoonverkeersrekening. *De Ingenieur*, 52:E15–E25, 1937.

[318] Harold W. Kuhn. The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1–2):83–97, 1955.

[319] Brian Kulis. Metric learning: a survey. *Foundations and Trends in Machine Learning*, 5(4):287–364, 2012.

[320] Debarshi Kundu, Archisman Ghosh, Swaroop Ghosh, and Vasant Honavar. Projection-free transformers via Gaussian kernel attention. *arXiv preprint arXiv:2605.02144*, 2026.

[321] Krzysztof Kurdyka. On gradients of functions definable in o-minimal structures. *Annales de l'Institut Fourier*, 48(3):769–783, 1998.

[322] Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. From word embeddings to document distances. In *International Conference on Machine Learning*, pages 957–966, 2015.

[323] Jean-Michel Lasry and Pierre-Louis Lions. Mean field games. *Japanese Journal of Mathematics*, 2(1):229–260, 2007.

[324] Hugo Lavenant and Filippo Santambrogio. The flow map of the Fokker–Planck equation does not provide optimal transport. *Applied Mathematics Letters*, 133:108225, 2022.

[325] Hugo Lavenant and Giuseppe Savaré. Continuous transformations of probability measures and their transport representations. *arXiv preprint arXiv:2604.16653*, 2026.

[326] Hugo Lavenant, Stephen Zhang, Young-Heon Kim, and Geoffrey Schiebinger. Towards a mathematical theory of trajectory inference. *arXiv preprint arXiv:2102.09204*, 2021.

[327] Khang Le, Dung Q. Le, Huy Nguyen, Dat Do, Tung Pham, and Nhat Ho. Entropic gromov–wasserstein between gaussian distributions. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 12164–12203. PMLR, 2022.

[328] Tam Le, Makoto Yamada, Kenji Fukumizu, and Marco Cuturi. Tree-sliced variants of Wasserstein distances. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

[329] Thibaut Le Gouic and Jean-Michel Loubes. Existence and consistency of Wasserstein barycenters. *Probability Theory and Related Fields*, 168:901–917, 2016.

[330] Hoai An Le Thi and Tao Pham Dinh. The DC (difference of convex functions) programming and DCA revisited with DC models of real world nonconvex optimization problems. *Annals of Operations Research*, 133(1–4):23–46, 2005.

[331] Michel Ledoux. *The Concentration of Measure Phenomenon*, volume 89 of *Mathematical Surveys and Monographs*. American Mathematical Society, 2001.

[332] Flavien Léger, Gabriele Todeschi, and Franccois-Xavier Vialard. Nonnegative cross-curvature in infinite dimensions: Synthetic definition and spaces of measures. *Geometric and Functional Analysis*, 35(6):1638–1711, 2025.

[333] Tobias Lehmann, Max-K. von Renesse, Adam Sambale, and André Uschmajew. A note on overrelaxation in the Sinkhorn algorithm. *Optimization Letters*, 16(8):2209–2220, 2022.

[334] Jan Lellmann, Dirk A Lorenz, Carola Schönlieb, and Tuomo Valkonen. Imaging with Kantorovich–Rubinstein discrepancy. *SIAM Journal on Imaging Sciences*, 7(4):2833–2859, 2014.

[335] Bas Lemmens and Roger Nussbaum. *Nonlinear Perron-Frobenius Theory*, volume 189. Cambridge University Press, 2012.

[336] Christian Léonard. From the Schrödinger problem to the Monge–Kantorovich problem. *Journal of Functional Analysis*, 262(4):1879–1920, 2012.

[337] Christian Léonard. A survey of the Schrödinger problem and some of its connections with optimal transport. *Discrete Continuous Dynamical Systems Series A*, 34(4):1533–1574, 2014.

[338] Christian Léonard. Revisiting Fortet's proof of existence of a solution to the Schrödinger system. *arXiv preprint arXiv:1904.13211*, 2019.

[339] Matthias Liero, Alexander Mielke, and Giuseppe Savaré. Optimal transport in competition with reaction: the Hellinger–Kantorovich distance and geodesic curves. *SIAM Journal on Mathematical Analysis*, 48(4):2869–2911, 2016.

[340] Matthias Liero, Alexander Mielke, and Giuseppe Savaré. Optimal entropy-transport problems and a new hellinger–kantorovich distance between positive measures. *Inventiones Mathematicae*, 211(3):969–1117, 2018.

[341] Tianyi Lin, Nhat Ho, Marco Cuturi, and Michael I. Jordan. On the complexity of approximating multimarginal optimal transport. *arXiv preprint arXiv:1910.00152*, 2019.

[342] Haibin Ling and Kazunori Okada. An efficient earth mover's distance algorithm for robust histogram comparison. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(5):840–853, 2007.

[343] Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023.

[344] Jingyuan Liu, Jianlin Su, Xingcheng Yao, Zhejun Jiang, Guokun Lai, Yulun Du, Yidao Qin, Weixin Xu, Enzhe Lu, Junjie Yan, Yanru Chen, Huabin Zheng, Yibo Liu, Shaowei Liu, Bohong Yin, Weiran He, Han Zhu, Yuzhi Wang, Jianzhou Wang, Mengnan Dong, Zheng Zhang, Yongsheng Kang, Hao Zhang, Xinran Xu, Yutao Zhang, Yuxin Wu, Xinyu Zhou, and Zhilin Yang. Muon is scalable for LLM training, 2025.

[345] Qiang Liu. Stein variational gradient descent as gradient flow. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

[346] Qiang Liu and Dilin Wang. Stein variational gradient descent: A general purpose bayesian inference algorithm. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

[347] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International Conference on Learning Representations*, 2023.

[348] Stuart Lloyd. Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28(2):129–137, 1982.

[349] Grégoire Loeper. On the regularity of maps solutions of optimal transportation problems. *Acta Mathematica*, 202(2):241–283, 2009.

[350] Grégoire Loeper. Regularity of optimal maps on the sphere: The quadratic cost and the reflector antenna. *Archive for Rational Mechanics and Analysis*, 199(1):269–289, 2011.

[351] Grégoire Loeper and Francesca Rapetti. Numerical solution of the Monge-Ampère equation by a Newton's algorithm. *C. R. Acad. Sci. Paris, Ser. I*, 340(4):319–324, 2005.

[352] Eliane Maria Loiola, Nair Maria Maia de Abreu, Paulo Oswaldo Boaventura-Netto, Peter Hahn, and Tania Querido. A survey for the quadratic assignment problem. *European Journal of Operational Research*, 176(2):657–690, 2007.

[353] Stanisław Łojasiewicz. Une propriété topologique des sous-ensembles analytiques réels. In *Les Équations aux Dérivées Partielles*, pages 87–89. CNRS, Paris, 1963.

[354] John Lott and Cédric Villani. Ricci curvature for metric-measure spaces via optimal transport. *Annals of Mathematics*, 169(3):903–991, 2009.

[355] Yiping Lu, Aoxiao Zhong, Quanzheng Li, and Bin Dong. Beyond finite layer neural networks: Bridging deep architectures and numerical differential equations. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 3276–3285, 2018.

[356] Frederike Lübeck, Charlotte Bunne, Gabriele Gut, Jacobo Sarabia del Castillo, Lucas Pelkmans, and David Alvarez-Melis. Neural unbalanced optimal transport via cycle-consistent semi-couplings. *arXiv preprint arXiv:2209.15621*, 2022.

[357] Yves Lucet. What shape is your conjugate? a survey of computational convex analysis and its applications. *SIAM Review*, 52(3):505–542, 2010.

[358] Vince Lyzinski, Donniell E Fishkind, Marcelo Fiori, Joshua T Vogelstein, Carey E Priebe, and Guillermo Sapiro. Graph matching: relax at your own risk. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(1):60–73, 2016.

[359] Shaojun Ma, Haodong Sun, Xiaojing Ye, Hongyuan Zha, and Haomin Zhou. Learning cost functions for optimal transport. *arXiv preprint arXiv:2002.09650*, 2020.

[360] Xi-Nan Ma, Neil S Trudinger, and Xu-Jia Wang. Regularity of potential functions of the optimal transportation problem. *Archive for Rational Mechanics and Analysis*, 177(2):151–183, 2005.

[361] Jan Maas. Gradient flows of the entropy for finite Markov chains. *Journal of Functional Analysis*, 261(8):2250–2292, 2011.

[362] Jan Maas, Martin Rumpf, Carola Schönlieb, and Stefan Simon. A generalized model for optimal transport of images including dissipation and density modulation. *ESAIM: Mathematical Modelling and Numerical Analysis*, 49(6):1745–1769, 2015.

[363] Jan Maas, Martin Rumpf, and Stefan Simon. Generalized optimal transport with singular sources. *arXiv preprint arXiv:1607.01186*, 2016.

[364] Guillaume Mahey, Laetitia Chapel, Gilles Gasso, Clément Bonet, and Nicolas Courty. Fast optimal transport through sliced Wasserstein generalized geodesics. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

[365] Ashok Vardhan Makkuva, Amirhossein Taghvaei, Sewoong Oh, and Jason D. Lee. Optimal transport mapping via input convex neural networks. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 6672–6681. PMLR, 2020.

[366] Stephan Mandt, Matthew D. Hoffman, and David M. Blei. Stochastic gradient descent as approximate bayesian inference. *Journal of Machine Learning Research*, 18(134):1–35, 2017.

[367] Tudor Manole, Sivaraman Balakrishnan, Jonathan Niles-Weed, and Larry Wasserman. Plugin estimation of smooth optimal transport maps. *The Annals of Statistics*, 52(3):966–998, 2024.

[368] Tudor Manole, Sivaraman Balakrishnan, and Larry Wasserman. Minimax confidence intervals for the sliced Wasserstein distance. *Electronic Journal of Statistics*, 16(1):2252–2345, 2022.

[369] Haggai Maron and Yaron Lipman. (probably) concave graph matching. In *Advances in Neural Information Processing Systems*, volume 31, pages 406–416, 2018.

[370] Bertrand Maury, Aude Roudneff-Chupin, and Filippo Santambrogio. A macroscopic crowd motion model of gradient flow type. *Mathematical Models and Methods in Applied Sciences*, 20(10):1787–1821, 2010.

[371] Robert J McCann. A convexity principle for interacting gases. *Advances in Mathematics*, 128(1):153–179, 1997.

[372] Robert J. McCann. Polar factorization of maps on Riemannian manifolds. *Geometric and Functional Analysis*, 11(3):589–608, 2001.

[373] Song Mei, Andrea Montanari, and Phan-Minh Nguyen. A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences*, 115(33):E7665–E7671, 2018.

[374] Facundo Mémoli. On the use of Gromov–Hausdorff distances for shape comparison. In *Symposium on Point Based Graphics*, pages 81–90. Eurographics Association, 2007.

[375] Facundo Mémoli. Gromov–Hausdorff distances in Euclidean spaces. In *Proceedings of CVPR Workshops*, pages 1–8, 2008.

[376] Facundo Mémoli. Gromov–Wasserstein distances and the metric approach to object matching. *Foundations of Computational Mathematics*, 11(4):417–487, 2011.

[377] Facundo Mémoli. The Gromov–Wasserstein distance: a brief overview. *Axioms*, 3(3), 2014.

[378] Facundo Mémoli and Tom Needham. Comparison results for gromov–wasserstein and gromov–monge distances. *ESAIM: Control, Optimisation and Calculus of Variations*, 30:78, 2024.

[379] Gonzalo Mena and Jonathan Niles-Weed. Statistical bounds for entropic optimal transport: Sample complexity and the central limit theorem. In *Advances in Neural Information Processing Systems*, volume 32, pages 4541–4551, 2019.

[380] Quentin Mérigot. A multiscale approach to optimal transport. *Computer Graphics Forum*, 30(5):1583–1592, 2011.

[381] Quentin Mérigot. A comparison of two dual methods for discrete optimal transport. In *Geometric science of information*, pages 389–396. Springer, 2013.

[382] Quentin Mérigot, Alex Delalande, and Frédéric Chazal. Quantitative stability of optimal transport maps and linearization of the 2-Wasserstein space. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 3186–3196. PMLR, 2020.

[383] Quentin Mérigot and Boris Thibert. Optimal transport: Discretization and algorithms. *arXiv preprint arXiv:2003.00855*, 2020.

[384] Alexander Mielke. Geodesic convexity of the relative entropy in reversible Markov chains. *Calculus of Variations and Partial Differential Equations*, 48(1-2):1–31, 2013.

[385] Jean-Marie Mirebeau. Discretization of the 3D Monge-Ampere operator, between wide stencils and power diagrams. *ESAIM: Mathematical Modelling and Numerical Analysis*, 49(5):1511–1523, 2015.

[386] Peyman Mohajerin Esfahani and Daniel Kuhn. Data-driven distributionally robust optimization using the wasserstein metric: Performance guarantees and tractable reformulations. *Mathematical Programming*, 171(1-2):115–166, 2018.

[387] Gaspard Monge. Mémoire sur la théorie des déblais et des remblais. *Histoire de l'Académie Royale des Sciences*, pages 666–704, 1781.

[388] Eduardo Fernandes Montesuma, Fred Ngolè Mboula, and Antoine Souloumiac. Recent advances in optimal transport for machine learning. *arXiv preprint arXiv:2306.16156*, 2023.

[389] Caroline Moosmüller and Alexander Cloninger. Linear optimal transport embedding: Provable Wasserstein classification for certain rigid transformations and perturbations. *Information and Inference: A Journal of the IMA*, 12(1):363–389, 2023.

[390] Jorge J. Moré and Werner C. Rheinboldt. On P- and S-functions and related classes of n-dimensional nonlinear mappings. *Linear Algebra and its Applications*, 6:45–68, 1973.

[391] Sebastien Motsch and Eitan Tadmor. Heterophilious dynamics enhances consensus. *SIAM Review*, 56(4):577–621, 2014.

[392] Krikamol Muandet, Kenji Fukumizu, Bharath Sriperumbudur, and Bernhard Schölkopf. Kernel mean embedding of distributions: a review and beyond. *Foundations and Trends in Machine Learning*, 10(1-2):1–141, 2017.

[393] Meinard Müller. *Information Retrieval for Music and Motion*. Springer, Berlin, Heidelberg, 2007.

[394] Ryan Murray, Brian Swenson, and Soummya Kar. Revisiting normalized gradient descent: Fast evasion of saddle points. *IEEE Transactions on Automatic Control*, 64(11):4818–4824, 2019.

[395] Kimia Nadjahi, Alain Durmus, Lénaïc Chizat, Soheil Kolouri, and Gustavo K. Rohde. Statistical and topological properties of sliced probability divergences. In *Advances in Neural Information Processing Systems*, volume 33, pages 20802–20812, 2020.

[396] Kimia Nadjahi, Alain Durmus, Umut Simsekli, and Roland Badeau. Asymptotic guarantees for learning generative models with the sliced-Wasserstein distance. In *Advances in Neural Information Processing Systems*, 2019.

[397] Arkadi Nemirovski, Anatoli Juditsky, Guanghui Lan, and Alexander Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009.

[398] Arkadi Nemirovski and Uriel Rothblum. On complexity of matrix scaling. *Linear Algebra and its Applications*, 302:435–460, 1999.

[399] Yurii Nesterov and Arkadii Nemirovskii. *Interior-point polynomial algorithms in convex programming*, volume 13. SIAM, 1994.

[400] Yurii E. Nesterov. A method for solving the convex programming problem with convergence rate $O(1/k^2)$. *Doklady Akademii Nauk SSSR*, 269(3):543–547, 1983.

[401] Jonathan Niles-Weed and Quentin Berthet. Minimax estimation of smooth densities in Wasserstein distance. *The Annals of Statistics*, 50(3):1519–1540, 2022.

[402] Lipeng Ning and Tryphon T Georgiou. Metrics for matrix-valued measures via test functions. In *53rd IEEE Conference on Decision and Control*, pages 2642–2647. IEEE, 2014.

[403] Lipeng Ning, Tryphon T Georgiou, and Allen Tannenbaum. On matrix-valued Monge–Kantorovich optimal mass transport. *IEEE Transactions on Automatic Control*, 60(2):373–382, 2015.

[404] Jorge Nocedal and Stephen J Wright. *Numerical Optimization*. Springer-Verlag, 1999.

[405] Nikolas Nüsken and D. R. Michiel Renger. Stein variational gradient descent: Many-particle and long-time asymptotics. *Foundations of Data Science*, 5(3):286–320, 2023.

[406] Marcel Nutz and Johannes Wiesel. Entropic optimal transport: Convergence of potentials. *Probability Theory and Related Fields*, 184:401–424, 2022.

[407] Vladimir Oliker and Laird D Prussner. On the numerical solution of the equation $\frac{\partial^2 z}{\partial x^2}\frac{\partial^2 z}{\partial y^2} - \left(\frac{\partial^2 z}{\partial x \partial y}\right)^2 = f$ and its discretizations, I. *Numerische Mathematik*, 54(3):271–293, 1989.

[408] Dean S Oliver. Minimization for conditional simulation: Relationship to optimal transport. *Journal of Computational Physics*, 265:1–15, 2014.

[409] James B. Orlin. A polynomial time primal network simplex algorithm for minimum cost flows. *Mathematical Programming*, 78(2):109–129, 1997.

[410] Ferdinand Österreicher and Igor Vajda. A new class of metric divergences on probability spaces and its applicability in statistics. *Annals of the Institute of Statistical Mathematics*, 55(3):639–653, 2003.

[411] Felix Otto. The geometry of dissipative evolution equations: the porous medium equation. *Communications in Partial Differential Equations*, 26(1–2):101–174, 2001.

[412] Felix Otto and Cédric Villani. Generalization of an inequality by Talagrand and links with the logarithmic Sobolev inequality. *Journal of Functional Analysis*, 173(2):361–400, 2000.

[413] Andrea Ottolini and Stefan Steinerberger. Greedy matching in optimal transport with concave cost. *Annales Fennici Mathematici*, 50(2):549–562, 2025.

[414] Nicolas Papadakis, Gabriel Peyré, and Edouard Oudet. Optimal transport with proximal splitting. *SIAM Journal on Imaging Sciences*, 7(1):212–238, 2014.

[415] Sangmin Park and Dejan Slepčev. Geometry and analytic properties of the sliced Wasserstein space. *Journal of Functional Analysis*, 289(7):110975, 2025.

[416] Brendan Pass. Uniqueness and Monge solutions in the multimarginal optimal transportation problem. *SIAM Journal on Mathematical Analysis*, 43(6):2758–2775, 2011.

[417] Brendan Pass. On the local structure of optimal measures in the multi-marginal optimal transportation problem. *Calculus of Variations and Partial Differential Equations*, 43(3-4):529–536, 2012.

[418] Brendan Pass. Multi-marginal optimal transport: theory and applications. *ESAIM: Mathematical Modelling and Numerical Analysis*, 49(6):1771–1790, 2015.

[419] Franccois-Pierre Paty and Marco Cuturi. Subspace robust Wasserstein distances. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5072–5081, 2019.

[420] Franccois-Pierre Paty and Marco Cuturi. Regularized optimal transport is ground cost adversarial. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 7532–7542. PMLR, 2020.

[421] Jan Peszek and David Poyato. Heterogeneous gradient flows in the topology of fibered optimal transport. *Calculus of Variations and Partial Differential Equations*, 62(9):258, 2023.

[422] Peter Petersen. Gromov–Hausdorff convergence of metric spaces. In *Differential Geometry: Riemannian Geometry*, volume 54, Part 3 of *Proceedings of Symposia in Pure Mathematics*, pages 489–504. American Mathematical Society, 1993.

[423] Thomas Pethick, Wanyun Xie, Kimon Antonakopoulos, Zhenyu Zhu, Antonio Silveti-Falls, and Volkan Cevher. Training deep learning models with norm-constrained LMOs. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 49069–49104, 2025.

[424] Valentin V. Petrov. *Sums of Independent Random Variables*, volume 82 of *Ergebnisse der Mathematik und ihrer Grenzgebiete*. Springer-Verlag, Berlin, 1975.

[425] Gabriel Peyré. Entropic approximation of Wasserstein gradient flows. *SIAM Journal on Imaging Sciences*, 8(4):2323–2351, 2015.

[426] Gabriel Peyré. Optimal and diffusion transports in machine learning. *arXiv preprint arXiv:2512.06797*, 2025. Proc. 2026 International Congress of Mathematicians.

[427] Gabriel Peyré. Muon dynamics as a spectral Wasserstein flow. *arXiv preprint arXiv:2604.04891*, 2026.

[428] Gabriel Peyré. Robust sublinear convergence rates for iterative Bregman projections. *arXiv preprint arXiv:2602.01372*, 2026.

[429] Gabriel Peyré, Lenaic Chizat, Francois-Xavier Vialard, and Justin Solomon. Quantum entropic regularization of matrix-valued optimal transport. *European Journal of Applied Mathematics*, 30(6):1079–1102, 2019.

[430] Gabriel Peyré and Marco Cuturi. Computational optimal transport: With applications to data science. *Foundations and Trends in Machine Learning*, 11(5–6):355–607, 2019.

[431] Gabriel Peyré, Marco Cuturi, and Justin Solomon. Gromov-Wasserstein averaging of kernel and distance matrices. In *International Conference on Machine Learning*, pages 2664–2672, 2016.

[432] Gabriel Peyré, Clarice Poon, and Oscar Tron. Curvature of optimal transport with respect to the cost and applications to inverse optimal transport. *arXiv preprint arXiv:2604.22670*, 2026.

[433] Tao Pham Dinh and Hoai An Le Thi. Convex analysis approach to D.C. programming: Theory, algorithms and applications. *Acta Mathematica Vietnamica*, 22(1):289–355, 1997.

[434] Robert R. Phelps. *Lectures on Choquet's Theorem*, volume 1757 of *Lecture Notes in Mathematics*. Springer, Berlin, Heidelberg, 2 edition, 2001.

[435] Mark S. Pinsker. *Information and Information Stability of Random Variables and Processes*. Holden-Day, San Francisco, 1964. Translated and edited by Amiel Feinstein.

[436] Franccois Pitié, Anil C. Kokaram, and Rozenn Dahyot. N-dimensional probability density function transfer and its application to color transfer. In *IEEE International Conference on Computer Vision*, pages 1434–1439, 2005.

[437] Robert J. Plemmons. M-matrix characterizations. I—nonsingular M-matrices. *Linear Algebra and Its Applications*, 18(2):175–188, 1977.

[438] Boris T. Polyak. Gradient methods for the minimisation of functionals. *USSR Computational Mathematics and Mathematical Physics*, 3(4):864–878, 1963.

[439] Boris T. Polyak. Some methods of speeding up the convergence of iteration methods. *USSR Computational Mathematics and Mathematical Physics*, 4(5):1–17, 1964.

[440] Aram-Alexandre Pooladian and Jonathan Niles-Weed. Entropic estimation of optimal transport maps. *arXiv preprint arXiv:2109.12004*, 2021.

[441] Ning Qian. On the momentum term in gradient descent learning algorithms. *Neural Networks*, 12(1):145–151, 1999.

[442] Zhaonan Qu, Alfred Galichon, Wenzhi Gao, and Johan Ugander. On Sinkhorn's algorithm and choice modeling, 2023. Accepted for publication in Operations Research.

[443] Julien Rabin, Julie Delon, and Yann Gousseau. Transportation distances on the circle. *Journal of Mathematical Imaging and Vision*, 41(1–2):147–167, 2011.

[444] Julien Rabin, Gabriel Peyré, Julie Delon, and Marc Bernot. Wasserstein barycenter and its application to texture mixing. In *International Conference on Scale Space and Variational Methods in Computer Vision*, pages 435–446. Springer, 2011.

[445] Svetlozar T Rachev and Ludger Rüschendorf. *Mass Transportation Problems: Volume I: Theory*. Springer, 1998.

[446] Svetlozar T Rachev and Ludger Rüschendorf. *Mass Transportation Problems: Volume II: Applications*. Springer, 1998.

[447] Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems*, volume 20, 2007.

[448] Aaditya Ramdas, Nicolás García Trillos, and Marco Cuturi. On Wasserstein two-sample testing and related families of nonparametric tests. *Entropy*, 19(2):47, 2017.

[449] Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006.

[450] Erik Reinhard, Michael Adhikhmin, Bruce Gooch, and Peter Shirley. Color transfer between images. *IEEE Computer Graphics and Applications*, 21(5):34–41, 2001.

[451] Emmanuel Rio. Asymptotic constants for minimal distance in the central limit theorem. *Electronic Communications in Probability*, 16:96–103, 2011.

[452] Ralph Tyrell Rockafellar. *Convex Analysis*. Princeton university press, 2015.

[453] Aviv A. Rosenberg, Sanketh Vedula, Yaniv Romano, and Alexander M. Bronstein. Fast nonlinear vector quantile regression. In *International Conference on Learning Representations*, 2023.

[454] Murray Rosenblatt. Remarks on a multivariate transformation. *The Annals of Mathematical Statistics*, 23(3):470–472, 1952.

[455] Grant M. Rotskoff, Samy Jelassi, Joan Bruna, and Eric Vanden-Eijnden. Neuron birth-death dynamics accelerates gradient descent and converges asymptotically. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5508–5517. PMLR, 2019.

[456] Grant M. Rotskoff and Eric Vanden-Eijnden. Trainability and accuracy of artificial neural networks: An interacting particle system approach. *Communications on Pure and Applied Mathematics*, 75(9):1889–1935, 2022.

[457] Yossi Rubner, Carlo Tomasi, and Leonidas J Guibas. The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision*, 40(2):99–121, 2000.

[458] Alessandro Rudi and Lorenzo Rosasco. Generalization properties of learning with random features. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

[459] Walter Rudin. *Real and Complex Analysis*. McGraw–Hill, third edition, 1987.

[460] Ludger Ruschendorf. Convergence of the iterative proportional fitting procedure. *Annals of Statistics*, 23(4):1160–1174, 1995.

[461] Ludger Rüschendorf and Wolfgang Thomsen. Closedness of sum spaces and the generalized Schrödinger problem. *Theory of Probability and Its Applications*, 42(3):483–494, 1998.

[462] Ludger Rüschendorf and Ludger Uckelmann. On the n-coupling problem. *Journal of Multivariate Analysis*, 81(2):242–258, 2002.

[463] Jayoung Ryu, Charlotte Bunne, Luca Pinello, Aviv Regev, and Romain Lopez. Cross-modality matching and prediction of perturbation responses with labeled Gromov-Wasserstein optimal transport. *arXiv preprint arXiv:2405.00838*, 2024.

[464] Hiroaki Sakoe and Seibi Chiba. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 26(1):43–49, 1978.

[465] Antoine Salmona, Julie Delon, and Agnès Desolneux. Gromov–wasserstein distances between gaussian distributions. *Journal of Applied Probability*, 59(4):1178–1198, 2022.

[466] Hans Samelson. On the perron-frobenius theorem. *Michigan Mathematical Journal*, 4(1):57–59, 1957.

[467] Stefan G. Samko, Anatoly A. Kilbas, and Oleg I. Marichev. *Fractional Integrals and Derivatives: Theory and Applications*. Gordon and Breach, 1993.

[468] Michael E. Sander, Pierre Ablin, Mathieu Blondel, and Gabriel Peyré. Sinkformers: Transformers with doubly stochastic attention. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 3515–3530. PMLR, 2022.

[469] Filippo Santambrogio. *Optimal Transport for Applied Mathematicians: Calculus of Variations, PDEs, and Modeling*. Birkhäuser, 2015.

[470] Filippo Santambrogio. Dealing with moment measures via entropy and optimal transport. *Journal of Functional Analysis*, 271(2):418–436, 2016.

[471] Filippo Santambrogio. Crowd motion and population dynamics under density constraints. *GMT preprint 3728*, 2018.

[472] Giuseppe Savaré and G. E. Sodini. A simple relaxation approach to duality for optimal transport problems in completely regular spaces. *Journal of Convex Analysis*, 29(1):1–12, 2022.

[473] Meyer Scetbon and Marco Cuturi. Linear time Sinkhorn divergences using positive features. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

[474] Meyer Scetbon, Marco Cuturi, and Gabriel Peyré. Low-rank sinkhorn factorization. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 9344–9354. PMLR, 2021.

[475] Meyer Scetbon, Gabriel Peyré, and Marco Cuturi. Linear-time Gromov–Wasserstein distances using low-rank couplings and costs. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 19347–19365. PMLR, 2022.

[476] Geoffrey Schiebinger, Jian Shu, Marcin Tabaka, Brian Cleary, Vidya Subramanian, Joshua Gould, Aryeh Solomon, Siyan Liu, Stacie Lin, Peter Berube, Lia Lee, Jenny Chen, Justin Brumbaugh, Philippe Rigollet, Konrad Hochedlinger, Rudolf Jaenisch, Aviv Regev, and Eric S. Lander. Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming. *Cell*, 176(4):928–943.e22, 2019.

[477] Bernhard Schmitzer and Christoph Schnörr. Modelling convex shape priors and matching based on the Gromov-Wasserstein distance. *Journal of Mathematical Imaging and Vision*, 46(1):143–159, 2013.

[478] R. Schnabl. Zur approximation durch Bernsteinpolynome auf gewissen räumen von wahrscheinlichkeitsmaßen. *Mathematische Annalen*, 180:326–330, 1969.

[479] Isaac J Schoenberg. Metric spaces and positive definite functions. *Transactions of the American Mathematical Society*, 44(3):522–536, 1938.

[480] Bernhard Schölkopf and Alexander J Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, 2002.

[481] Erwin Schrödinger. Über die Umkehrung der Naturgesetze. *Sitzungsberichte Preuss. Akad. Wiss. Berlin. Phys. Math.*, 144:144–153, 1931.

[482] Othmane Sebbouh, Marco Cuturi, and Gabriel Peyré. Structured transforms across spaces with cost-regularized optimal transport. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pages 586–594. PMLR, 2024.

[483] Vivien Seguy and Marco Cuturi. Principal geodesic analysis for probability measures under the optimal transport metric. In *Advances in Neural Information Processing Systems 28*, pages 3294–3302, 2015.

[484] Thibault Séjourné, Franccois-Xavier Vialard, and Gabriel Peyré. The unbalanced gromov–wasserstein distance: Conic formulation and relaxation. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

[485] Soroosh Shafieezadeh Abadeh, Peyman Mohajerin Esfahani, and Daniel Kuhn. Distributionally robust logistic regression. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 1576–1584, 2015.

[486] Bin Shi, Simon S. Du, Michael I. Jordan, and Weijie J. Su. Understanding the acceleration phenomenon via high-resolution differential equations, 2018.

[487] Umut cSimcsekli, Levent Sagun, and Mert Gürbüzbalaban. A tail-index analysis of stochastic gradient noise in deep neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5827–5837. PMLR, 2019.

[488] Shashank Singh and Barnabás Póczos. Minimax distribution estimation in Wasserstein distance. *arXiv preprint arXiv:1802.08855*, 2018.

[489] Aman Sinha, Hongseok Namkoong, and John Duchi. Certifying some distributional robustness with principled adversarial training. In *International Conference on Learning Representations*, 2018.

[490] Richard Sinkhorn. A relationship between arbitrary positive matrices and doubly stochastic matrices. *The Annals of Mathematical Statistics*, 35(2):876–879, 1964.

[491] Richard Sinkhorn. Diagonal equivalence to matrices with prescribed row and column sums. *American Mathematical Monthly*, 74:402–405, 1967.

[492] Richard Sinkhorn and Paul Knopp. Concerning nonnegative matrices and doubly stochastic matrices. *Pacific Journal of Mathematics*, 21(2):343–348, 1967.

[493] Dejan Slepčev and Andrew Warren. Nonlocal wasserstein distance: Metric and asymptotic properties. *arXiv preprint arXiv:2209.08407*, 2022.

[494] Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 2256–2265. PMLR, 2015.

[495] Justin Solomon, Fernando De Goes, Gabriel Peyré, Marco Cuturi, Adrian Butscher, Andy Nguyen, Tao Du, and Leonidas Guibas. Convolutional Wasserstein distances: efficient optimal transportation on geometric domains. *ACM Transactions on Graphics*, 34(4):66:1–66:11, 2015.

[496] Max Sommerfeld and Axel Munk. Inference for empirical wasserstein distances on finite spaces. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(1):219–238, 2018.

[497] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

[498] Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021.

[499] Bharath K Sriperumbudur, Kenji Fukumizu, Arthur Gretton, Bernhard Schölkopf, and Gert RG Lanckriet. On integral probability metrics, φ-divergences and binary classification. *arXiv preprint arXiv:0901.2698*, 2009.

[500] Bharath K Sriperumbudur, Kenji Fukumizu, Arthur Gretton, Bernhard Schölkopf, and Gert RG Lanckriet. On the empirical estimation of integral probability metrics. *Electronic Journal of Statistics*, 6:1550–1599, 2012.

[501] Bharath K Sriperumbudur, Arthur Gretton, Kenji Fukumizu, Gert Lanckriet, and Bernhard Schölkopf. Injective Hilbert space embeddings of probability measures. In *Proceedings of the 21st Annual Conference on Learning Theory*, pages 111–122. Citeseer, 2008.

[502] Sanvesh Srivastava, Volkan Cevher, Quoc Dinh, and David Dunson. WASP: scalable bayes via barycenters of subset posteriors. In *Artificial Intelligence and Statistics*, pages 912–920, 2015.

[503] Sanvesh Srivastava, Cheng Li, and David B Dunson. Scalable bayes via barycenter in wasserstein space. *The Journal of Machine Learning Research*, 19(1):312–346, 2018.

[504] Matthew Staib, Sebastian Claici, Justin M Solomon, and Stefanie Jegelka. Parallel streaming wasserstein barycenters. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 2647–2658, 2017.

[505] Stefan Stanojevic, Yijun Li, and Lana X. Garmire. Computational methods for single-cell multi-omics integration and alignment. *arXiv preprint arXiv:2201.06725*, 2022.

[506] Volker Strassen. The existence of probability measures with given marginals. *The Annals of Mathematical Statistics*, 36(2):423–439, 1965.

[507] Karl-Theodor Sturm. On the geometry of metric measure spaces. I. *Acta Mathematica*, 196(1):65–131, 2006.

[508] Karl-Theodor Sturm. On the geometry of metric measure spaces. II. *Acta Mathematica*, 196(1):133–177, 2006.

[509] Karl-Theodor Sturm. The space of spaces: curvature bounds and gradient flows on the space of metric measure spaces. Preprint 1208.0434, arXiv, 2012.

[510] Weijie Su, Stephen Boyd, and Emmanuel J. Candès. A differential equation for modeling Nesterov's accelerated gradient method: Theory and insights. *Journal of Machine Learning Research*, 17(153):1–43, 2016.

[511] Mohamed M Sulman, JF Williams, and Robert D Russell. An efficient approach for the numerical solution of the monge–ampère equation. *Applied Numerical Mathematics*, 61(3):298–307, 2011.

[512] Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. On the importance of initialization and momentum in deep learning. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 1139–1147, Atlanta, Georgia, USA, 2013. PMLR.

[513] Gábor J Székely and Maria L Rizzo. Testing for equal distributions in high dimension. *InterStat*, 5(16.10), 2004.

[514] Michel Talagrand. Transportation cost for Gaussian and other product measures. *Geometric and Functional Analysis*, 6(3):587–600, 1996.

[515] Carla Tameling, Max Sommerfeld, and Axel Munk. Empirical optimal transport on countable metric spaces: Distributional limits and statistical applications. *The Annals of Applied Probability*, 29(5):2744–2781, 2019.

[516] Eloi Tanguy, Laetitia Chapel, and Julie Delon. Sliced transport plans. *arXiv preprint arXiv:2508.01243*, 2025.

[517] Alexis Thibault, Lénaïc Chizat, Charles Dossal, and Nicolas Papadakis. Overrelaxed Sinkhorn–Knopp algorithm for regularized optimal transport. *Algorithms*, 14(5):143, 2021.

[518] Matthew Thorpe, Serim Park, Soheil Kolouri, Gustavo K Rohde, and Dejan Slepčev. A transportation l^p distance for signal analysis. *Journal of Mathematical Imaging and Vision*, 59(2):187–210, 2017.

[519] John F. Toland. Duality in nonconvex optimization. *Journal of Mathematical Analysis and Applications*, 66(2):399–415, 1978.

[520] Alexander Tong, Jessie Huang, Guy Wolf, David van Dijk, and Smita Krishnaswamy. TrajectoryNet: A dynamic optimal transport network for modeling cellular dynamics. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 9526–9536. PMLR, 2020.

[521] Huy V. Tran, Hicham Janati, Nicolas Courty, Rémi Flamary, Ievgen Redko, Pinar Demetci, and Ritambhara Singh. Unbalanced co-optimal transport. *arXiv preprint arXiv:2205.14923*, 2022.

[522] Alain Trouvé and Laurent Younes. Metamorphoses through Lie group action. *Foundations of Computational Mathematics*, 5(2):173–198, 2005.

[523] Neil S Trudinger and Xu-Jia Wang. On the monge mass transfer problem. *Calculus of Variations and Partial Differential Equations*, 13(1):19–31, 2001.

[524] Paul Tseng. Convergence of a block coordinate descent method for nondifferentiable minimization. *Journal of Optimization Theory and Applications*, 109(3):475–494, 2001.

[525] Ananya Uppal, Shashank Singh, and Barnabás Póczos. Nonparametric density estimation and convergence rates for GANs under Besov IPM losses. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

[526] Théo Uscidda and Marco Cuturi. The Monge gap: A regularizer to learn all transport maps. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 34709–34733. PMLR, 2023.

[527] Adrien Vacher, Boris Muzellec, Alessandro Rudi, Francis Bach, and Franccois-Xavier Vialard. A dimension-free computational upper-bound for smooth optimal transport estimation. *arXiv preprint arXiv:2101.05380*, 2021.

[528] S. S. Vallender. Calculation of the wasserstein distance between probability distributions on the line. *Theory of Probability and Its Applications*, 18(4):784–786, 1974.

[529] Aad W. van der Vaart and Jon A. Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer Series in Statistics. Springer, New York, 1996.

[530] Sathamangalam R Srinivasa Varadhan. On the behavior of the fundamental solution of the heat equation with variable coefficients. *Communications on Pure and Applied Mathematics*, 20(2):431–455, 1967.

[531] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

[532] Christophe Vauthier, Quentin Mérigot, and Anna Korba. Highly data parallelizable estimation of the Sliced-Wasserstein distance using cumulative distribution functions. *arXiv preprint arXiv:2606.30310*, 2026.

[533] Titouan Vayer. *A Contribution to Optimal Transport on Incomparable Spaces*. PhD thesis, Université Bretagne Sud, Vannes, France, 2020.

[534] Titouan Vayer. On sparsity, extremal structure, and monotonicity properties of wasserstein and gromov–wasserstein optimal transport plans. *arXiv preprint arXiv:2602.16265*, 2026.

[535] Titouan Vayer, Nicolas Courty, Romain Tavenard, Laetitia Chapel, and Rémi Flamary. Optimal transport for structured data with application on graphs. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 6275–6284. PMLR, 2019.

[536] Nina Vesseron, Louis Béthune, and Marco Cuturi. Sample and map from a single convex potential: Generation using conjugate moment measures. *arXiv preprint arXiv:2503.10576*, 2025.

[537] Cédric Villani. A review of mathematical topics in collisional kinetic theory. In *Handbook of Mathematical Fluid Dynamics*, volume 1, pages 71–305. North-Holland, 2002.

[538] Cédric Villani. *Topics in Optimal Transportation*, volume 58 of *Graduate Studies in Mathematics Series*. American Mathematical Society, 2003.

[539] Cédric Villani. *Optimal Transport: Old and New*, volume 338. Springer, 2009.

[540] Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural Computation*, 23(7):1661–1674, 2011.

[541] Taras K. Vintsyuk. Speech discrimination by dynamic programming. *Cybernetics*, 4(1):52–57, 1968.

[542] John von Neumann. A certain zero-sum two-person game equivalent to the optimal assignment problem. In Harold W. Kuhn and Albert W. Tucker, editors, *Contributions to the Theory of Games II*, volume 28 of *Annals of Mathematics Studies*, pages 5–12. Princeton University Press, Princeton, NJ, 1953.

[543] Max-K. von Renesse and Karl-Theodor Sturm. Transport inequalities, gradient estimates, entropy and Ricci curvature. *Communications on Pure and Applied Mathematics*, 58(7):923–940, 2005.

[544] James Vuckovic, Aristide Baratin, and Rémi Tachet des Combes. A mathematical theory of attention. *arXiv preprint arXiv:2007.02876*, 2020.

[545] Sinong Wang, Belinda Z. Li, Madian Khabsa, Han Fang, and Hao Ma. Linformer: Self-attention with linear complexity. *arXiv preprint arXiv:2006.04768*, 2020.

[546] Wei Wang, Dejan Slepčev, Saurav Basu, John A Ozolek, and Gustavo K Rohde. A linear optimal transportation framework for quantifying and visualizing variations in sets of images. *International Journal of Computer Vision*, 101(2):254–269, 2013.

[547] Andrew Warren. Gradient flow structure for some nonlocal diffusion equations. *arXiv preprint arXiv:2412.20969*, 2024.

[548] Jonathan Weed and Francis Bach. Sharp asymptotic and finite-sample rates of convergence of empirical measures in Wasserstein distance. *Bernoulli*, 25(4A):2620–2648, 2019.

[549] Holger Wendland. *Scattered Data Approximation*. Cambridge University Press, 2005.

[550] Andre Wibisono, Ashia C. Wilson, and Michael I. Jordan. A variational perspective on accelerated methods in optimization. *Proceedings of the National Academy of Sciences*, 113(47):E7351–E7358, 2016.

[551] Diyuan Wu, Vyacheslav Kungurtsev, and Marco Mondelli. Mean-field analysis for heavy ball methods: Dropout-stability, connectivity, and global convergence, 2022.

[552] Gui-Song Xia, Sira Ferradans, Gabriel Peyré, and Jean-Franccois Aujol. Synthesizing and mixing stationary Gaussian texture models. *SIAM Journal on Imaging Sciences*, 7(1):476–508, 2014.

[553] Huang Xiao, Michael Herman, Joerg Wagner, Sebastian Ziesche, Jalal Etesami, and Thai Hong Linh. Wasserstein adversarial imitation learning. *arXiv preprint arXiv:1906.08113*, 2019.

[554] Yunyang Xiong, Zhanpeng Zeng, Rudrasis Chakraborty, Mingxing Tan, Glenn Fung, Yin Li, and Vikas Singh. Nyströmformer: A Nyström-based algorithm for approximating self-attention. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14138–14148, 2021.

[555] Huan Xu and Shie Mannor. Distributionally robust markov decision processes. *Mathematics of Operations Research*, 37(2):288–300, 2012.

[556] Insoon Yang. A convex optimization approach to distributionally robust markov decision processes with Wasserstein distance. *IEEE Control Systems Letters*, 1(1):164–169, 2017.

[557] Jiaping Yang and Yunxin Zhang. Closed forms for Gaussian Kullback–Leibler unbalanced optimal transport without coupling entropy, 2026.

[558] G Udny Yule. On the methods of measuring association between two attributes. *Journal of the Royal Statistical Society*, 75(6):579–652, 1912.

[559] Yoav Zemel and Victor M Panaretos. Fréchet means and Procrustes analysis in Wasserstein space. *Bernoulli*, 25(2):932–976, 2019.

APPENDIX A

Notation Table

This appendix collects the main notation used throughout the book. The last column points to the first section, equation, definition, proposition or theorem where the notation is defined or first used in a mathematically meaningful way. The global convention is that α, β denote the main source and target measures. A single generic measure is also denoted by α; auxiliary measures use contextual letters such as ν, γ or η.

Notation	Meaning	First reference
Ambient spaces, measures and elementary objects		
$\mathbb{R}^d$	Euclidean ambient space.	Section 2.1
$\mathcal{X}, \mathcal{Y}$	Source and target spaces.	Eq. (2.5)
$\mathcal{M}(\mathcal{X})$	Finite signed Radon measures on $\mathcal{X}$.	Section 2.1
$\mathcal{M}_+(\mathcal{X}), \mathcal{M}_+^1(\mathcal{X})$	Positive finite measures and probability measures.	Section 2.1
$\mathcal{P}(\mathcal{X}), \mathcal{P}_p(\mathcal{X})$	Probability measures, with finite p-moment for $\mathcal{P}_p$.	Section 3.3
Σ_n	Probability simplex of histograms of length n.	Definition 2.1
δ_x	Dirac mass at x.	Definition 2.2
α, β, γ	Source, target and auxiliary probability measures.	Eq. (2.5)
ρ_α, ρ_β	Densities of α, β with respect to a reference measure; write ρ when the measure is unambiguous.	Definition 2.6
$\mathrm{d}\alpha, \mathrm{d}x$	Integration against α and against Lebesgue measure.	Section 2.1
$\mathbb{E}$	Expectation of a random variable.	Section 2.1
$\mathrm{supp}(\pi)$	Topological support of a measure.	Definition 2.4
$\mathrm{Supp}(\mathrm{b})$	Index support of a histogram.	Eq. (7.8)
$\mathcal{C}(\mathcal{X})$	Continuous real-valued functions on $\mathcal{X}$.	Section 2.1
$\|\cdot\|$	Euclidean norm or the norm indicated by a subscript.	Chapter 1
$\langle \cdot, \cdot \rangle$	Euclidean/Frobenius pairing or measure-function pairing.	Section 2.1
Discrete matching and discrete Kantorovich OT		
$(x_i)_i, (y_j)_j$	Source and target point clouds.	Eq. (1.1)
$\mathrm{C} = (\mathrm{C}_{i,j})$	Cost matrix between source and target points.	Eq. (1.1)
$\sigma \in \mathrm{Perm}(n)$	Permutation encoding a one-to-one matching.	Eq. (1.1)
$\mathrm{P}_\sigma, \mathcal{P}_n^{\mathrm{perm}}$	Permutation matrix and the set of all such matrices.	Definition 3.11
$\mathcal{B}_n$	Birkhoff polytope of bistochastic matrices.	Definition 3.12
a, b	Discrete probability histograms.	Eq. (3.1)
P	Discrete transport/coupling matrix.	Eq. (3.1)
$\mathrm{U}(\mathrm{a}, \mathrm{b})$	Polytope of discrete couplings with marginals a, b.	Eq. (3.1)
$\mathbb{1}_n, \mathrm{P}^\top$	All-ones vector and transpose of P.	Eq. (3.1)
$\mathrm{L}_\mathrm{C}(\mathrm{a}, \mathrm{b})$	Discrete Kantorovich optimal value with cost C.	Eq. (3.2)
D	Ground distance matrix for discrete Wasserstein distances.	Definition 4.2
$\mathrm{W}_p(\mathrm{a}, \mathrm{b})$	Discrete p-Wasserstein distance.	Definition 4.2
Monge maps, one-dimensional OT and Gaussians		
T, T	Transport map.	Eq. (2.5)
$T_\sharp \alpha$	Push-forward of α by T.	Definition 2.10
$T^\sharp g$	Pullback of a test function, $T^\sharp g = g \circ T$.	Remark 2.11
Id	Identity map.	Definition 2.10

Notation	Meaning	First reference
$\tilde{\mathcal{W}}_p$	Directed Monge transport distance.	Eq. (2.6)
$\nabla\varphi$	Brenier map for quadratic cost.	Theorem 2.23
C_α	Cumulative distribution function of a 1-D measure.	Eq. (2.8)
C_α^{-1}	Quantile function of a 1-D measure.	Eq. (2.9)
s_+	Positive part $\max(s,0)$ in the Bobkov–Ledoux cumulative formula.	Proposition 2.45
$\mathsf{T}=(V,E,\ell)$	Weighted tree with vertices, edges and edge lengths.	Proposition 2.48
d_T, V_e, F_e	Tree geodesic distance, edge subtree and subtree mass imbalance.	Eq. (2.14)
$\mathcal{N}(m,\Sigma)$	Gaussian law with mean m and covariance Σ.	Eq. (2.15)
m_α, Σ_α	Mean and covariance of a Gaussian measure α.	Eq. (2.19)
$\mathcal{B}(\Sigma_\alpha,\Sigma_\beta)$	Bures covariance distance.	Definition 2.53
$\Phi_2(\alpha)$	Raw second-moment matrix of a probability measure.	Proposition 2.57
d_{FR}	Fisher–Rao distance on Gaussian covariance or scale parameters.	Section 2.7
$\mathrm{tr}(\Sigma)$	Trace of a matrix.	Eq. (2.19)
Continuous Kantorovich OT and Wasserstein distances		
π	Coupling or transport plan.	Definition 3.26
$\mathcal{U}(\alpha,\beta)$	Set of couplings between α and β.	Eq. (3.6)
$\mathcal{L}_c(\alpha,\beta)$	Kantorovich value with ground cost c, extended homogeneously to finite equal-mass measures.	Eqs. (3.7) and (11.2)
$\mathcal{M}_\alpha^c(T)$	Monge optimality gap of a map relative to its source law.	Definition 3.40
d	Ground distance on the underlying metric space.	Eq. (4.3)
$\mathcal{W}_p(\alpha,\beta)$	p-Wasserstein distance.	Definition 4.6
$\mathcal{W}_\infty(\alpha,\beta)$	Worst-displacement Wasserstein distance.	Eq. (4.5)
$\mathfrak{A},\mathfrak{B}$	Probability laws over probability measures.	Eq. (13.5)
$\widehat{\mathfrak{A}}_p$	Empirical law of p random probability measures.	Proposition 12.13
$\bar{\alpha}_\mathfrak{A}$	Collapsed mixture associated with a law over measures.	Definition 13.9
$\widetilde{\alpha}_\mathfrak{A}$	Wasserstein-barycenter flattening of a law over measures.	Section 12.1
$\mathbb{W}_p$	p-Wasserstein distance on the Wasserstein space.	Eq. (13.7)
Γ	c-cyclically monotone subset of $\mathcal{X}\times\mathcal{Y}$.	Definition 3.45
ρ	Glued or composed coupling.	Lemma 4.4
$\rightharpoonup$	Weak* convergence of measures.	Definition 4.18
$\mathrm{TV}, \|\cdot\|_{\mathrm{TV}}$	Total variation divergence/norm.	Section 2.1
Duality, transforms and weak norms		
f,g	Discrete dual potentials.	Eq. (5.2)
f,g	Continuous dual potentials.	Eq. (5.6)
$\mathcal{D}_0(f,g), \mathcal{D}_\varepsilon(f,g)$	Hard and entropic full dual objectives, for functions or vectors.	Eqs. (6.1) and (8.23)
$\mathcal{E}_0(g), \mathcal{E}_\varepsilon(g)$	Hard and entropic semi-dual objectives, for functions or vectors.	Eqs. (6.2) and (9.20)
$\mathrm{R}(\mathrm{C})$	Feasible set of discrete dual potentials for cost C.	Eq. (5.1)
$\mathcal{R}(c)$	Feasible set of continuous dual potentials.	Eq. (5.5)
$f^c, g^{\bar{c}}$	c- and $\bar{c}$-transforms of dual potentials.	Definition 5.8
$\mathbb{L}_j(\mathrm{g})$	Laguerre/power cell in semi-discrete OT.	Eq. (6.14)
$\mathbb{V}_j(Y)$	Voronoi cell of codepoint y_j.	Proposition 6.12
$\mathcal{Q}_m(\alpha)$	Optimal m-point quantization error.	Eq. (6.19)
$\mathcal{F}(Y), \mathcal{F}_{\mathrm{eq}}(Y)$	Free-mass and equal-weight quantization energies.	Section 6.4
$\mathrm{Lip}(f)$	Lipschitz constant of f.	Eq. (6.20)

Notation	Meaning	First reference
$\mathcal{W}_1$	Kantorovich–Rubinstein distance/norm.	Eq. (6.21)
$s, \|s\|_{\mathrm{TV}}$	Vector-valued Beckmann flux and its total variation norm.	Definition 6.17
$\mathcal{W}_{1,G}$	Graph Wasserstein-1/transshipment distance.	Proposition 6.22
$d_G, \nabla_G, \mathrm{div}_G$	Graph geodesic distance, gradient and divergence.	Proposition 6.22
$\|\cdot\|_B$	Extended dual seminorm induced by a discriminator class B.	Eq. (7.1)
$\mathcal{H}, k$	Reproducing kernel Hilbert space and its kernel.	Definition 7.9
MMD_k	Maximum mean discrepancy/kernel seminorm for k.	Definition 7.9
$\mathcal{D}_\varphi, \mathrm{D}_\varphi$	Continuous and discrete φ-divergences.	Eq. (7.6)
φ'_∞	Recession slope of an entropy function.	Definition 7.15
$\varphi^\star$	Legendre transform of φ.	Eq. (7.13)
KL, KL	Continuous and discrete Kullback–Leibler divergences.	Definitions 8.7, 8.11
$\mathfrak{h}$	Hellinger distance.	Section 7.3
JS	Jensen–Shannon distance.	Section 7.3
Entropic regularization and Sinkhorn algorithms		
ε	Entropic regularization strength.	Eq. (8.1)
H(P)	Shannon–Boltzmann entropy of a matrix.	Definition 8.1
$\mathrm{L}_\mathrm{C}^\varepsilon(\mathrm{a}, \mathrm{b})$	Discrete entropic OT value.	Eq. (8.1)
$\mathcal{L}_c^\varepsilon(\alpha, \beta)$	Continuous entropic OT value.	Eq. (8.13)
$\mathcal{I}(X, Y)$	Mutual information of a coupled pair.	Definition 8.13
$\mathcal{I}_{\mathrm{geo}}(\alpha, \beta)$	Integrated Fisher information along the quadratic Wasserstein geodesic.	Proposition 8.17
K	Gibbs kernel $e^{-\mathrm{C}/\varepsilon}$.	Eq. (8.2)
u, v	Left and right Sinkhorn scalings.	Eq. (8.2)
diag(u)K diag(v)	Scaling form of the entropic coupling.	Eq. (9.1)
$\odot$	Entrywise product of vectors.	Eq. (8.4)
$\mathrm{u}^{(\ell)}, \mathrm{u}^{(\ell+1)}$	Current and next Sinkhorn iterates.	Eq. (8.5)
$f^{c,\varepsilon}, g^{\bar{c},\varepsilon}$	Continuous soft c-transforms.	Definition 8.21
$\mathrm{SB}_\varepsilon(\alpha, \beta)$	Dynamic Schrödinger bridge value.	Eq. (8.30)
$\mathrm{prox}_\mathfrak{h}^{\mathrm{KL}}$	KL-proximal map of a marginal penalty.	Eq. (8.39)
$\|\cdot\|_V$	Variation seminorm on potentials modulo constants.	Definition 9.6
$d_\mathcal{H}$	Hilbert projective metric on positive vectors.	Definition 9.14
$\Delta^{(\ell)}$	Entropic dual suboptimality after ℓ Sinkhorn cycles.	Proposition 9.12
$\eta(\mathrm{K}), \lambda(\mathrm{K})$	Projective cross-ratio and Birkhoff contraction factor.	Theorem 9.16
$T_\pi, T_\pi^*, T_\varepsilon, \sigma_\varepsilon$	Conditional coupling operators and maximal correlation at the entropic optimum.	Definition 9.18
$\mathrm{Proj}^{\mathrm{KL}}$	KL/Bregman projection.	Eq. (9.1)
$\bar{\mathcal{L}}_c^\varepsilon(\alpha, \beta)$	Debiased Sinkhorn divergence.	Eq. (8.57)
$\mathcal{V}_{c,\varepsilon}, \mathcal{V}_{\mathrm{C},\varepsilon}$	KL-normalized continuous and finite entropic OT values.	Propositions 12.29, 10.17
$z = (u, v), Q(z)$	Log-scaling variables and clearing map in the M-function view.	Definition 9.30
$\underline{z}, \overline{z}, z^\star$	Lower/upper order barriers and the monotone-clearing fixed point.	Theorem 9.31
$\eta_{i,j}, \sigma_i, \tau_j$	Loss factors and outside-option coefficients in lossy Sinkhorn clearing.	Example 9.33
$u_t, F, G, \bar{r}_t$	Continuous ε-Sinkhorn potential, log-densities and gauge term.	Definition 9.27
$v_k, S_k, \rho_{k,u}$	Scaled log-Sinkhorn transforms and multiplicative increment.	Eq. (9.25)

Notation	Meaning	First reference
B_t, q_t	Gaussian linear part and mean shift in continuous Sinkhorn closure.	Section 9.7
Statistical OT, empirical maps and kernel sketching		
X_i, Y_j	I.i.d. source and target samples.	Eq. (10.1)
$\hat{\alpha}_n, \hat{\beta}_m$	Empirical source and target laws.	Eq. (10.1)
$\mathcal{Q}_j$	Dyadic partition at scale 2^{-j}.	Proposition 10.6
$B_{n,m}(\mathcal{V}), Z_{n,m}(\mathcal{V})$	Bias and centered fluctuation of an empirical OT value.	Section 10.3
$\mathcal{D}^\star(a, b)$	Optimal dual face for a finite exact OT problem.	Proposition 10.16
$\mathrm{P}^\varepsilon_{n,m}$	Empirical entropic coupling between sampled measures.	Eq. (10.5)
$T^\varepsilon_{n,m}, \omega^\varepsilon_j$	Soft barycentric map estimator and its barycentric weights.	Eq. (10.5)
$k, \varphi(x, z), \widetilde{k}_r$	Kernel, feature map and Monte-Carlo kernel sketch.	Section 10.4
Φ, Φ_X, Φ_Y	Sampled feature matrices for symmetric and rectangular kernel sketches.	Section 10.4
$\widetilde{K}, \widetilde{c}_{i,j}$	Sketched Sinkhorn kernel and effective sketched cost.	Proposition 10.18
$r, R = rp$	Number of sampled latent variables and flattened sketch rank.	Section 10.4
$\mathrm{DNN}_n, \mathrm{CP}_n$	Doubly nonnegative and completely positive matrix cones.	Section 10.4
$k_{p,\varepsilon}, \varphi_p, \nu_{a,\varepsilon}$	Generalized Gaussian kernel, positive feature and stable mixing law.	Proposition 10.21
Extensions of OT		
ψ_1, ψ_2	Entropy functions penalizing marginal mismatch.	Eq. (11.1)
$\mathrm{UW}_c, \mathrm{UW}_{c,\tau}$	Relaxed unbalanced OT value with marginal penalties.	Eq. (11.1)
$\mathcal{E}_{c,\tau_0,\tau_1}$	Mass-independent normalized shape envelope for KL relaxation.	Eq. (11.4)
GHK_τ	Gaussian Hellinger–Kantorovich endpoint value for quadratic KL-relaxed transport.	Eq. (11.13)
C_τ, Ψ	Gaussian covariance envelope and soft row-Stiefel penalty.	Eqs. (11.17) and (11.24)
P_τ, Q_τ, L_τ	Adjusted Gaussian covariances and their optimal linear transport map.	Eqs. (11.20)–(11.21)
$\mathrm{UW}^{\mathrm{KL}}_{\tau,\varepsilon}, M_{\varepsilon,\tau}$	Entropic KL–UOT value and transported mass for unnormalized Gaussian inputs.	Eqs. (11.31)–(11.32)
POT_m	Partial OT value transporting exactly mass m.	Definition 11.19
$\mathrm{POT}^{\mathrm{TV}}_\lambda$	TV-penalized partial OT value with mass price λ.	Definition 11.19
L_c	Reverse-formulation local unbalanced cost.	Eq. (11.8)
H_c	Homogeneous perspective of the local cost L_c.	Eq. (11.9)
HW	Homogeneous unbalanced formulation.	Eq. (11.10)
$\mathfrak{C}[\mathcal{X}]$	Cone over the metric space $\mathcal{X}$.	Section 11.1
CW	Cone formulation of unbalanced OT.	Theorem 11.10
$A_\kappa, J_\kappa, \mathbb{J}_\kappa$	Unbalanced velocity action, flux/source perspective, and measure action for transport and growth.	Eqs. (14.38)–(14.40)
WFR_κ	Wasserstein–Fisher–Rao dynamic distance with growth scale κ.	Eq. (14.41)
β_s, λ_s	Input measures and weights in barycenter problems.	Eq. (12.1)
$\alpha^\star$	Optimal measure, often a barycenter.	Eq. (12.1)
$\mathcal{B}_c(\mathfrak{N})$	Set of barycenters of a law over measures.	Section 12.1
B, c_{bar}	Barycentric map and induced multi-marginal barycenter cost.	Proposition 12.20

Notation	Meaning	First reference
SW_p	Sliced Wasserstein distance.	Definition 11.22
$\mathbb{S}^{d-1}$	Unit sphere of projection directions.	Definition 11.22
P_θ	Projection on direction θ.	Definition 11.22
$\mathfrak{R}\alpha$	Measure-valued Radon transform of α.	Definition 11.23
$R\rho$	Density Radon transform of ρ.	Remark 11.24
$R^\dagger h$	Least-squares Radon pseudoinverse density reconstructed from a sinogram h.	Proposition 12.14
γ_θ	One-dimensional projected/Radon-domain barycenter law.	Section 12.1
$\mathrm{SW}_{p,q,k}$	L^q aggregate of p-Wasserstein distances over k-dimensional projections.	Definition 11.30
$\mathrm{SW}_{p,q}, \mathrm{MaxSW}_{p,k}$	Line-sliced and max-sliced abbreviations, with $\mathrm{MaxSW}_{p,k} = \mathrm{SW}_{p,\infty,k}$.	Definition 11.30
Min-SW_2	Min-SW lifted-plan discrepancy, upper-bounding $\mathcal{W}_2$.	Definition 11.32
$\mathcal{W}_{p,\mathcal{G}}$	Quotient Wasserstein distance modulo a group action.	Definition 11.34
$\mathrm{E}(d), \mathrm{O}(d), \mathrm{SO}(d)$	Euclidean, orthogonal and orientation-preserving rigid-motion groups.	Section 11.3
R, t, M_{P}	Rigid motion and weighted cross-covariance in Wasserstein–Procrustes.	Proposition 11.37
$\mathcal{W}_\gamma$	Spectral Wasserstein distance associated with a matrix gauge γ.	Eq. (11.52)
$\mathcal{B}_\gamma$	Polar set defining the robust projected form of $\mathcal{W}_\gamma$.	Eq. (11.54)
$\mathcal{W}_{2,A}$	Quadratic Wasserstein pseudodistance after projection by $A^{1/2}$.	Eq. (11.53)
$\mathrm{SRW}_{2,k}$	Paty–Cuturi subspace robust Wasserstein distance.	Section 11.5
$\mathrm{LOT}_{\alpha_{\mathrm{ref}}}$	Linear OT distance around the reference measure α_{ref}.	Definition 11.38
MMOT_c	Multimarginal optimal-transport value.	Eq. (12.29)
$\mathrm{Q}, \mathrm{R}, g$	Low-rank OT factors and latent mass vector.	Definition 12.24
$\mathrm{P}(\mathrm{Q}, \mathrm{R}, g)$	Coupling induced by a low-rank factored representation.	Eq. (12.36)
$\eta = \sum_k g_k \delta_{z_k}$	Abstract intermediate measure in low-rank OT.	Definition 12.24
$\mathcal{L}_c^u, u$	Capacity-constrained OT value and capacity density.	Definition 12.26
$\overline{\mathrm{P}}$	Discrete upper-capacity matrix for a capped coupling.	Eq. (12.39)
$\mathcal{V}_c$	Unregularized OT value used in sensitivity formulas.	Proposition 12.28
$\mathcal{L}_{\mathrm{iOT}}$	Inverse-OT primal–dual gap loss.	Section 12.5
$\bar{T}_\pi$	Barycentric projection of a coupling π.	Eq. (12.43)
$\bar{\beta}_\pi$	Push-forward of α by the barycentric projection.	Eq. (12.43)
WOT_C	Weak OT value with conditional-law cost C.	Eq. (12.44)
g^C	Weak C-transform in weak OT duality.	Proposition 12.35
$\mathcal{U}_{\mathrm{mart}}(\alpha, \beta)$	Martingale couplings between α and β.	Definition 12.37
$\preceq_{\mathrm{st}}, \preceq_{\mathrm{cx}}$	Stochastic order and convex order.	Section 12.6.1
C_{bar}	Quadratic barycentric weak-transport cost.	Proposition 12.36
u_t, V_t	Positive vector-valued density and spatial flux.	Eqs. (13.1), (13.2)
$\mathcal{W}_\Phi$	Dynamic vector-valued BB-type cost.	Eq. (13.1)
D, D'	Intra-domain distance matrices in discrete GW.	Eq. (13.8)
Δ	Discrepancy between intra-domain distances.	Eq. (13.8)
$\mathcal{E}_{\mathrm{D},\mathrm{D}'}(\mathrm{P})$	Discrete GW distortion energy.	Eq. (13.8)
GW	Discrete Gromov–Wasserstein cost.	Eq. (13.8)
$\mathbb{X}, \mathbb{Y}$	Metric-measure spaces.	Definition 13.14

Notation	Meaning	First reference
$\mathcal{GW}$	Continuous Gromov–Wasserstein distance.	Eq. (13.9)
$\mathcal{GW}_p^{\mathrm{M}}, \mathcal{GW}_\infty^{\mathrm{M}}$	Directed Gromov–Monge distances.	Eqs. (13.35), (13.36)
$\mathcal{GW}_\infty$	Infinite-order Gromov–Wasserstein distance.	Eq. (13.10)
$\mathfrak{D}_{\mathbb{X}}$	Law of local distance profiles of a metric-measure space.	Proposition 13.25
C(P)	Half-gradient of the squared discrete GW distortion.	Eq. (13.51)
$d_{\mathrm{H}}, d_{\mathrm{GH}}$	Hausdorff and Gromov–Hausdorff distances.	Section 13.3
$\mathrm{FGW}_{\lambda,p}$	Fused Gromov–Wasserstein distance.	Section 13.3
$\mathbb{S}^m, \mathbb{S}_+^m$	Real symmetric matrices and their positive semidefinite cone.	Definition 13.5
A_t, P_t	Positive matrix-valued density and spatial matrix flux.	Eqs. (13.3), (13.4)
$\mathcal{W}_{\mathrm{mat}}$	Conservative matrix-valued BB-type cost.	Eq. (13.3)
$\mathbb{H}_n, \mathbb{H}_n^+, \mathbb{H}_n^{+,1}$	Hermitian matrices, positive semidefinite Hermitian matrices and density matrices.	Definition 13.47
$\mathrm{Tr}_A, \mathrm{Tr}_B$	Partial traces of a bipartite matrix.	Eq. (13.57)
$\mathrm{QOT}_C(A, B)$	Finite-dimensional quantum OT value with cost observable C.	Eq. (13.58)
$\mathrm{QOT}_C^\varepsilon(A, B)$	Entropically regularized quantum OT value.	Eq. (13.60)
$D_H(T\|K)$	Quantum relative entropy used for Bregman projections.	Proposition 13.53
$T_e(F, G), T_s(F, G)$	Exact Gibbs coupling and symmetric Gurvits-scaling surrogate.	Eqs. (13.62), (13.63)
$\Omega_{n,m}, A_\omega$	Monotone warping paths and their incidence matrices.	Definition 13.56
$\mathrm{DTW}_c(x, y)$	Dynamic time-warping value between two feature sequences.	Eq. (13.65)
$\Gamma_\uparrow, \mathrm{CDTW}_c(x, y)$	Monotone clock pairs and continuous time-warping value.	Eq. (13.67)
$\mathrm{sDTW}_{c,\varepsilon}, \mathbb{P}_\varepsilon, E_\varepsilon$	Soft-DTW free energy, Gibbs path law and expected alignment matrix.	Eqs. (13.71)–(13.74)
Dynamic OT and Wasserstein gradient flows		
α_t	Time-dependent curve of probability measures.	Eq. (14.2)
ρ_t	Density of α_t with respect to the relevant reference measure.	Eq. (14.12)
v_t	Eulerian velocity field transporting α_t.	Eq. (14.2)
$\mathcal{H}_\alpha$	$L^2(\alpha)$ vector fields with zero weighted divergence.	Section 14
$\omega_t = \alpha_t v_t$	Vector-valued momentum/flux measure in convex dynamic formulations.	Eq. (14.12)
$m_t = \rho_t v_t$	Density of ω_t when $\alpha_t = \rho_t \, \mathrm{d}x$.	Eq. (14.12)
$\sigma_t = s_t \, \mathrm{d}x$	Source measure in dynamic unbalanced transport.	Eq. (14.41)
$J(a, m)$	Quadratic perspective $\|m\|^2/a$ with lower-semicontinuous extension at $a = 0$.	Eq. (14.10)
$\mathbb{J}(\alpha, \omega)$	Measure-level perspective action induced by J.	Eq. (14.11)
T_t	Lagrangian particle flow map.	Eq. (14.1)
P_t	Interpolant map in flow matching.	Eq. (16.1)
$\mathcal{W}_2^2$ via action	Benamou–Brenier dynamic formulation.	Eq. (14.8)
$\|\dot{x}_t\|$	Metric derivative of an absolutely continuous curve.	Section 15
$\mathbb{A}(\alpha, w)$	Tangent action defining a dynamic length distance.	Eq. (14.15)
$\mathsf{D}_{\mathbb{A}}$	Length-space distance generated by the tangent action $\mathbb{A}$.	Eq. (14.15)
$\mathrm{PMO}_{\mathbb{A},\alpha}$	Penalized minimization oracle associated with $w \mapsto \mathbb{A}(\alpha, w)$.	Eq. (15.33)

Notation	Meaning	First reference
Q_α	Positive operator representing a quadratic local tangent action.	Eq. (14.16)
D_Q	Geodesic distance induced by a quadratic tangent operator Q_α.	Eq. (14.16)
$A(a,w)$	Local velocity action density, with scalar density value a and velocity $w \in \mathbb{R}^d$.	Eq. (14.17)
$J_A(a,m)$	Momentum perspective of A, with pointwise momentum $m \in \mathbb{R}^d$.	Eq. (14.19)
$\mathbb{J}_{A,\lambda}(\alpha,\omega)$	Measure-level momentum action associated with A relative to a reference measure λ; written $\mathbb{J}_A$ only in intrinsic cases.	Eq. (14.20)
$\mathsf{D}_{A,\lambda}$	Dynamic distance induced by the homogeneous momentum action J_A, with reference measure suppressed only in intrinsic cases.	Proposition 14.11
$A_p, J_p, \mathbb{A}_p$	Velocity action, momentum perspective and tangent action for $\mathcal{W}_p$.	Section 14.3
$\theta(a), J_\theta, \mathbb{A}_{\theta,\lambda}, \mathsf{W}_{\theta,\lambda}$	Concave mobility, its momentum action, reference-dependent tangent action and associated mobility distance.	Section 14.3
(s_t, g_t)	Source and relative growth variables in dynamic unbalanced OT.	Eq. (14.41)
$\mathbb{A}_\gamma(\alpha,v)$	Spectral tangent action induced by the gauge γ.	Eq. (14.23)
$\mathsf{W}_{\gamma,\mathrm{dyn}}$	Dynamic path-length representation of $\mathcal{W}_\gamma$.	Eq. (14.24)
Σ_n	Finite-state probability simplex in Markov-chain geometries.	Section 14.4
$K_{i,j}, \pi$	Reversible Markov transition rates and invariant law.	Section 14.4
$\theta(a,b)$	Logarithmic mean used as Markov/nonlocal mobility.	Eq. (14.26)
$\mathcal{K}_\rho$	Onsager operator for a discrete reversible Markov chain.	Eq. (14.32)
$\mathbb{A}_K(a,\psi), \mathcal{W}_K$	Markov-chain tangent action and associated discrete Wasserstein distance.	Eqs. (14.33), (14.34)
$\mathfrak{m}, K(x,\mathrm{d}y), \mathsf{J}$	Reference measure, reversible jump kernel and symmetric jump measure.	Eq. (14.27)
$\bar{\nabla}\varphi$	Nonlocal gradient/increment $\varphi(y)-\varphi(x)$.	Eq. (14.28)
$\mathbb{A}_K(\alpha,v), \mathcal{W}_K$	Nonlocal jump tangent action and associated Wasserstein distance.	Eq. (14.29)
$\mathcal{W}_k, \mathcal{H}_k^d, \kappa_k$	Kernelized Benamou–Brenier distance, vector-valued velocity RKHS, and uniform evaluation bound.	Eq. (14.25), Proposition 14.18
$\mathbb{W}f(\alpha)$	Wasserstein gradient of a functional.	Proposition 15.7
$\mathcal{I}(\alpha\|\beta)$	Relative Fisher information of α with respect to β.	Eq. (15.26)
ζ	Probability law of input-label pairs in the MLP and ResNet models.	Eq. (15.28)
$\mathcal{P}_{p,\lambda}(S\times\Omega)$	Conditional probability laws with fixed condition marginal λ and finite pth moment.	Section 11.6
$\mathcal{U}_\lambda(\alpha,\beta)$	Conditional couplings that keep the condition variable fixed.	Eq. (11.55)
$\mathcal{L}_c^\lambda, \mathcal{W}_{p,\lambda}$	Conditional OT value and, for $c_s = d^p$, its pth-root metric.	Eqs. (11.55), (11.56)
$\mathrm{PMO}_{\mathbb{A}_\gamma,\alpha}(g)$	Spectral specialization of the penalized minimization oracle.	Proposition 15.55
$S_\alpha(g), A_\alpha^\star$	Gradient covariance and active polar preconditioner in spectral flow.	Proposition 15.55
$\mathcal{P}_{2,\mathbf{m}}(\Omega;\mathbb{R}_+^p)$	Positive multi-species measures with fixed component masses.	Eq. (15.14)

Notation	Meaning	First reference
$\mathcal{W}_{2,\oplus}$	Mass-weighted product Wasserstein distance for independent species transport.	Eq. (15.15)
$\delta f(\alpha)$	First vertical variation of f at α.	Definition 15.2
$D^j f(\alpha)$, $\mathrm{Tay}^r_\alpha f$	Higher-order vertical derivative and vertical Taylor polynomial.	Definition 15.3, Remark 15.4
$\lvert\partial f\rvert(\alpha)$	Wasserstein or metric slope of an energy functional.	Definition 15.27
$\partial_t \alpha + \mathrm{div}(\alpha v) = 0$	Continuity equation.	Eq. (14.2)
$\alpha_{t+\tau}$	One JKO/minimizing-movement step.	Eq. (15.1)
$\mathcal{S} = C([0,1];\mathbb{R}^d), e_t$	Path space and evaluation map in the superposition formulation.	Section 14.2
$\Delta_\kappa, \mathrm{CW}_\kappa$	Scaled cone metric and static cone value for Wasserstein–Fisher–Rao transport.	Eqs. (14.42), (14.43)
$X, S, \alpha_X, F(X)$	Particle configuration, particle velocity, empirical law and lifted energy.	Eq. (15.61)
$s, \eta_t, \eta_t^n, \pi_x$	Velocity variable, phase-space laws and spatial projection in inertial flows.	Proposition 15.66
Generative models, neural dynamics and Gaussian closures		
u, π, P_t	Latent variable, latent law and interpolant map in flow matching.	Eq. (16.1)
$v_t, \partial_t P_t$	Eulerian flow-matching velocity and latent path velocity.	Proposition 16.3
$I_t(X_0, X_1, Z)$	Static-noise stochastic interpolant.	Definition 16.4
D_t	Conditional diffusion tensor in noisy interpolant/Fokker–Planck dynamics.	Eq. (16.9)
$f_\theta, \alpha_\theta, H_\beta(\theta)$	Generator, generated law and parameter-domain discrepancy objective.	Section 16
$\alpha(\mathrm{d}s, \mathrm{d}\theta) = \alpha_s(\mathrm{d}\theta)\mathrm{d}s$	Continuous-depth conditional law of neuron parameters.	Eq. (15.53)
z_s, α_s, Θ	ResNet state, depth-indexed neuron law and neuron parameter space.	Eq. (15.54)
$k, \nabla \log \beta, \varphi^\star_\alpha$	SVGD kernel, target score and RKHS steepest-descent field.	Section 16.2
$\varphi, \nabla\varphi, \varphi^\star$	Convex potential, its gradient map and Legendre transform in moment-measure models.	Section 16
Q, K, V	Query, key and value maps in attention dynamics.	Section 16.4
m_t, Σ_t	Gaussian mean and covariance evolving under a closed flow.	Section 16

APPENDIX B

Index

A

absolute continuity, 19, 20, 25, 58, 115, 129, 140, 257
absolutely continuous path, 318, 321
absorptive matching, 181
action density, 322
active mass, 208, 218, 220
additively homogeneous map, 162
admissible potential, 82, 84, 86
ADMM, 320
adversarial
    loss, 118
    perturbation, 74
affine
    closure, 412, 413, 417, 419
    constraint, 159, 274
    gradient, 412, 418
    map, 36, 84, 395
    marginal constraint, 158, 160
affine-difference cost, 59
affine-invariant metric, 39
Alexandrov solution, 24
algorithm
    Birkhoff–von Neumann decomposition, 49
algorithmic convergence, 158
alternating
    optimization, 88
    projection, 124, 158, 160, 163
    tree, 8, 10
alternating minimization, 304
angular covering, 197
antisymmetric flux, 377
antisymmetric velocity, 329
assignment problem, 1, 2, 7, 8, 41, 46, 48, 50, 51, 92, 285
Assouad lemma, 189
atom-preserving map, 78
atomic measure, 328
atomless measure, 17, 28
attention, 399, 406, 412, 419, 420
    causal, 79
    Gaussian kernel, 407
    linear, 206
    Lipschitz stability, 80
    map, 79
    masking, 79
    non-causal, 79
    operator, 80, 406
    self-attention, 79
    single-head, 406
auction algorithm, 7, 92, 94
    coordinate ascent, 92
augmented Lagrangian, 337

B

Baker-Campbell-Hausdorff formula, 310
Bakry–Émery criterion, 363
balance equation, 333, 334, 380
balanced
    transport between relaxed marginals, 207
balanced interpolation, 334
balanced OT, 142
Banach dual, 12, 13
Banach fixed-point theorem, 214
Banach space, 12
Banach-space embedding, 30
barrier
    entropy, 122
    logarithmic, 53, 122, 125
    LP, 122
    parameter, 53
barycenter, 243, 403
    fixed support, 243
    free support, 244
    law of large numbers, 252
    mean, 245
    support, 244, 245, 260
barycentric
    attention, 407
    constraint, 275
    mixture, 252, 283, 284
    projection, 43, 194, 196, 236, 273–275, 278
    transport, 275
basic feasible solution, 45, 260
Beckmann
    flow, 108
    formulation, 91, 107, 109, 325
    problem on graphs, 108, 109
belief propagation, 262
Benamou–Brenier, 138, 140, 279–282, 315, 317, 318, 321, 323, 326, 334
    distance, 323
    dual, 319
    formula, 315, 317
    formulation, 138, 317, 321
    kernelized, 327, 399
    viscous, 140
Bernstein–Schnabl operator, 76
Berry-Esseen
    bound, 184
    theorem, 71, 184
Besov smoothness, 190
Besov space, 191
Bessel function, 112
bias, 197, 198
bias-variance decomposition, 198
bias-variance tradeoff, 183, 190

biconvex
relaxation, 301
bilevel problem, 268
bilinear cost, 271
Birkhoff
contraction theorem, 166, 167, 169, 408
polytope, 46
von Neumann decomposition, 48
von Neumann theorem, 46–48, 58
birth-death dynamics, 379, 381
bistochastic matrix, 46, 48
block-coordinate
ascent, 251
block-coordinate descent, 264
block-coordinate method, 233
block-quotient norm, 164
block-stationary point, 264
Bobkov-Ledoux formula, 30
Bochner spectral measure, 114
Bochner theorem, 201
Boltzmann entropy, 120
Boltzmann equation, 386
book-shifting, 4, 19, 55
Borel measure, 11, 12, 20
Borel set, 11, 12, 14, 23
bounded confidence, 407
box constraint, 266
Bregman
divergence, 60, 114, 117, 126, 149, 158, 159
iteration, 308
penalty, 159
projection, 122, 151, 158–161, 163, 266, 309
regularization, 149–151
three-point formula, 160
Brenier
factor, 21
map, 20, 22, 23, 29, 33, 57, 82, 87, 88, 129, 176, 194, 234–236, 245, 349, 352, 353, 361, 362, 395, 406, 415
perturbation, 224
potential stability, 195
source hypothesis, 20
theorem, 19–21, 24, 25, 34, 36, 57, 58, 62, 87, 97, 234, 257
Brenier theorem, 301
Brenier–Strassen projection, 277
bridge
Brownian, 138, 140–142
linear, 396
OU, 394, 397
overshooting, 396, 397
speed, 396
variance-preserving, 394, 396, 397
Brownian bridge, 105, 138, 141
Brunn-Minkowski inequality, 361, 362
Bures
covariance, 36
metric, 34, 36–39, 254
Procrustes representation, 38
smoothed term, 176
soft Procrustes limit, 217
Wasserstein geometry, 34, 37, 175, 176, 178, 247, 284, 412, 415
Burg entropy, 117
Burgers equation, 147, 148

C

c-concavity, 89, 137
c-convexity, 27
c-cyclical monotonicity, 61
c-splitting set, 257
c-transform, 82, 84–89, 91, 96, 106, 128, 135
Caffarelli regularity, 23, 24
capacity-constrained OT, 265–267
capacity-constrained OT feasibility, 266
capacity-constrained Sinkhorn, 266
capped transportation polytope, 266
Carathéodory theorem, 48
Catalan number, 7
causality, 406
CDF formula, 30
cell mass, 96, 97
centered fluctuation, 198, 200
centered measure, 66
central limit theorem, 71, 183, 184, 253
empirical process, 186
entropic OT, 199
optimal transport, 197, 198
quantitative, 71
central path, 53
centroidal Voronoi, 103
centroidal Voronoi tessellation, 101
chain graph, 261
change of variables formula, 15, 24, 25, 39, 178, 246
characteristic kernel, 344
Choi matrix, 310
Cholesky factorization, 145
circle
transport, 5, 6
unfolded interval, 5, 6
classical convexity, 371
clustering, 407
co-motion function, 258
co-motion map, 258
collapsed mixture, 252, 283, 284
column constraint, 180
column normalization, 124
commutator, 310
compact sublevel set, 209
compactness, 330
complementary slackness, 7, 9, 82–85, 94, 273, 320
complete graph, 261
complete metric space, 12
completely positive cone, 203
completely positive kernel, 203
completely positive matrix, 203
complex
epsilon, 155
Gibbs kernel, 156
matrix scaling, 156
measure, 156
Sinkhorn coupling, 157
Sinkhorn iteration, 156
Sinkhorn scaling, 156

complexity
Birkhoff–von Neumann decomposition, 49
composition constraint, 349
computer vision, 51
concave
minimization, 292
concave cost, 2
concave envelope, 152
concavity
discrete OT, 43
Kantorovich value, 56
conditional
coupling, 241
expectation, 169, 276, 389–391, 393, 395
geodesic, 241
gradient flow, 383
independence, 263
Jensen inequality, 211
law, 32, 33, 140, 273–275, 321
mean, 274
neuron law, 382
optimal transport, 240, 381
probability, 12, 65, 283, 390
quantile, 234
transport, 240
variance, 275
Wasserstein distance, 240, 241, 381
Wasserstein flow, 381
conditionally negative definite kernel, 112, 155, 292
conditionally positive definite kernel, 112
cone
formulation, 212
lifting, 211, 334
metric, 211, 334
congestion, 265, 335, 346
congruence normalization, 310
conic lifting, 211
conjugate moment measure, 406
consensus, 407
Dobrushin contraction, 409
constant-speed geodesic, 23, 67, 290, 349
contact set, 27
continuity equation, 140, 279, 281, 282, 315, 317–320, 333, 338, 341, 375, 389–391, 393, 398, 401, 402, 412
continuous
bounded function, 70
coupling, 53, 54
depth, 383, 406
dual, 83, 135
path, 12
Sinkhorn dual, 149
continuous DTW, 312
continuous-depth neural network, 406
contraction ratio, 169
contractive Gaussian projection, 420
controlled transport, 412
convergence
in law, 31, 70, 155
in probability, 70
law, 69
mode, 70
convergence-determining class, 183
convex
action density, 323, 325
cone of functions, 59
conjugate, 87, 149, 163, 406
envelope, 88, 147
function, 20, 21, 25, 29, 34, 57, 58, 69, 74, 83, 87, 88, 118, 148, 149, 158, 343, 403
neural network, 137
order, 276–278
potential, 19–25, 61, 62, 85, 91, 138, 394, 403
potential generator, 406
regularizer, 148, 272
relaxation, 68
convex envelope
Kantorovich cost, 57
convex hull, 245, 409
convexity
along geodesics, 370
classical, 370, 371
constraint, 25
discrete OT, 43
geodesic, 338, 349, 350, 352–354, 356, 358, 361
Kantorovich value, 56
perspective transform, 323
strong geodesic, 358
convolution, 81
Wasserstein contraction, 81
convolutional Sinkhorn, 145
coordinate ascent, 92, 163
coordinate clearing, 180
copositive certificate, 205
copositive matrix, 203
correspondence, 43, 279, 286, 291, 307
cost
assignment, 50
bilinear, 27
derivative, 268
learning, 268
matrix, 1, 42, 45, 46, 63, 83, 123, 125, 164, 244, 292
reduced, 8, 9, 52, 109
regularized optimal transport, 295
transport, 56
unbalanced, 211
cost-sectional curvature, 27
Coulomb cost, 257–259
countable dense subset, 12, 283
counterexample, 258
coupling measure, 64
covariance
commuting, 395
cone, 37
diagonal, 39
Loewner order, 278
matrix, 29, 36, 39, 420
normalization, 375
operator, 236
subblock, 417
term, 36, 37
covering number, 187, 189
Cramér condition, 185
Cramer-von Mises distance, 31
Cramer-Wold theorem, 222, 255

created mass, 208
cross-covariance, 416
cross-curvature, 90
cross-moment, 297
crowd motion, 346, 348
Csiszár divergence, 352
cumulant
  fourth, 185
cumulative distribution function, 28–30
cumulative function, 4, 28, 31, 32
curse of dimensionality, 100, 187, 198
curve of maximal slope, 341, 372
cut locus, 27, 28
cycle graph, 261
cyclic
  Bregman projection, 159
  co-motion plan, 258
  monotonicity, 61, 62, 274
  projection, 159
  shift, 5
cylindrical Gaussian, 253

D

Dacorogna–Moser
  construction, 316
  inversion, 316
damped iteration, 248
damped second-order equation, 384
Danskin theorem, 199, 268
data-processing inequality, 118
DCA algorithm, 296
de la Vallée–Poussin criterion, 184
Delaunay graph, 109
delta method, 197, 199
demographic parity, 244
denoising score matching, 389
dense linear system, 122
density
  closure, 387
  constraint, 336, 346, 348, 349, 353
  estimation, 189, 190
  formula, 15
  matrix, 307, 308, 311
  ratio, 12, 114, 115, 117–119, 148–152, 265
  relative, 13
density-functional theory, 257–259
depth evolution, 406
depth variable, 382
derivative
  horizontal, 340
  vertical, 338
destroyed mass, 208
detailed balance, 378, 379
deterministic interpolant, 389
diagonal coupling, 276
diagonal dominance, 180
diagonal isotonicity, 179
diameter
  contraction, 409
difference-of-convex programming, 296
diffusion model, 78, 81, 389, 393, 396
  Gaussian noising, 131
diffusive map, 81
dimension dependence, 189
Dirac mass, 11, 14, 17, 29, 34, 41, 63, 68, 78, 245, 263, 338, 342, 349, 390
Dirac measure, 257
directional derivative, 268
Dirichlet energy, 343
discrepancy minimization, 397
discrete
  barycenter, 244, 259
  c-transform, 86
  coupling, 41
  Laplacian, 145
  measure, 11, 67, 70, 120
  space, 73
  target, 96
  Wasserstein geometry, 331, 332, 379
discretization error, 249, 390
discriminator class, 110, 111, 119
disintegration, 12, 32, 65, 140, 212, 240, 257, 273–275, 391, 398
displacement convexity, 22, 26, 351, 361, 367, 370, 403, 404
displacement covariance, 236, 238, 240
displacement interpolation, 22, 23, 31, 32, 240, 353
distance profile, 291
distance residual, 291
distortion, 285, 288, 291, 292, 305, 307
distributional robustness, 73
divergence-free field, 413
Dobrushin coefficient, 408
domain adaptation, 51
Doob transform, 139
doubly nonnegative cone, 203
doubly nonnegative matrix, 203
doubly positive kernel, 203
doubly stochastic normalization, 412
Douglas-Rachford algorithm, 320
Douglas-Rachford splitting, 320, 321
drifting
  field, 401, 412
  model, 399, 401, 402
dual
  alternate maximization, 151
  ascent, 143
  attainment, 84–86, 308
  block maximization, 137
  certificate, 7, 89
  formula, 118
  gap, 125, 158, 164, 165, 272, 273
  gradient, 98
  Hessian, 170
  norm, 110–113, 119
  objective gap, 163
  pairing, 12
  potential, 82, 83, 86, 88, 96, 124, 125, 128, 137, 148, 150, 153–155, 164, 169, 176, 187, 194, 197, 199, 200, 268, 272, 308
  price, 8
  problem, 82, 86, 91, 108, 121, 134, 137, 251
  rate, 164
  Sinkhorn, 143
  slack, 85

twist condition, 26
weight, 96, 97
duality
entropic OT, 136
Fenchel-Rockafellar, 148, 210, 274
strong, 108, 251, 307
weak, 10, 83, 274
Dudley entropy, 194
Dudley metric, 110, 111
Dvoretzky-Kiefer-Wolfowitz inequality, 197
dyadic cube, 189
dyadic partition, 188
Dykstra algorithm, 161, 266
dynamic
action, 318
distance, 324, 326
formulation, 317, 321, 324
optimal transport, 315
unbalanced OT, 333
value, 318
Wasserstein flow, 372
dynamic programming, 311
dynamic time warping, 311
algorithm, 312
numerics, 314
dynamical optimal plan, 68, 360
dynamical plan, 241

E

earth mover's distance, 51
edge flux, 332
edge imbalance, 31
Edgeworth expansion, 185
effective cost, 264
effective dimension, 202
effective rank, 202
eikonal equation, 421
electron density, 258
elliptical law, 36
elliptically contoured distribution, 36
embedding
alignment, 232
empirical
context, 406
distribution, 346
dual potential, 194
energy, 387
law, 73, 75, 183, 346
measure, 3, 11, 18, 44, 48, 70, 76, 104, 154, 183, 187, 191, 229, 267, 270, 289, 315, 344, 345, 376, 382, 384, 398, 406
measure over measures, 252
Monge map, 16
OT, 100, 189
OT rate, 104, 189, 290
process, 186, 194, 197, 200
token measure, 79
Wasserstein metric, 384
endpoint coupling, 139–142
energy
decay, 356
dissipation, 355–357, 385
interaction, 153, 343–345, 350, 415
energy distance, 112, 155, 344, 345, 351, 386
energy-dissipation identity, 355
energy-dissipation inequality, 342, 355
entropic
barycenter, 246, 251
barycentric map, 195
bias, 125, 132, 152, 153
map estimation, 196
OT, 120–122, 131, 134, 153, 163, 175, 197, 199, 269, 273, 304, 308, 401
path, 122
plan, 206
potential, 138, 153
regularization, 88, 89, 96, 120, 122, 153, 158, 187, 192, 197, 213, 260, 264, 266, 270, 301, 308
smoothing, 121, 163, 196, 250
temperature, 178
entropy, 404
closure, 387
decay, 363
function, 114–116, 148, 207
penalty, 120
score, 387
envelope theorem, 121, 199, 268, 269
ε-scaling, 95
equal marginal, 258
equality constraint, 52, 108
essential continuity, 403
Euclidean group, 232
Eulerian
description, 315
velocity, 316, 318, 319
exact OT, 24, 187, 189, 196
explicit Euler scheme, 384, 406
extreme minimizer, 47
extreme point, 46–48, 52

F

factored coupling, 263
fairness, 244
farthest-point sampling, 387
fast diffusion equation, 374
fast Legendre transform, 146, 147
feasible coupling, 52, 55, 235, 290
feature cost, 304
feature sketch, 200
feature term, 305
Fenchel duality, 143, 148, 149, 210, 274
Fenchel inequality, 138
Fenchel-Rockafellar duality, 319, 320
Fenchel-Young loss, 270
FFT, 147
fiber, 241
fiberwise optimal plan, 241
fiberwise transport, 240
fill-in edge, 261
filtered back-projection, 255
finite difference scheme, 24, 388
finite measure, 12, 110, 111, 209, 279
finite metric length, 355
finite space, 197, 198
finite-action component, 324, 328
finite-dimensional linear program, 106, 108

finite-sample bias, 398
finite-state transport, 331
first variation, 149, 268, 269, 309, 338, 340, 356, 369–371, 383, 401, 402, 412, 415, 417, 419
Fisher information, 133, 140, 362, 364, 367, 415
Fisher-Rao metric, 34, 35, 39, 40
fixed-point
  Gaussian barycenter, 248
flattening map, 252
flow
  conservation, 52
  continuous normalizing, 390
  cross-covariance, 416
  map, 315, 395
  minimum-cost, 52, 109
  momentum, 384
  normalizing, 390
  one-step, 398
  parameter-domain, 397
  phase-space, 385
  probability ODE, 389, 390, 392–394, 396
  rectified, 389, 390, 394
  reverse, 397
  second-order, 384
flow matching, 317, 389–391, 393–395, 419
flux, 279, 281, 316, 341
Fokker–Planck particle closure, 401
Fokker–Planck
  equation, 345–347, 363, 368, 374, 378, 380, 381, 392, 395, 396
  fractional equation, 378
Fortet iteration, 163
Fourier features, 200
Fourier multiplier, 112
Fourier random features, 201
Fourier transform, 254
Fourier-slice theorem, 255
fractional diffusion, 378
fractional heat equation, 378
fractional Laplacian, 377, 378
fractional PDE, 377
Frechet inception distance, 114
Frechet mean, 243, 245
Frobenius inner product, 37
Frobenius norm, 158, 238
frozen surrogate, 402
functional
  derivative, 338
  on Wasserstein space, 75
functional inequality, 361, 362, 365
fused Gromov-Wasserstein, 304, 305

G

Gamma convergence, 131
GAN, 118
gauge direction, 180
gauge fixing, 177
gauge invariance, 181
Gaussian
  barycenter, 247, 248
  blurring, 131
  closure, 178, 217, 399, 412, 417, 419
  complex Sinkhorn, 157
  convex order, 277
  convolution, 146, 147
  covariance, 176, 414
  denoiser, 393
  flow matching, 394, 395
  kernel, 201
  kernel attention, 407
  kernel drifting, 402, 419
  kernel rank, 202
  manifold, 412–414, 417
  marginals, 138, 174
  mean, 414
  measure, 18, 22, 34, 35, 247, 278, 364, 403, 412, 418, 419
  mixture, 3, 97, 102, 124, 125, 128, 152, 208, 246, 265, 267, 284, 394, 396
  one-dimensional, 417
  optimal map, 29
  preservation, 412, 413, 419, 420
  preserving flow, 419
  projection, 417, 420
  push-forward, 35
  scale mixture, 205
  Sinkhorn, 174–176, 178, 416
  Sinkhorn divergence, 176
  Sinkhorn flow, 178
  Sinkhorn problem, 178
  sliced Wasserstein, 226
  submanifold, 412, 418
  unbalanced, 215
Gaussian Hellinger-Kantorovich, 215
Gaussian measure
  Gromov–Wasserstein, 298
Gelbrich inequality, 420
Gelbrich theorem, 420
generalized
  inverse, 17, 28
  Wasserstein, 207
generalized Gaussian kernel, 204, 205
generalized quantile, 32
generalized Sinkhorn algorithm, 142, 144
generalized Wasserstein flow, 372
generative model, 110, 227, 315, 354, 389, 397, 406
generator manifold, 398
geodesic, 329
  convexity, 349, 353, 356
  distance, 25, 32, 145, 322
  semiconcavity, 354
  space, 67, 360
geodesics in heat, 144, 145
Gibbs
  coupling, 308, 309
  formula, 308
  kernel, 124–126, 129, 143, 155, 159, 161, 169, 201
    push-forward, 130
    sketching, 205
  path law, 313
  reference, 159
Girsanov theorem, 139
global
  minimizer, 370, 371
  optimality, 371

global invariance, 232
gluing lemma, 63–66, 213, 288
gradient
  field, 20
  flow, 22, 286, 290, 315, 317, 338, 342–344, 346, 349, 355, 356, 369, 370, 389, 399, 412, 413, 419, 420
  inequality, 356
  kernelized flow, 399
  noise, 378
  projection, 402
  structure, 280, 317, 411
  velocity, 316, 317, 413
graph
  augmenting path, 10, 49
  bipartite, 45, 48
  coupling, 55, 68
  equality, 8–10
  geodesic distance, 108
  matching, 285
  support, 45, 49
  transport, 109
graphical model, 260
Grassmann manifold, 227
gravity interaction, 386
greedy sweep, 45
Gromov–Wasserstein
  Gaussian measures, 298
Gromov-Hausdorff distance, 286, 305, 306
Gromov–Monge, 300, 307
Gromov–Wasserstein, 232, 279, 285–288, 290, 304, 307
  compactness, 289
  completeness, 289
  conditional concavity, 292
  distance, 279, 285
  geodesic, 290
  metric, 288
  sample complexity, 289
gross substitutes, 179
ground cost, 63, 106, 252, 268, 270, 285, 290
  adversarial, 295
ground metric learning, 66
group action, 231
growth term, 333
Gurvits scaling, 309, 310

H

Hadamard differentiability, 200
Hall theorem, 49
Hamilton-Jacobi equation, 140, 147, 319, 320
Hamilton-Jacobi-Bellman equation, 336
hard congestion, 336
Hausdorff distance, 286, 305–307
Hausdorff measure, 20
heat equation, 147, 342–344, 387
heat flow, 81, 331
heat kernel, 144–146
heavy-ball method, 384, 385
heavy-tailed noise, 378
Hegselmann–Krause model, 407
Hellinger
  divergence, 117
  geodesic, 217
  geometry, 39
Hellinger-Kantorovich, 334
Helmholtz decomposition, 413
Hermitian matrix, 307, 309
hidden convexity, 404
Hilbert
  diameter, 168
  embedding, 226
  metric, 165
  projective metric, 165, 167, 168
Hilbert space, 354
Hilbertian
  embedding, 235
  limit, 155
Hilbertian tangent norm, 322
Hilbertian Wasserstein space, 30
histogram, 4, 11, 30, 32, 63, 64, 121, 156, 164, 207, 244, 279, 285
Hoeffding inequality, 202
Holder curve, 355
Hölder inequality, 238
holomorphic continuation, 155
holomorphic implicit function theorem, 156
homogeneous action, 324
homogeneous formulation, 210, 211
homogeneous Sobolev norm, 155
homogenization, 211
Hopf-Cole transform, 144, 145, 147, 148
Hopf-Lax formula, 145, 147
Horn matrix, 205
Hungarian primal-dual method, 1, 7–9
HWI inequality, 367, 368
hyperbolic geodesic, 26
hyperbolic metric, 34
hypersurface, 20

I

ice-cream cone, 37
imitation learning, 119
implicit Euler scheme, 338, 372
in-context map, 406
incompressible Euler, 262
increasing rearrangement, 60
induced width, 261
inertia, 386
inertial particle method, 385
infinite-depth limit, 406
infinite-dimensional linear programming, 56
inner product
  Gromov–Wasserstein, 297
input-convex neural network, 90, 137, 138, 406
integer multiplicity, 51
integral probability metric, 109–111, 117–119, 398
interacting particle system, 368, 371
interaction energy, 360
interaction kernel, 386
interaction polynomial, 76
interior-point method, 52, 122, 125
intermediate measure, 263
internal energy, 351
interpolant, 396
intrinsic dimension, 183, 187, 189
intrinsic length metric, 225
inverse isotonicity, 179

inverse OT, 268, 270, 272
inverse OT gap loss, 270
inverse square root, 310
irreducibility, 330
isometric embedding, 306
isoperimetric inequality, 361, 362
isotonic regression, 60
iterative closest point, 232
iterative proportional fitting, 122, 123

J

Jacobi iteration, 180
Jacobian determinant, 15, 352
Jacobian equation, 27, 178
Jensen inequality, 69, 115, 127, 184, 192, 195, 275, 318, 327
Jensen-Shannon divergence, 117, 119
JKO scheme, 142, 338, 341, 342, 348, 372, 375, 379, 380, 384, 418, 420
joint
    law, 53, 56, 259
    measure, 53
Jordan decomposition, 13
jump kernel, 328
jump process, 328, 377
junction tree, 261

K

Kantorovich
    axiomatic characterization, 57
    coupling, 58
    dual, 320
    duality, 7, 20, 27, 82–84, 87, 272, 274, 307, 404
    finite equal-mass measures, 207
    interpolation, 22
    potential, 83, 85, 257, 268
    problem, 43, 48, 50, 52, 55–58, 62, 68, 82, 128, 138, 141, 265, 274, 282
    relaxation, 1, 16, 20, 41, 43, 48, 49, 53, 57, 68, 207
Kantorovich–Rubinstein
    duality, 71, 80, 111, 119, 184
    formula, 106–108
    norm, 31, 106
kernel
    attention, 206
    characteristic, 113
    density estimate, 387, 399, 407
    drifting, 402
    energy, 112, 113
    Gaussian, 112, 126, 401, 411, 419
    Laplacian, 112
    logarithmic accuracy, 202
    low-rank approximation, 200
    Matérn, 112
    mean embedding, 113, 191, 235
    norm, 112, 113, 119, 155, 350, 401
    positive, 112, 153, 165, 401
    positive definite, 155, 191, 226, 344
    positivity, 202
    quadrature, 344
    regression, 407, 411
    ridge regression, 61, 201
    Riesz, 112
    sample complexity, 200
    sketching, 200
    smoothing, 190
    universal, 113, 155
kernelized self-interaction, 344
kernelized Stein discrepancy, 399
kernelized transport distance, 327, 328
kinetic action, 140
kinetic equation, 386
kinetic mean-field limit, 385
KL
    projection, 109, 143, 158–160, 163, 164, 250, 266, 310
    proximal map, 143
    relaxation, 213
KL barycenter, 248
KL proximal operator, 142–144
Knothe–Rosenblatt rearrangement, 32, 33
Kolmogorov forward equation, 379
Kolmogorov-Smirnov distance, 31
Kullback-Leibler divergence, 34, 39, 116, 126, 127, 129, 142, 358
Kurdyka–Łojasiewicz inequality, 356, 358–360

L

L2 attention, 407
L2 self-attention, 407
Lagrange multiplier, 347
Lagrangian
    description, 315
    flow, 315
Lagrangian velocity, 385
Laguerre cell, 18, 87, 96–98, 103
    discrete, 92
Langevin
    dynamics, 139, 347
    free energy, 345
Laplace method, 177, 178
Laplace-Beltrami operator, 144
large-temperature collapse, 153
large-temperature limit, 155
latent
    atom, 263
    mass, 264
    variable, 263, 390, 392
latent measure, 263
lattice distribution, 185
Lavenant criterion, 420
law of large numbers, 183, 252
law over measures, 252
layer normalization, 406, 412
layer-cake formula, 30
layerwise matching, 383
least squares
    Radon reconstruction, 254
least-square
    inversion, 317
    velocity, 317
Lebesgue decomposition, 114, 210
Lebesgue measure, 13, 17, 20, 21, 57, 127
Legendre function, 158
Legendre transform, 88, 118, 251, 404
Legendre-Fenchel transform, 146, 147
length metric, 225
length space, 322

Lévy process, 378
linear
  attention, 206
  convergence, 125, 161, 163, 165, 167, 214, 357
  form, 12, 13, 47, 371
  functional, 342
  linearization, 25, 303
  objective, 48
  operator, 44
  OT, 234, 235, 237
  potential, 342
  programming, 43, 44, 47, 52, 122, 243, 266
  tilt, 159
linear-time Sinkhorn, 200
linearized Wasserstein metric, 365
Liouville equation, 385
Lipschitz
  constant, 106
  function, 19, 91, 106, 108, 110, 112
  map, 78
  stability, 86
  test function, 71
Lloyd
  algorithm, 100, 354
  flow, 101, 102
local action, 322, 372
local distance distribution, 285
local matching indicator, 2
local metric slope, 341
local profile, 285
local tangent action, 322
localization, 24
Loewner order, 277, 327
log-concave
  density, 403
  measure, 403
log-concave generator, 406
log-concavity, 363
log-domain Sinkhorn, 121
log-scaling, 179
log-sum-exp, 96, 128, 135, 137, 145, 153, 262
logarithmic error, 202
logarithmic kernel approximation, 202
logarithmic mean, 329, 331, 332
logarithmic Sobolev inequality, 362, 364, 366–368
LogDet divergence, 215
look-ahead point, 384
Lorentz cone, 37
loss factor, 181
lossy Sinkhorn clearing, 181, 182
low-rank
  coupling, 206, 263
  optimal transport, 262, 263, 265
lower semicontinuity, 56, 114, 115, 154, 209, 210, 252, 274, 279, 330, 334
lower semicontinuous envelope, 68, 69
lower semicontinuous relaxation, 324
lower-dimensional set, 20
Lyapunov equation, 413

M

M-function, 158, 179, 180, 182
M-matrix, 180, 181
M0-function, 180
Ma-Trudinger-Wang condition, 25, 27
Maas distance, 331
manifold support, 187, 189
many-token regime, 79
map estimation, 196
map extrapolation, 194
marginal
  constraint, 42, 45, 53, 56, 109, 123, 124, 127, 129, 141, 148, 150, 158, 160, 163, 169, 212, 274, 276, 307
  divergence, 208–210
  penalty, 142, 144, 209, 210
  relaxation, 208–210
  relaxed, 207
  residual, 121, 164
marginal-dependent problem, 142, 143
market clearing, 179
Markov
  averaging, 407
Markov chain, 331, 379
Markov kernel, 32, 81, 166, 285
  Wasserstein stability, 81
Markov semigroup, 377, 379
Markov-chain Wasserstein geometry, 377, 379
martingale coupling, 275, 276
martingale OT, 275–278
masking, 406
mass
  balance, 2, 97
  conservation, 41, 123, 207
  creation, 207
  free, 100
  shape separation, 208
  splitting, 26, 41, 69, 81
  variation, 209
mass flux, 336
matching
  algorithm, 7
  non-crossing, 6, 7
  optimal, 1, 62
  partial, 9
  perfect, 7, 9, 49
  uniform, 46, 48, 50, 51, 57, 62
matrix
  cone, 281
  incidence, 109
  permutation, 44, 46–48, 285
  perspective, 326
  polar decomposition, 22
  scaling, 120, 122, 123, 155, 156, 158
  square root, 174
  subproblem, 282
matrix-valued measure, 279, 281, 282
matrix-vector iteration, 122
matrix-vector product, 201
max-flow min-cut theorem, 266
maximal correlation, 169, 404
maximal density constraint, 346
maximum mean discrepancy, 111–114, 118, 119, 187, 191, 192, 235, 344, 350, 351, 365, 386, 398, 414, 415
McCann
  displacement convexity, 351

interpolation, 23, 26, 32, 33, 67, 240, 244, 321, 349
McKean-Vlasov equation, 345, 346
McKean-Vlasov limit, 345, 346
mean
Wasserstein decomposition, 66
mean field game, 335
mean shift, 407
blurring, 407
iteration, 407
vector, 407
mean-field limit, 80, 346, 361, 406
mean-field momentum, 385
mean-field neural network, 338, 361, 369, 383
mean-preserving splitting, 275
mean-shift PDE, 407
measurable function, 12, 110
measurable selection, 138, 139, 241, 285
measure
atomless, 17, 68
evolution, 315, 316, 342
isomorphism, 17
measure convolution, 71
measure-function duality, 83
measure-preserving
isometry, 286, 288
map, 19, 21, 22, 300
measure-to-measure map, 77
measure-to-vector map, 75
measure-valued Radon transform, 221
mechanical energy, 385
Mémoli profile, 285, 290, 291
mesh size, 145
message passing, 261
metric
derivative, 224, 341, 355
entropy, 187, 189
equivalence, 238
learning, 74, 268, 269
length, 355
length bound, 355
pull-back, 221
slope, 356, 359
space, 12
triangle inequality, 66
metric-measure space, 279, 285, 286, 288, 290, 360, 361
min-plus algebra, 89
Min-SW, 229, 231
minibatch, 397
minibatch bias, 397
minibatch noise, 379
minimal velocity, 342
minimax, 236, 239
minimax lower bound, 189
minimax optimality, 189
minimizer
sparse, 44
minimizing movement, 338, 342, 375, 420
minimizing movement scheme, 338
Minkowski inequality, 19, 64, 66, 79, 222, 288
Minkowski problem, 403
mixed Hessian, 27
MMD-GAN, 397
mobility, 325, 373
concave, 325, 326, 373
convexity mechanism, 325
nonlocal, 329
modulo isometry, 288
modulus of continuity, 86
moment
convergence, 72
moment closure, 421
moment condition, 183
moment energy, 360
moment functional, 413, 415
moment matching, 412, 420
moment measure, 403, 404, 406
momentum, 279, 280, 318–320
momentum formulation, 318, 322, 323, 336
momentum variable, 322
Monge
distance, 18, 19, 23, 68, 69
geodesic, 22, 350
interpolation, 22, 30
map, 11, 16, 17, 19, 257
problem, 1, 11, 16, 17, 19, 20, 22, 28, 29, 33, 34, 41, 55, 57, 66, 68, 317, 318, 349
Monge gap, 59
convexity, 59
isotonicity, 60
kernel regression, 61
one-dimensional formula, 60
Monge map
Gromov–Wasserstein, 301
Monge–Ampère
complex equation, 178
equation, 23–25, 177, 235
geometric equation, 177
parabolic equation, 177, 179
solver, 24
Monge-Kantorovich equivalence, 57, 257
monotone
alignment, 311
convergence, 162
field, 20
fixed point, 163
matching, 1, 5, 229
rearrangement, 3, 4, 20, 29, 32, 33, 58, 62
spectral gauge, 236, 326
transport map, 399
Moore-Penrose pseudo-inverse, 254
MTW condition, 27
multi-marginal
cost, 259
formulation, 259
Monge map, 256, 257
OT, 243, 244, 256, 259, 260
Sinkhorn, 259
multi-omics, 305
multi-species
flow, 348
gradient flow, 348, 349
multi-valued subdifferential, 20
multinomial distribution, 199
multiplicative scaling, 143

multiscale coupling, 188
Muon algorithm, 236, 239, 327, 374, 376, 377
Muon polar factor, 375
mutual information, 129

N

Nash equilibrium, 335
natural gradient, 398
natural language processing, 67
near-linear Sinkhorn, 202
nearest neighbor, 232
negative Shannon entropy, 158–160, 343
negative Sobolev norm, 366
Nesterov acceleration, 384
network simplex, 109
neural
    dual solver, 137, 138
    Sinkhorn solver, 137
neural optimal transport, 90
neuron, 368–371, 382, 417
neuron law, 369, 416, 417
Newton dynamics, 386
Newton step, 53, 122
Newton–Schulz iteration, 376
Newtonian dynamics, 386
node feature, 305
noising path, 396
noising schedule, 394
noisy gradient descent, 371
non-convex domain, 24
non-convexity, 354
non-Gaussian limit, 200
non-variational scaling, 179, 181
nonconvex
    optimization, 244
nonlinear flattening, 252
nonlocal
    action, 328
    continuity equation, 329
    gradient, 329
nonlocal Wasserstein
    distance, 328, 329, 377
    flow, 377
nonnegative
    orthant, 160
nonnegative factorization, 203
nonnegative measure, 149, 150, 207, 212
nonnegative rank, 263
nonnegative transport matrix, 49
norm
    flat, 110, 111
    RKHS, 113
    Wasserstein-1, 106
normal approximation, 184
normal cone, 346
normalized dynamics, 374
normalized gradient flow, 373, 374
normalized SGD, 374
north-west corner
    plan, 45
    rule, 45, 52
Nystrom approximation, 202

O

ODE limit, 384
one-dimensional
    Monge map, 28
    projection, 18, 207, 221, 223, 416
    transport, 21, 28, 31, 45, 62, 350
one-form, 181
one-step generative model, 389, 397, 406
Onsager operator, 322, 331, 332
operator bound, 239
operator gauge, 327, 374
operator norm, 13
operator scaling, 307, 309, 310
operator-valued coupling, 310
opinion dynamics, 407
optimal coupling, 28–30, 44, 48, 50, 55, 57, 58, 62, 65, 67, 82, 83, 88, 100, 245, 250, 259, 268, 270, 273, 288, 290, 318, 349
optimal face, 199
optimal plan, 20, 27, 42, 44, 45, 62, 84, 88, 124, 148, 273, 288, 349
optimal transport map
    consistency, 194
    estimation, 194
optimality
    first-order, 97, 160, 309
orbit space, 231
order interval, 180
order-constrained coupling, 276
order-preserving map, 162
Ornstein-Uhlenbeck flow, 364, 368
Ornstein-Uhlenbeck process, 387, 394, 396
OT map estimation, 194
OT on trees, 31
OU bridge, 396
outside option, 181
overshooting bridge, 396

P

pair space, 329
pair-space action, 328
pairwise distance, 286
parabolic optimal transport, 177
parameter space, 397
partial matching, 8
partial transport, 209, 212, 218, 220
    active region, 221
    fixed mass, 220
    TV penalization, 219
particle method, 78
particle ODE, 315, 345, 368
particle polynomial, 339
particle splitting, 328
particle Wasserstein descent, 384
path action, 322
path entropy, 313
path measure, 318, 327
path metric, 225
path space, 12, 68, 138–142
path-space formulation, 120, 138–141, 321
path-space problem, 139
path-space transport, 138, 141

Pearson divergence, 117
penalized minimization oracle, 372
permutation
  equivariance, 79
permutation matrix, 46
Perron-Frobenius theorem, 162, 165, 166
perspective function, 318–321, 323
perspective recession, 279
perspective transform, 211, 323
perturbation response, 398
phase space, 385
phase-space density, 388
phase-space equation, 387
phase-space measure, 385
phi-divergence, 110, 114, 115, 117, 118, 126, 148–150, 353, 365
phi-divergence regularized OT, 148–151, 162
piecewise affine map, 197
Pinsker inequality, 116, 164
PL convergence, 356
plan
  interpolation, 67
  lifted, 229, 231, 240
  optimal, 67
  sparse, 42, 44
  transport, 41, 50, 61, 67, 83, 240, 270, 279
plug-in estimator, 183, 190, 191
PMO, 372
Poincaré disk, 26
Poincaré inequality, 365, 366
Poisson equation, 145
polar factor, 376
polar factorization, 21, 22
polar formula, 239
polar set, 236, 238, 239, 374
Polish space, 12, 17, 64, 65, 75, 274, 283
Polyak momentum, 384
Polyak-Lojasiewicz inequality, 356
population law, 335
porous medium equation, 343, 374
Portmanteau theorem, 56
positional encoding, 79
positive
  autocorrelation, 205
  cone, 279, 280, 309
  feature map, 203
  features, 203
  linear map, 167
  matrix-valued measure, 281
  measure, 380
  operator, 279, 307
  random features, 200, 204–206
  semidefinite cone, 37
  semidefinite kernel, 200, 201
  semidefinite matrix, 34, 36, 37, 236, 279, 280, 282, 308
  sketch, 205
  vector-valued measure, 279
potential energy, 360
potential game, 335
power diagram, 96, 98
power divergence, 352
power-law jump kernel, 378
Prékopa
  inequality, 405
  Leindler inequality, 362
pressure field, 347–349
primal-dual
  method, 92
principal component analysis, 235–237
principal value, 378
probabilistic coupling, 14, 15
probability integral transform, 28
probability measure, 11, 12, 14, 17, 19, 30, 33, 53, 56, 63, 70, 75, 84, 111–113, 117, 150, 221, 226, 234, 243, 256, 283, 287, 290, 317, 318, 344, 370
probability path, 390
probability simplex, 11, 122, 150, 166
Procrustes
  alternating minimization, 233
  Bures metric, 38
  orthogonal, 232
  problem, 232
  rigid update, 232
  soft, 217
product
  coupling, 42, 54, 56, 65, 68, 153, 265, 394
  measure, 124, 129, 151, 260
product Wasserstein metric, 348
profile lower bound, 232, 285, 290
projected formulation, 236
projected measure, 221
projection direction, 221
projection-free attention, 407
projection-robust Wasserstein, 227
projective
  cone, 165
  diameter, 166
  geometry, 162
  metric, 165
Prokhorov theorem, 56, 75
propagation of chaos, 346
proximal
  operator, 320
  splitting, 320
pullback, 14, 15
push-forward, 11, 14–18, 28, 29, 35, 36, 41, 316, 340, 389, 391
  measure, 79
  stability, 78
Pythagorean identity, 158, 160, 164

Q

quadratic
  action, 322
  assignment problem, 285
  barycenter, 245, 259
  closure, 174
  confinement, 387
  cost, 3, 17, 20, 22, 23, 25, 27, 29, 33, 36, 83–85, 87, 96, 97, 100, 137, 169, 175, 207, 236, 243, 273, 318
  energy, 386
  form, 292
  perspective, 318, 323
  potential, 20, 35, 36, 174, 413, 415

regularization, 161
scaling, 157
tangent norm, 322
quantile, 28
barycenter, 246
coupling, 58
formula, 284
function, 4, 28–31, 221, 226, 234–237, 246, 254
map, 16, 29, 30, 280
process, 105, 196, 197
push-forward, 28
quantization, 104
quantization
energy, 101, 102
fixed weights, 103
high-resolution limit, 102
mean-field limit, 102
non-uniform weights, 100
optimal, 99, 354
rate, 100
quantum
Bregman projection, 309
coupling, 307
Kantorovich duality, 307
OT, 279, 307, 308
OT duality, 308
Sinkhorn, 309
quotient
metric property, 231
Wasserstein distance, 231, 232

R

Rademacher bound, 194
radial coupling, 21
radial generator, 36
radial measure, 20
Radon barycenter, 254
Radon measure, 12, 13, 61, 84
Radon pseudo-inverse, 254
Radon range condition, 222, 254
Radon sinogram, 254
Radon structure, 223
Radon transform, 221, 254
Radon Wasserstein barycenter, 254
Radon Wasserstein distance, 222
Radon-Nikodym derivative, 391
ramp filter, 255
random features, 200, 201
random measure, 252
random probability measure, 252, 283
random variable, 14, 16, 56, 70, 184, 413
random walk, 332
rank, 44
rank constraint, 236
rank-deficient covariance, 39
rate
parametric, 183, 191, 196, 198
sublinear, 99, 163
rational weights, 49–51, 283
raw second moment, 37
reaction term, 380
reaction–transport action, 333
reaction-transport geometry, 333
recession convention, 209, 210, 279, 319
rectifiable set, 20
reference
distribution, 389
measure, 13, 22, 118, 126, 127, 149, 150, 210, 234, 235, 249, 389
reflow, 390
regression loss, 389
regular conditional distribution, 32
regular vector field, 328
regularity theory, 23, 24, 27, 28
regularization
Monge gap, 60
regularization bias, 183, 192, 198, 200
reinforcement learning, 119
relative
entropy, 116, 120, 126, 129, 140–142, 175, 345, 352, 353, 358, 360, 361, 377, 380, 415, 417
error, 202
Fisher information, 363, 366, 417
relative entropy
regularity, 76
reproducing kernel Hilbert space, 112, 113
rescaled convolution, 71
residual attention, 406
residual block, 381
residual network, 383
ResNet, 381
continuous-depth, 382
finite-depth, 381
finite-depth mean-field, 382
finite-width, 381
infinite-depth, 381, 383
particle discretization, 383
resolvent, 145
reverse formulation, 210
reverse KL divergence, 117
reversible jump kernel, 328
reversible Markov chain, 331
Riccati equation, 178
Ricci curvature, 361
Richardson–Romberg extrapolation, 339
Riemann sum, 70
Riemannian manifold, 25, 361
Riemannian metric, 322
Riemannian tensor, 322
Riesz potential, 112
Riesz-Markov-Kakutani theorem, 12, 13
rigid motion, 232
rigid registration, 232
rigid update, 233
RKHS, 111–114, 119, 191, 192, 235
cost regularization, 296
vector field, 327
velocity, 327
robust
envelope, 75
finance, 275
optimal transport, 295
representation, 238
Wasserstein, 236, 239
Wasserstein distance, 236, 239

rotational component, 29, 395
row constraint, 180
row normalization, 124
running-intersection property, 261

S

Sakoe–Chiba band, 311
sample complexity, 100, 183, 187, 253, 289
sampling, 78
scalar bridge, 396
scaling
    algorithm, 120, 260
    column, 124, 158, 160
    form, 122, 123, 134, 160, 250, 310
    potential, 125
    row, 124, 158, 160
    vectors, 161, 165
Schatten norm, 238
Schoenberg theorem, 205
Schrödinger
    bridge, 138–141
    interpolation, 140
    problem, 120, 133, 138–141, 155
    system, 163
score function, 343, 393, 394, 396, 401
score-based generative modeling, 389
score-SDE, 389
second boundary condition, 24, 25
second moment, 37, 415, 416
second-moment quotient, 37
second-order limit, 200
sectional curvature, 28
self Sinkhorn, 155
self-concordance, 122
self-corrected field, 402
self-distance, 198
semi-coupling, 211
semi-discrete
    dual, 97, 98
    Monge map, 17, 18
    OT, 18, 55, 87, 91, 96–99, 136
semi-dual, 91, 96, 97, 250
    discrete, 91
    entropic, 173
semi-relaxed problem, 86
semidefinite program, 307
semidefinite programming, 191
separable space, 12
separator, 262
Shampoo optimizer, 374
Shannon entropy, 115, 116, 120, 308, 331, 343, 344, 349, 353
Shannon-Boltzmann entropy, 120
shape space, 231
shearing component, 29, 395
signed
    measure, 13, 31, 70, 106, 112, 113, 338, 340, 344
    negative part, 109
    positive part, 109
simplex, 332
    projection, 161
simplex network, 52
single-cell genomics, 51, 52, 66, 209, 305, 342, 398
Sinkhorn
    acceleration, 169
    algorithm, 88, 96, 109, 120, 122, 123, 125, 135, 167, 169
    barycenter, 236, 249, 335
    closed-form coupling, 130
    continuous dual iteration, 137
    continuous flow, 177, 179
    continuous iteration, 129
    continuous limit, 177
    convergence, 158, 162, 163, 165
    coupling, 141
    divergence, 120, 152–155, 176, 187, 192, 200, 346, 398, 399, 416, 417
    dual, 134, 149, 164
    fixed point, 156, 158
    full-cycle map, 168
    GAN, 397
    half-step, 125, 126, 160, 164
    holomorphic continuation, 155–157
    iteration, 125, 153, 161, 174, 250, 262, 273, 304
    Jacobian, 170
    kernel, 201
    linear-time, 200
    local convergence, 169
    low-rank, 263, 264
    over-relaxation, 172
    potential, 177
    scaling, 120, 122, 124, 142, 148, 151, 159, 168, 179–181, 214, 310, 401
    sketching, 200
    statistical convergence, 183
    update, 309
skip connection, 406
slack variable, 8
Slater condition, 307, 308
sliced barycenter, 254
sliced length metric, 225
sliced tangent norm, 224
sliced Wasserstein
    barycenter, 254
    CDF estimator, 30
    compact comparison, 223
    directional aggregation, 227
    distance, 196, 221, 222, 227, 235, 416
    flow, 399
    Gaussian, 226
    length, 225
    max, 227
    Monte Carlo directions, 197
    reverse bound, 223
    sample complexity, 196
    tangent norm, 224
slope
    metric, 341
    Wasserstein, 356, 358
small-jump limit, 331
smooth density, 191
smooth OT estimation, 191
smoothed plug-in estimator, 190
smoothing estimator, 190
smoothness, 190

Sobolev norm, 155
Sobolev space
  learned cost, 296
soft
  barycentric projection, 194
  c-transform, 134–137, 146, 151–153, 162, 164, 174, 178, 194, 213, 214, 269
  conjugation, 194
  minimum, 135
  transform, 120, 135–138, 158, 174, 176, 213
soft-DTW, 312
  expected alignment, 314
softmax attention, 206
source term, 333
spanning tree, 52, 109
sparse linear program, 109
sparsity, 44
spectral
  dynamics, 239
  gap, 366
  gauge, 207, 236, 238, 240, 374
  steepest descent, 374
  tangent action, 326, 374
spectral Wasserstein
  distance, 236, 238, 239, 326, 327
  dynamic distance, 326
  flow, 374, 375, 377
sphere
  c-concave potential, 90
spherical average, 221
spherical cap, 197
spherical geodesic, 26
splitting obstruction, 17, 20, 24
splitting set, 256, 257
squared local action, 322
stable process, 378
static-dynamic equivalence, 327, 334
stationarity, 88, 303, 308, 370, 371
stationary condition, 370
stationary density, 370
stationary point, 292
stationary potential, 178
statistical bias, 200
statistical optimal transport, 183, 253
Stein force, 399
Stein geometry, 327
Stein method, 184
Stein operator, 399
Stein variational gradient descent, 327, 399, 401
step-size normalization, 376
Stiefel manifold, 227
stochastic
  control, 139
  differential equation, 378
  gradient, 98, 119, 391, 398
  gradient descent, 378
  interpolant, 389, 392
  optimization, 98
  order, 276
  particle, 345
Strang splitting, 310
Strassen theorem, 276
strict
  convexity, 24, 27, 121, 126, 128, 137, 160
  positivity, 155, 163
strictly correlated electrons, 258
strong law of large numbers, 183
strong MTW condition, 27
strong upper gradient, 342
structured data, 285
subdifferential, 20, 24, 57, 59
subgaussian measure, 192
subgradient, 269
subgradient inequality, 371
sublinear convergence, 359
subsolution, 180
subspace
  projection, 227
  robust Wasserstein, 239
subspace-sliced Wasserstein, 227
substitutability, 179
successive over-relaxation, 172
sum of squares, 191
sum-of-squares relaxation, 191
sum-product algorithm, 261
supergradient, 269
superposition principle, 318, 327
supersolution, 180
support, 12, 14, 27, 48, 57, 83, 84, 108
SVGD, 399
symmetric plan, 258
synthetic Ricci curvature, 361

T

tail index, 379
Talagrand inequality, 367, 368
tangent action, 322, 372
tangent coordinate, 253
tangent norm, 322
teacher distribution, 344
teacher kernel mean, 344
temporal noise, 389
tensor product coupling, 54
Thompson metric, 214
tightness, 10, 12, 31, 57, 65, 223, 283
time change, 396
time reparametrization, 396
time series
  alignment, 311
time-dependent vector field, 315, 389
token limit, 80, 406
token measure, 79, 389, 406
Toland duality, 296, 405
tomography, 255
topical map, 162
topology
  strong, 70, 72, 115, 116
  weak, 56, 69, 70, 72, 73, 84, 116, 117
total unimodularity, 50
total variation, 11, 13, 63, 70, 73, 110, 111, 115–117, 208, 209, 332
  vector-valued measure, 107
totally positive kernel, 203
trace
  entropy, 308

gauge, 238, 240
norm, 239
partial, 307, 308, 311
partial constraint, 307
trace-class operator, 311
trajectory inference, 51, 52
transfer learning, 51
transformer, 78, 79, 206, 406, 412, 419, 420
translation-invariant kernel, 112, 114
transport
cost, 285
coupling, 311
equation, 412
inequality, 361
map, 4, 17, 22, 32, 41, 83, 273, 391, 395
metric, 377
PDE, 411
representation, 78
transport-entropy inequality, 367
transportation polytope, 41, 42, 45, 50, 52–54, 121, 127, 159, 160, 165, 266, 285
transportation simplex, 45, 52
tree
decomposition, 261
earth mover distance, 31
embedding, 32
geodesic, 32
interpolation, 32
metric, 31
transport, 31
tree-sliced Wasserstein, 32
tree-Wasserstein distance, 31
treewidth, 261
triangle inequality, 6, 19, 23, 39, 64, 65, 106, 165, 189, 192, 213, 222, 239, 286, 288, 290, 330
triangular map, 32
triangular rearrangement, 32, 33
trivial coupling, 390
tropical
barycenter, 89
truncated marginal, 220
Tweedie identity, 393, 396
twist condition, 25, 26, 256, 257
learned bilinear cost, 301
twist on splitting sets, 256, 257
two-layer neural network, 369–371, 377, 416
two-sample statistic, 199
two-sample testing, 114

U

ultrafast diffusion, 102
unbalanced
barycenter, 335
gradient flow, 379
interpolation, 334
JKO step, 379
OT, 127, 142, 207–210, 212, 213, 270, 333, 334
OT dual, 210
Sinkhorn, 144
Sinkhorn convergence, 214
tangent space, 333
uniform integrability, 72, 184
upper hemisphere, 26
upper matrix, 266

V

Varadhan formula, 144
variable elimination, 261
variable projection, 173
variance, 197
variation seminorm, 162, 164, 165, 408
variational
mean field game, 335
variational dual formula, 118
vector quantile, 234
vector-valued measure, 107, 279, 280, 348
velocity covariance, 326
velocity field, 315, 316, 321, 340, 341, 389, 406, 412, 419
Vlasov equation, 386
von Neumann entropy, 308
Voronoi cell, 18, 97, 100, 103

W

Waddington landscape, 342
Waddington-OT, 51
warping path, 311
Wasserstein
barycenter, 142, 235, 243, 244, 246, 250, 252, 253, 259, 260
consistency, 252
population, 252
convergence, 183
coordinate, 234
distance, 30, 63–67, 69–73, 117, 118, 184, 207, 213, 221, 222, 236, 238, 239, 283–285, 287, 290, 317, 321, 325, 356, 404, 420
empirical rate, 188
flow, 356
flow of discrepancy, 398
formula, 29, 30
GAN, 397
Gaussian geometry, 417
geodesic, 22, 67, 236, 244, 290
gradient, 317, 338, 340–342, 344, 346, 356, 369, 370, 384, 389, 399, 401, 402, 412, 413, 418–420
gradient flow, 22, 103, 149, 317, 338, 341–347, 356, 359, 370, 381, 383, 389, 398, 399, 412, 420
Hölder functional, 75
infinity distance, 74
infinity robustness, 74
Kurdyka–Łojasiewicz inequality, 358
Lipschitz functional, 340
Lipschitz map, 78, 80, 81
minimax rate, 189
over Wasserstein, 252, 283, 285, 291
p-action, 325
p-distance, 25
PL inequality, 356–358, 360, 361, 363, 365, 368
polynomial, 76
Procrustes, 231–233, 287, 292
space, 63, 75, 77, 283, 284, 342, 419
space CLT, 253
squared distance, 354
Taylor regularity, 339
topology, 72, 184, 283
tower, 90

training, 369
W1 non-uniqueness, 4, 19, 55
Wasserstein–Fisher–Rao, 334, 379–381
wavelet estimator, 191
weak
C-transform, 274
convergence, 12, 56, 69–72, 111, 113, 114, 117, 183, 184, 222, 223, 239, 283
cost, 274, 275
Kantorovich duality, 274
limit, 17, 34, 70
map, 24
optimal transport, 273, 274
transport, 274, 275
weight clipping, 119
weighted
Laplacian, 366
Poisson equation, 25, 317
sweep, 45
wide stencil, 24
word mover's distance, 67

Z

Z-function, 179
Z-matrix, 180
zero mass, 34, 97, 112, 122
zero mean, 317