# Optimal Transport for Machine Learners

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

June 14, 2026

# Abstract

Modern machine learning repeatedly manipulates probability measures: empirical datasets, generated samples, latent distributions, class-conditional laws, particle systems, weights 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 therefore combines a statistically meaningful notion of discrepancy with a geometry of interpolation, dual certificates and variational dynamics. This makes OT a common language for losses, generative modeling, domain adaptation, robust learning, barycenters, gradient flows and mean-field descriptions of learning algorithms.

This book presents the main OT techniques with these machine-learning uses in mind. It starts from finite assignment and the Monge map viewpoint, passes to Kantorovich couplings and dual potentials, and then explains the algorithmic ideas that make transport usable: linear programming, semi-discrete cells, Sinkhorn scaling and low-dimensional projections. The same objects are then reused as a geometry of measures, giving Wasserstein distances, barycenters, gradient flows, dynamic formulations and Gaussian/Bures formulas. The final chapters emphasize the variants most relevant to modern ML: divergences and adversarial losses, entropic and unbalanced relaxations, robust or spectral ground geometries, Gromov and quantum extensions, and transport-based views of generative models, mean-field networks and attention dynamics. The goal is to keep the mathematics explicit while exposing the computational and geometric intuitions needed to turn OT into a working toolbox for machine learners.

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 [90]. 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.

# Guide to the Literature and Scope

Several books already cover optimal transport from complementary viewpoints. The two-volume monograph of Rachev and Rüschendorf [190, 191] gives a broad probabilistic treatment of mass transportation and its applications. Villani's books [225, 226] are the standard references for the modern mathematical theory, from Kantorovich duality to curvature, concentration and geometric analysis. Santambrogio's text [202] offers a concise applied-mathematics route through the same foundations, with a strong emphasis on PDEs and variational arguments. Ambrosio, Gigli and Savaré [7] develop the metric-space theory of gradient flows that underlies the dynamical part of the subject.

On the computational side, Peyré and Cuturi [185] provide the reference account of numerical OT, entropic regularization and applications in data sciences. Galichon's book [96] explains the economic and matching-theoretic viewpoint, while the statistical theory of OT is developed in the recent lecture notes of Chewi, Niles-Weed and Rigollet [62]. Recent surveys complement these books by emphasizing scalable algorithms and machine-learning applications [131, 168], as well as the role of OT in imaging and graphics [39]. These references remain the natural places to find exhaustive proofs, historical details and specialized variants.

The aim here is different and more selective. The book keeps the core mathematics explicit, but organizes it around the questions that repeatedly arise in machine learning: how to compare singular empirical measures, how to compute differentiable transport losses, how regularization changes optimization and statistics, how dual potentials become discriminators, and how transport geometry produces flows of particles, neurons and tokens. The intended contribution is therefore not a replacement for the references above, but a compact bridge between rigorous OT and the geometric intuitions needed to use it in modern ML.

# Contents

CHAPTER 1

# Optimal Matching between Point Clouds

This opening chapter isolates the simplest form of optimal transport: pairing two finite point clouds. The stakes are algorithmic and geometric at once: one sees the combinatorial nature of transport, the special simplicity of the line, and the first hints that convex relaxation will be necessary in higher dimension. Classical assignment algorithms such as the Hungarian method and auction methods [136, 26] provide the computational backdrop, while the geometric examples prepare 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.

**Matching Problem** Given a cost matrix $(\mathrm{C}_{i,j})_{i\in[\![n]\!],j\in[\![m]\!]}$ and assuming $n = m$, the optimal assignment problem aims to find a bijection $\sigma$ within the set $\mathrm{Perm}(n)$ of permutations of $n$ elements that solves

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

One could naively evaluate the cost function above using all permutations in the set $\mathrm{Perm}(n)$. However, this set has size $n!$, which becomes enormous even for small values of $n$. In general, the optimal $\sigma$ is not unique.

**1D Case** In 1D, for convex cost, the matching defines a monotonic map.

**Proposition 1.1** (Monotone matching on the line)**.** *Assume that the points $(x_i)_i$ and $(y_j)_j$ are pairwise distinct. If the cost is of the form $\mathrm{C}_{i,j} = h(x_i - y_j)$, where $h : \mathbb{R} \to \mathbb{R}^+$ is strictly convex (for example, $\mathrm{C}_{i,j} = |x_i - y_j|^p$ for $p > 1$), then any optimal $\sigma$ defines a strictly increasing map $x_i \mapsto y_{\sigma(i)}$ (and thus is unique), i.e.,*

$$\forall (i,i'), \quad (x_i - x_{i'})(y_{\sigma(i)} - y_{\sigma(i')}) > 0.$$

*Proof.* Indeed, if this property is violated, i.e., there exists $(i,i')$ such that $(x_i - x_{i'})(y_{\sigma(i)} - y_{\sigma(i')}) < 0$, then one can define a permutation $\tilde{\sigma}$ by swapping the match, i.e., $\tilde{\sigma}(i) = \sigma(i')$ and $\tilde{\sigma}(i') = \sigma(i)$, yielding a strictly better cost, as proved in the following fact. Let $h : \mathbb{R} \to \mathbb{R}$ be strictly convex and let $x < x'$ and $y < y'$. Then

$$h(x-y) + h(x'-y') < h(x-y') + h(x'-y).$$

We set the gap $d := y' - y > 0$ and define for every $s \in \mathbb{R}$

$$D(s) := \frac{h(s) - h(s-d)}{d} \quad \text{and} \quad \Delta := h(x-y') + h(x'-y) - h(x-y) - h(x'-y') = d(D(x'-y) - D(x-y)).$$

Because $h$ is strictly convex, $D$ is strictly increasing. Since $x - y < x' - y$, monotonicity yields $D(x-y) < D(x'-y)$, that is $\Delta > 0$. □

This property extends by continuity to convex (not strictly convex) costs such as $|x-y|$, but in this case, the matching is not necessarily unique. For convex costs, the algorithm to compute an optimal transport is therefore to sort the points, i.e., find some pair of permutations $\sigma_X, \sigma_Y$ 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 then an optimal match is mapping $x_{\sigma_X(k)} \mapsto y_{\sigma_Y(k)}$, i.e., an optimal transport is $\sigma = \sigma_Y \circ \sigma_X^{-1}$. The total computational cost is thus $O(n \log(n))$, using, for instance, the quicksort algorithm.

**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 $\sigma$.
**Sort** source and target points: $x_{\sigma_X(1)} \leqslant \cdots \leqslant x_{\sigma_X(n)}$, $y_{\sigma_Y(1)} \leqslant \cdots \leqslant y_{\sigma_Y(n)}$.
**For** $k = 1, \ldots, n$ **do**:
  **Match** $x_{\sigma_X(k)}$ with $y_{\sigma_Y(k)}$.
**Return** $\sigma = \sigma_Y \circ \sigma_X^{-1}$.

If the distance profile is concave instead of convex, the geometry changes. For costs such as $c_p(x,y) = |x-y|^p$ with $0 < p < 1$, splitting a displacement into several smaller displacements is expensive, so optimal matchings tend to create long exchanges rather than the monotone equal-rank pairing; see Figure 1.1. This is the strictly concave regime studied by Gangbo and McCann [99].

The real line still gives special structure. After sorting all red and blue points together, the ordered sequence decomposes into maximal alternating chains, and local matching indicators can certify pairs that must be matched in an optimum. Removing such certified pairs and repeating yields an exact hierarchical algorithm for the unit-mass balanced assignment problem, with worst-case complexity $O(n^2)$ in the framework of Delon, Salomon and Sobolevski [77]. Very concave costs also motivate simpler greedy heuristics, studied for instance by Ottolini and Steinerberger [177]. The point is that these methods are not generic linear-programming solvers; they use the one-dimensional order and the concavity of the distance profile.

The resulting constructive rule is summarized as Algorithm 1.2.

**Algorithm 1.2** Concave line matching by local indicators

**Input:** Unit-mass source and target points on $\mathbb{R}$; strictly concave distance cost.
**Output:** Optimal concave-cost matching $M$.
**Sort** combined red-blue sequence on the line.
**Decompose** it into maximal alternating chains.
**Initialize:** Set $M = \emptyset$.
**While** unmatched points remain **do**:
  **For** each active alternating chain **do**:
    **Compute** local matching indicators [77].
    **If** an indicator certifies a pair $(i, j)$ **then**:
      **Update** $M \leftarrow M \cup \{(i, j)\}$.
      **Remove** points $i$ and $j$.
  **Recompute** only chains affected by removals.
**Return** $M$.

Note that if $\varphi : \mathbb{R} \to \mathbb{R}$ is an increasing map, one can apply this technique to costs of the form $h(|\varphi(x) - \varphi(y)|)$ with a change of variable. A typical application is grayscale histogram equalization. The empirical cumulative distribution of the luminance values of an image is transported to a prescribed target histogram, for instance a concentrated or reference-image histogram. In one dimension, the monotone rearrangement above gives the exact transport map, so the operation is both computationally simple and geometrically faithful: it matches distributions of intensities rather than individual pixels.

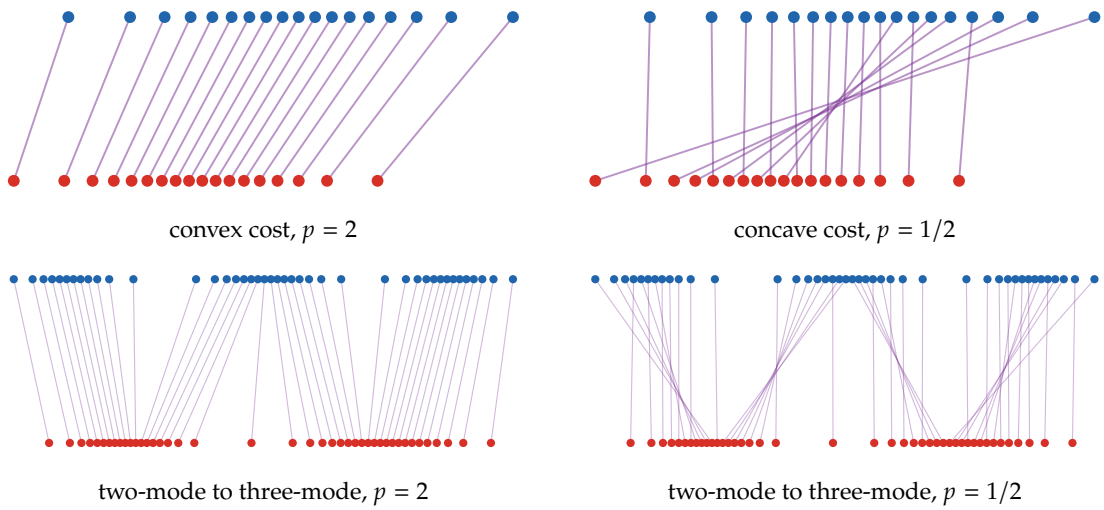



*Figure 1.1:* One-dimensional assignments for ordered source and target clouds with costs $c_p(x, y) = |x - y|^p$. The top row uses single-Gaussian source and target clouds; the bottom row uses a denser two-component source and three-component target. For the convex quadratic cost, equal ranks are matched and the segments do not cross. For the concave cost, the optimum creates long crossing exchanges; the ordered line remains useful, but through the alternating-chain structure of concave transport rather than through monotone rearrangement.

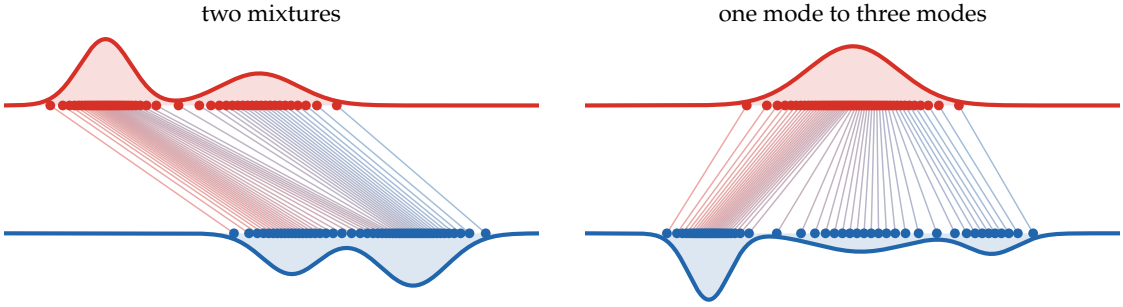



*Figure 1.2:* One-dimensional optimal matching by quantile sorting. The red and blue curves are smooth laws used to generate equal-weight empirical measures; the dots are inverse-CDF samples at common quantile levels. The monotone assignment connects equal ranks, both for two Gaussian mixtures and for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

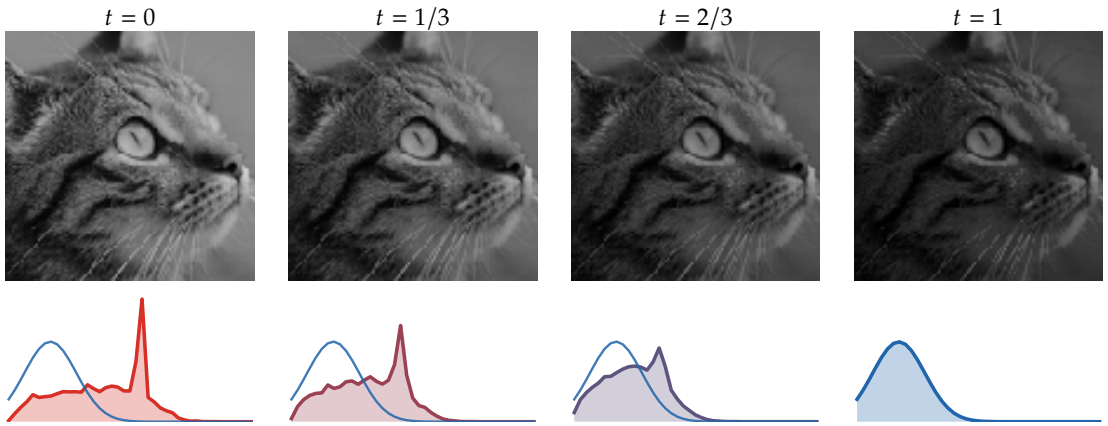



*Figure 1.3:* Histogram equalization as one-dimensional Monge transport on pixel intensities. The map is the monotone rearrangement $T = Q_\beta \circ F_\alpha$, using the cumulative and quantile functions defined later in Definition 2.29; here $\beta$ is a truncated Gaussian concentrated near dark intensities. 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.

Note that if $h$ is strictly convex, then all optimal assignments are increasing, and if the points are all distinct, this increasing map is unique. If $h$ is not strictly convex, for instance $c(x, y) = |x - y|$, non-increasing optimal assignments can also exist. This happens, for example, in the book-shifting problem with overlapping uniform distributions, where the mass in the intersection can stay fixed.

**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 1.2** (Discrete circle transport by a cut)**.** *Let $x_1, \dots, x_n$ and $y_1, \dots, y_n$ be two families of distinct points on $\mathbb{S}^1$, with equal weights. Let $x_{(1)}, \dots, x_{(n)}$ and $y_{(1)}, \dots, y_{(n)}$ denote any fixed cyclic orderings, with the convention $y_{(k+n)} = y_{(k)}$. For the cost $c_p$, $p > 1$, an optimal assignment is one of the cyclic shifts*

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

*and is found by minimizing*

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

*over the $n$ possible shifts. Equivalently, for an optimal shift one can choose a cut $\theta \in \mathbb{S}^1 \setminus (\{x_i\}_i \cup \{y_j\}_j)$ so that, after lifting all points to $(\theta, \theta + 1)$ and sorting them, the optimal matching is the equal-rank monotone matching on this interval.*

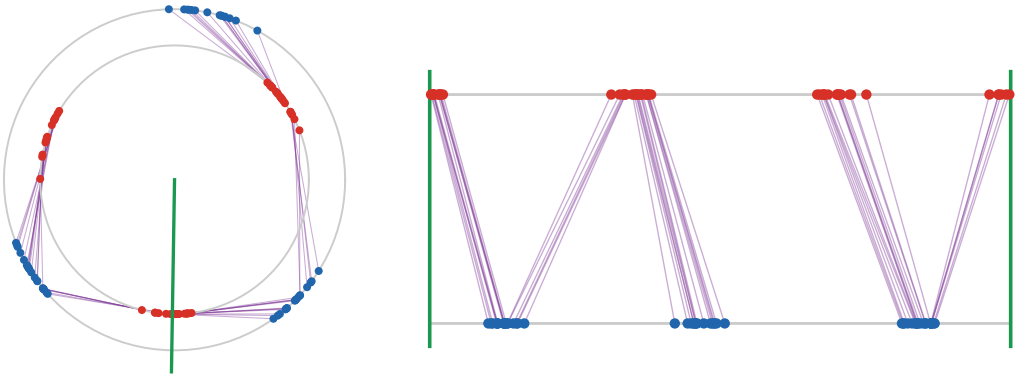
*Proof.* Call two matched pairs cyclically inverted if the circular order of their source endpoints is opposite to the circular order of their target endpoints. Among optimal assignments, choose one with the smallest number of such inversions. The elementary exchange step is the circular analogue of the line argument in Proposition 1.1: if two matched pairs are inverted, then cutting the circle in a gap which does not meet the four endpoints and choosing integer lifts realizes the four geodesic distances involved in the exchange as ordinary distances between two ordered source lifts and two oppositely ordered target lifts. The one-dimensional Monge inequality for the strictly convex function $r \mapsto |r|^p$ then shows that swapping the two targets cannot increase the cost, and decreases it unless the four endpoints are in a degenerate tie configuration.

Thus an optimal assignment can be chosen with no cyclic inversion. A bijection between two finite cyclically ordered sets with no cyclic inversion is a rotation of the order, hence a cyclic shift. This shift specifies how the two cyclic orderings should be opened; after this cut, the rotation becomes an ordinary linear order and the matching is the equal-rank monotone assignment on the unfolded interval. Conversely, each cut gives one such cyclic shift, so minimizing over the finitely many shifts gives an optimal discrete circle assignment. Repeated points or ties are obtained by the same argument after an arbitrarily small perturbation and a limiting passage. This is the discrete form of the fast circle-Monge construction of [76]. □

**Algorithm 1.3** Circle assignment by cutting

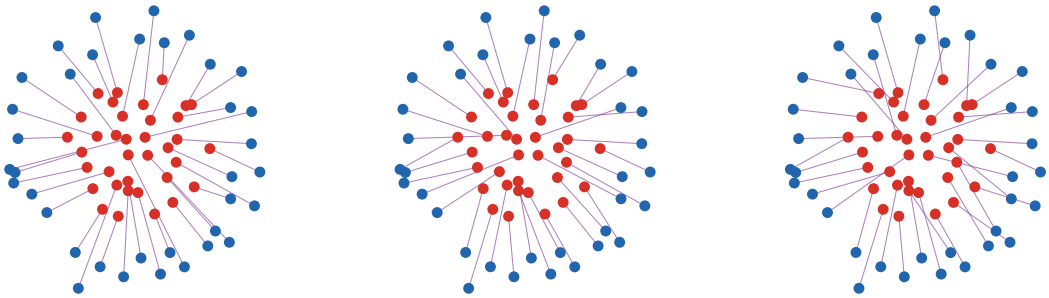
**Input:** Equal-weight points $(x_i)_{i=1}^n$, $(y_j)_{j=1}^n$ on $\mathbb{S}^1$; cost $d_{\mathbb{S}^1}^p$.
**Output:** Optimal cyclic assignment.
**Let** $x_{(1)}, \dots, x_{(n)}$ and $y_{(1)}, \dots, y_{(n)}$ be the points sorted by increasing angle from a fixed origin.
**For** $s = 0, \dots, n-1$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

**Set** $s^\star = \min \operatorname{argmin}_{0\leqslant s<n} E_s$.
**Set** $\theta_{\text{cut}}$ in an empty arc separating two consecutive matched pairs for the shift $s^\star$.
**Replace** every angle by its representative in $[\theta_{\text{cut}}, \theta_{\text{cut}} + 2\pi)$. **Return** $x_{(k)} \mapsto y_{(k+s^\star)}$.



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.



$c(x,y) = \|x-y\|$ $c(x,y) = \|x-y\|^2$ $c(x,y) = \|x-y\|^6$

*Figure 1.5:* Optimal assignments between the same two point clouds for three 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 increasing $p$ penalizes the longest edges more strongly and changes the global organization of the permutation.

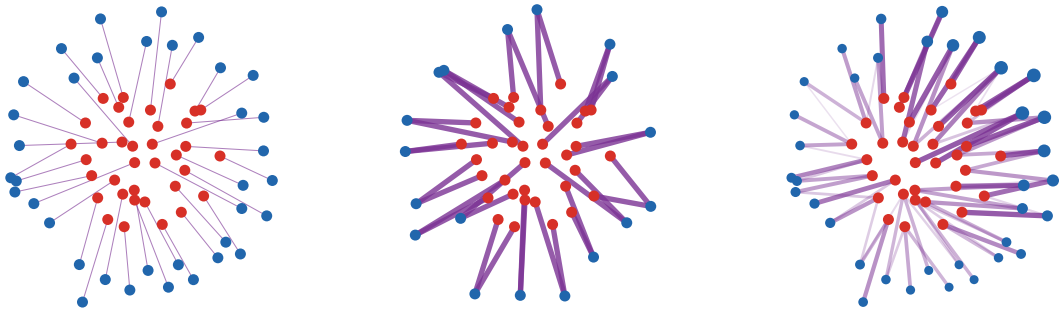
**Rational weights.** The strict assignment model is also tied to equal cardinalities and equal weights. As soon as the target resolution changes or the weights are not uniform, a permutation no longer describes the feasible transports. One instead needs a nonnegative transport matrix with prescribed row and column sums, as illustrated in Figure 1.6; this is the finite-dimensional Kantorovich relaxation developed in Chapter 3.

**Proposition 1.3** (Rational weights as duplicated uniform matching)**.** *Let*

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

*with $k_i, \ell_j \in \mathbb{N}$. The discrete Kantorovich problem between $(\mu, \nu)$ is equivalent to the uniform assignment problem obtained by replacing each $x_i$ by $k_i$ identical copies and each $y_j$ by $\ell_j$ identical copies. More precisely, after multiplying transport masses by $N$, optimal couplings correspond to optimal integer count matrices $(n_{ij})$ with row sums $k_i$ and column sums $\ell_j$, and these count matrices are exactly the collapsed form of assignments between the duplicated clouds.*

*Proof.* Any assignment between the duplicated source and target clouds defines integers $n_{ij}$ counting how many copied particles of type $x_i$ are matched to copied particles of type $y_j$. These counts satisfy



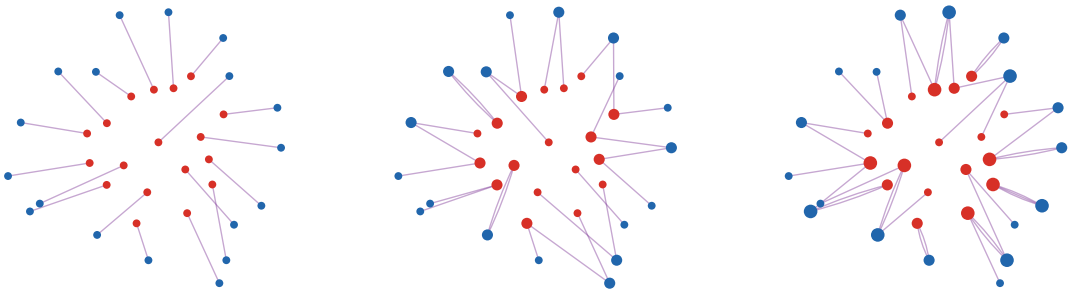
$n = m$ $m < n$ $n = m$ but $b_j \neq 1/n$

*Figure 1.6:* 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.

$\sum_j n_{ij} = k_i$ and $\sum_i n_{ij} = \ell_j$, and the associated coupling $P_{ij} = n_{ij}/N$ has marginals $k_i/N$ and $\ell_j/N$. The assignment cost is

$$\frac{1}{N}\sum_{i,j} n_{ij}c(x_i, y_j) = \sum_{i,j} P_{ij}c(x_i, y_j).$$

Conversely, any nonnegative integer count matrix with those row and column sums can be realized by matching the corresponding duplicated particles. Finally, the transportation constraint matrix is totally unimodular, so the linear problem with integer supplies and demands has an optimal integer count matrix. Thus the optimum of the rational-weight Kantorovich problem is the same as the optimum of the duplicated uniform assignment problem. □



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

*Figure 1.7:* Rational weights as duplicated uniform matchings, using the same disk-to-annulus geometry as Figure 1.6 with fewer displayed atoms. The red and blue locations are kept fixed, while disk areas encode the integer multiplicities $k_i$ and $\ell_j$. Solving the assignment problem after duplicating particles produces several collapsed segments attached to high-multiplicity atoms; this is the integer count matrix of Proposition 1.3.

**2D case.** This efficient strategy to compute the OT in 1-D does not extend to higher dimensions. In 2-D, as already noted by Monge, one has the following property.

**Proposition 1.4** (Non-crossing optimal matchings)**.** *In dimension 2, for $c(x, y) = \|x - y\|$, if $\sigma$ is an optimal assignment, then segments $[x_i, y_{\sigma(i)}]$ cannot cross.*

*Proof.* If two segments $[x_i, y_{\sigma(i)}]$ and $[x_j, y_{\sigma(j)}]$ cross at an interior point $z$, then the triangle inequality gives

$$\|x_i - y_{\sigma(j)}\| + \|x_j - y_{\sigma(i)}\| < \|x_i - y_{\sigma(i)}\| + \|x_j - y_{\sigma(j)}\|.$$

The assignment obtained by swapping $(\sigma(i), \sigma(j))$ therefore has a strictly smaller cost, which contradicts optimality. □

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 1.5 (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

$$M_0 = 1, \qquad M_n = \sum_{k=0}^{n-1} M_k M_{n-1-k}.$$

This recurrence characterizes the Catalan numbers, hence $M_n = C_n$.

Thus even after forbidding crossings, an exhaustive search remains exponential. The two-segment swap in the proof above is nevertheless useful: it explains why a crossing matching cannot be optimal, but it does not select among the exponentially many planar matchings that survive this local test.

## 1.2 Matching Algorithms

This section briefly locates matching within classical combinatorial optimization. Its main point is that efficient algorithms exist, but their cleanest analysis is obtained only after introducing the linear-programming viewpoint.

Efficient algorithms exist to solve the optimal matching problem. The most well-known are the Hungarian method and auction algorithms [136, 25, 26]. Auction algorithms use prices on the target points: each source bids for the target with largest reduced profit, the target price is increased, and the process terminates when the $\varepsilon$-complementary slackness conditions are satisfied. For integer costs, choosing $\varepsilon < 1/n$ gives an exact optimum after a finite number of bids [26]. Section 4.2 revisits this algorithm after Kantorovich duality and explains why it is a dual price method, parallel in spirit to Sinkhorn scaling.

**Hungarian primal-dual method.** The Hungarian method is best understood as a certificate-building algorithm for the assignment linear program. It maintains a partial matching $M$ and dual prices $(u_i, v_j)$ satisfying

$$u_i + v_j \leqslant C_{i,j} \qquad \forall i, j.$$

The equality graph $E(u, v) = \{(i, j) : u_i + v_j = C_{i,j}\}$ contains the edges whose reduced cost is zero. The algorithm only augments $M$ along alternating paths made of equality edges. Starting from an unmatched source, it grows an alternating tree with source set $S$ and target set $T$. If the tree reaches an unmatched target, the matching is augmented along the path. If no such edge exists, the dual variables are shifted by the smallest slack

$$\delta = \min_{i \in S,\ j \notin T} \big(C_{i,j} - u_i - v_j\big), \qquad u_i \leftarrow u_i + \delta\ (i \in S), \qquad v_j \leftarrow v_j - \delta\ (j \in T).$$

This update preserves all inequalities $u_i + v_j \leqslant C_{i,j}$, keeps the current alternating tree tight, and creates at least one new equality edge leaving $S$. Maintaining these slacks incrementally gives the standard $O(n^3)$ implementation for an $n \times n$ assignment problem. Algorithm 1.4 summarizes the primal-dual loop. Figure 1.8 displays actual iterates by showing only the evolving partial assignment: unmatched rows are shown as flat rows to keep a fixed matrix format, and matched rows are shown as one-hot rows.

**Algorithm 1.4** Hungarian primal-dual augmentation

**Input:** Square cost matrix $C \in \mathbb{R}^{n \times n}$.
**Output:** Minimum-cost perfect matching $M$.
**Initialize:** Set $u_i = \min_j C_{ij}$ and $v_j = 0$.
**Set** $M = \emptyset$.
**While** $M$ is not perfect **do**:
  **Build** equality graph: $E(u, v) = \{(i, j) : u_i + v_j = C_{i,j}\}$.
  **Set** root $i_0 = \min\{i : \ i \text{ is unmatched in } M\}$.
  **Set** reached sets $S = \{i_0\}$ and $T = \emptyset$; clear parent pointers.
  **While** $T$ contains no unmatched target **do**:
    **If** $N_E(S) \setminus T = \emptyset$ **then**:
      **Compute** $\delta = \min_{i \in S,\ j \notin T} \big(C_{i,j} - u_i - v_j\big)$.
      **Update** $u_i \leftarrow u_i + \delta$ for $i \in S$ and $v_j \leftarrow v_j - \delta$ for $j \in T$.
      **Refresh** equality graph $E(u, v)$.
    **Set** $J = N_E(S) \setminus T$.
    **For** each $j \in J$ in increasing order **do**:
      **Add** $j$ to $T$ and set parent row $p(j) = \min\{i \in S : (i, j) \in E(u, v)\}$.
      **If** $j$ is matched to $i'$ in $M$ **then set** $S \leftarrow S \cup \{i'\}$ and $q(i') = j$.
  **Set** $j_0 = \min\{j \in T : \ j \text{ is unmatched in } M\}$.
  **Set** $j = j_0$.
  **While** $j$ is defined **do**:
    **Set** $i = p(j)$.
    **Set** $M \leftarrow M \cup \{(i, j)\}$.
    **Set** $j_{\text{old}} = q(i)$.
    **If** $j_{\text{old}}$ is defined **then set** $M \leftarrow M \setminus \{(i, j_{\text{old}})\}$. **Set** $j = j_{\text{old}}$.
**Return** $M$.

**Proposition 1.6** (Correctness and complexity of the Hungarian primal-dual method)**.** *Assume the Hungarian method terminates with a perfect matching $\sigma$ contained in the equality graph*

$$E(u, v) = \big\{(i, j) \ ; \ u_i + v_j = C_{i,j}\big\} \, ,$$

*where $(u, v)$ is dual feasible, i.e. $u_i + v_j \leqslant C_{i,j}$ for all $(i, j)$. Then $\sigma$ is an optimal assignment. Moreover, the usual Hungarian updates terminate after finitely many augmentations. With maintained slacks, the method uses $O(n^3)$ arithmetic operations.*

*Proof.* For any permutation $\tau$, dual feasibility gives

$$\sum_i C_{i,\tau(i)} \geqslant \sum_i (u_i + v_{\tau(i)}) = \sum_i u_i + \sum_j v_j .$$

This is the weak duality lower bound. If $\sigma$ is contained in the equality graph, then

$$\sum_i C_{i,\sigma(i)} = \sum_i u_i + \sum_j v_j ,$$

so the primal cost of $\sigma$ reaches the dual lower bound and is optimal.

It remains to justify finite termination and the complexity bound. At each successful augmentation, the matching cardinality increases by one, so there are at most $n$ augmentations. During one augmentation phase, the algorithm grows an alternating tree in the equality graph. If no augmenting path is available, the dual update uses the smallest slack of an edge leaving the current tree. For edges inside the tree, adding $\delta$ to source labels and subtracting $\delta$ from target labels preserves tightness; for edges from $S$ to $T^c$, the definition of $\delta$ preserves feasibility and makes at least one new edge tight; all other inequalities are unchanged or become looser. Thus the reachable sets strictly grow between two failed augmentation attempts, and they can grow at most $n$ times within one phase. If the current slacks $\min_{i \in S}(C_{i,j} - u_i - v_j)$ are updated when a source enters $S$, each tree expansion costs $O(n)$. A phase therefore costs $O(n^2)$, and the $n$ phases give $O(n^3)$ operations. Hence the method reaches a perfect optimal matching. □

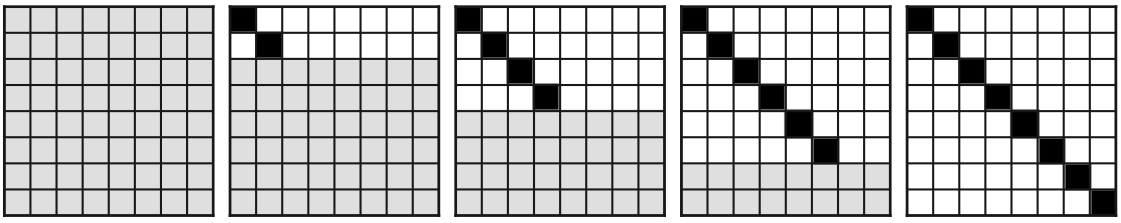



*Figure 1.8:* Matrix view of actual Hungarian primal-dual iterates on a diagonally dominant ordered one-dimensional squared-distance assignment. Each panel records the current partial assignment state: unassigned rows are kept flat, while assigned rows are one-hot. The snapshots are taken at initialization and after two, four, six and eight augmentations; for this pedagogical instance the partial assignments grow along the diagonal, and the final matrix is the identity assignment certified by complementary slackness.

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 [167] and modern treatments [225, 226, 202, 190] are the conceptual background for this transition.

The presentation of the previous chapter could only handle two sets with the same number of points. To relax this to a more general setting, one needs to consider probability distributions, so that the points are weighted by masses.

## 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 $\delta_x$ 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 $\mathrm{a}_i$ are nonnegative.

The Dirac mass should be thought of as a unit of mass infinitely concentrated at one location. An "empirical" probability distribution is uniform on a point cloud, i.e. $\alpha = \frac{1}{n} \sum_i \delta_{x_i}$. In practice, in many applications, it is useful to be able to manipulate both the positions $x_i$ ("Lagrangian" discretization) and the weights $\mathrm{a}_i$ ("Eulerian" discretization). Lagrangian modification is usually more powerful (because it leads to adaptive discretization) but it breaks the convexity of most problems.

**General measures** We consider Borel measures $\alpha \in \mathcal{M}(\mathcal{X})$ on a metric space $(\mathcal{X}, d)$, meaning that $\alpha(A)$ is defined for every Borel set $A$ (obtained from open sets by countable unions, countable intersections and complements). Unless otherwise stated, the measures are finite. A Dirac measure $\delta_x$ 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

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

We denote $\mathcal{M}_+(\mathcal{X})$ the subset of all positive measures on $\mathcal{X}$, i.e. $\alpha(A) \geqslant 0$ (and $\alpha(\mathcal{X}) < +\infty$ for the measure to be finite). The set of probability measures is denoted $\mathcal{M}_+^1(\mathcal{X})$, which means that any $\alpha \in \mathcal{M}_+^1(\mathcal{X})$ is positive, and that $\alpha(\mathcal{X}) = 1$.

**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.

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 3.46.

**Definition 2.4** (Support of a measure)**.** The support $\mathrm{supp}(\alpha)$ of a Borel measure $\alpha$ on a metric space $\mathcal{X}$ is the smallest closed set of full $\alpha$-mass. Equivalently, $x \in \mathrm{supp}(\alpha)$ if every open ball centered at $x$ has positive $\alpha$-mass.

**Radon measures** Using Lebesgue integration, a Borel measure can be used to compute the integral of measurable functions (i.e. such that level sets $\{x \,;\, f(x) < t\}$ are Borel sets), and we denote this pairing as

$$\langle f, \alpha \rangle \coloneqq \int f(x) \mathrm{d}\alpha(x).$$

Integration of such a measurable $f$ against a discrete measure $\alpha$ 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 [197, 35]. 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 2.5** (Relative density)**.** If $\alpha$ is absolutely continuous with respect to a reference measure $\lambda$, its relative density is the Radon–Nikodym derivative

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

Equivalently, for all $h \in C(\mathcal{X})$,

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

On $\mathbb{R}^d$ the reference $\lambda$ 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 2.6** (Total variation)**.** For a finite signed Radon measure $\alpha$ on a compact space $\mathcal{X}$,

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

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

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

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 2.7** (Dual and measure definitions of total variation)**.** *For a finite signed Radon measure $\alpha$ on a compact space $\mathcal{X}$,*

$$\|\alpha\|_{\mathrm{TV}} = |\alpha|(\mathcal{X}).$$

*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 written on the same union of supports, $\alpha = \sum_k a_k \delta_{z_k}$ and $\beta = \sum_k b_k \delta_{z_k}$, with missing coefficients set to zero,

$$\|\alpha - \beta\|_{\mathrm{TV}} = \sum_k |a_k - b_k|.$$

**Probabilistic interpretation.** Radon probability measures represent the laws of random variables. A random variable with values in $\mathcal{X}$ is a measurable map $X : \Omega \to \mathcal{X}$ from an abstract probability space $(\Omega, \mathbb{P})$. Its law is the Radon probability measure $\alpha$ defined by

$$\alpha(A) = \mathbb{P}(\{\omega \in \Omega \,;\, X(\omega) \in A\}) \qquad \text{for Borel sets } A \subset \mathcal{X}.$$

Integrals with respect to this law are expectations:

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

## 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 some continuous map $T : \mathcal{X} \to \mathcal{Y}$, we define the pushforward operator $T_\sharp : \mathcal{M}(\mathcal{X}) \to \mathcal{M}(\mathcal{Y})$. For a Dirac mass, one has $T_\sharp \delta_x = \delta_{T(x)}$, and this formula is extended to arbitrary measures by linearity. In some sense, moving from $T$ to $T_\sharp$ is a way to linearize any map at the price of moving from a (possibly) finite-dimensional space $\mathcal{X}$ to the infinite-dimensional space $\mathcal{M}(\mathcal{X})$, and this idea is central to many convex relaxation methods, most notably Lasserre's relaxation. For discrete measures (2.1), the pushforward operation consists simply in moving the positions of all the points in the support of the measure

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

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 2.8** (Push-forward). For $T : \mathcal{X} \to \mathcal{Y}$, the push forward measure $\beta = T_\sharp \alpha \in \mathcal{M}(\mathcal{Y})$ of some $\alpha \in \mathcal{M}(\mathcal{X})$ satisfies

$$\forall\, h \in C(\mathcal{Y}), \quad \int_{\mathcal{Y}} h(y) \mathrm{d}\beta(y) = \int_{\mathcal{X}} h(T(x)) \mathrm{d}\alpha(x). \tag{2.2}$$

Equivalently, for any measurable set $B \subset \mathcal{Y}$, one has

$$\beta(B) = \alpha(\{x \in \mathcal{X} \;;\; T(x) \in B\}). \tag{2.3}$$

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 2.9 (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:

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

In pairing notation,

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

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 2.10** (Push-forward formula for densities). *Let $\alpha$ and $\beta$ have densities $\rho_\alpha$ and $\rho_\beta$ on open subsets of $\mathbb{R}^d$. Assume that $T$ is a $C^1$ diffeomorphism and that $\beta = T_\sharp \alpha$. Then, for all $x$,*

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

*Equivalently, for $y = T(x)$,*

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

*Proof.* Explicitly doing the change of variable $y = T(x)$, so that $\mathrm{d}y = |\det(T'(x))| \mathrm{d}x$ in formula (2.2) for measures with densities $(\rho_\alpha, \rho_\beta)$ on $\mathbb{R}^d$, one has for all $h \in C(\mathcal{Y})$

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

which shows that

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

Since $T$ is a diffeomorphism, one obtains equivalently

$$\rho_\alpha(x) = |\det(T'(x))|\rho_\beta(T(x))$$

where $T'(x) \in \mathbb{R}^{d\times d}$ is the Jacobian matrix of $T$ (the matrix formed by taking the gradient of each coordinate of $T$). This implies, denoting $y = T(x)$

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

□

**Remark 2.11 (Probabilistic interpretation).** The law, i.e. probability distribution, of a random variable $X$ 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 (2.2), is equivalent to defining another random variable $Y = T(X)$, namely $\omega \in \Omega \mapsto T(X(\omega)) \in \mathcal{Y}$. Thus $\beta$ is the law of $Y$. Drawing a random sample $y$ from $Y$ is thus simply achieved by computing $y = T(x)$ where $x$ is drawn from $X$.

## 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 [167]. 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 section.

**Monge problem.** Given $\alpha \in \mathcal{M}_+^1(\mathcal{X})$, $\beta \in \mathcal{M}_+^1(\mathcal{Y})$ and a cost $c : \mathcal{X}\times\mathcal{Y} \to \mathbb{R}_+$, the Monge problem is

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

The constraint $T_\sharp\alpha = \beta$ means that $T$ pushes the mass of $\alpha$ onto $\beta$ in the sense of Definition 2.8.

**Proposition 2.12** (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 $\beta$, 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 $\alpha$, and in the distinct-target case*

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

*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 $\beta$; 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 2.13** (Existence of transport maps from atomless sources)**.** *Let $\alpha$ and $\beta$ be Borel probability measures on Polish spaces, and assume that $\alpha$ is atomless. Then there exists a measurable map $T$ such that $T_\sharp \alpha = \beta$.*

*Proof.* A standard measure-isomorphism theorem identifies the atomless probability space generated by $\alpha$ with Lebesgue measure on $[0,1]$, modulo null sets [35]. It is therefore enough to construct a map from $[0,1]$ to the target law $\beta$. Choose a Borel isomorphism $i$ from the support of $\beta$ onto a Borel subset of $[0,1]$, set $\nu = i_\sharp \beta$, and use the generalized inverse of the cumulative distribution function of $\nu$. This map sends Lebesgue measure on $[0,1]$ to $\nu$ and takes values in $i(\operatorname{supp}\beta)$ almost surely. Composing with $i^{-1}$ and then with the source isomorphism gives a measurable transport map from $\alpha$ to $\beta$. □

**Remark 2.14 (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 $\beta$ is not a single Dirac mass. Proposition 2.13 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 2.15 (A splitting obstruction).** A classical example, discussed for instance in [202], takes $\alpha$ uniform on a vertical segment and $\beta$ equal to the average of the uniform measures on two parallel vertical segments placed symmetrically to the left and to the right. For the quadratic cost, the relaxed Kantorovich optimizer splits each source point into its two symmetric targets. A deterministic Monge map cannot split one point into two destinations, so minimizing sequences must oscillate between the two sides and the Monge problem has no optimizer.

**Example 2.16 (Semi-discrete Monge maps).** The Monge formulation is not symmetric in $\alpha$ and $\beta$. It makes sense, for instance, when $\alpha$ has a density with respect to Lebesgue measure and $\beta$ 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

$$C_j := T^{-1}(y_j), \qquad \alpha(C_j) = b_j.$$

This is the semi-discrete setting. Chapter 5 explains how the cells become Laguerre cells for prescribed masses and ordinary Voronoi cells when the masses are free. If one exchanges the roles of $\alpha$ and $\beta$ so that $\alpha$ is discrete, then no valid $T$ exists in general: it is not possible to push forward a discrete measure to a measure with density.

Figure 2.1 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 [193, 188]; replacing these projections by a genuine three-dimensional OT map gives a more intrinsic way to match palettes while preserving their geometry [189].

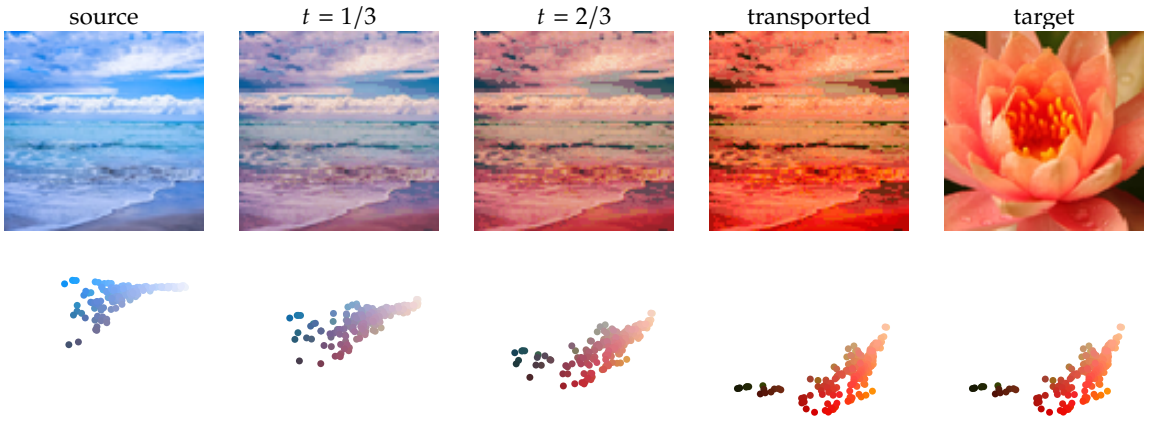



*Figure 2.1:* Color transfer as a Monge map in RGB space, from a beach photograph to a flower photograph. The top row applies the palette map to the source image; the bottom row shows the empirical color clouds in the RGB cube. Only colors are transported here, not pixel locations.

**Monge distance.** When $\mathcal{X} = \mathcal{Y}$ and $c(x, y) = d(x, y)^p$ for a metric $d$, set

$$\mathcal{E}_\alpha(T) \coloneqq \int_{\mathcal{X}} d(x, T(x))^p \mathrm{d}\alpha(x).$$

The Monge value defines the directed quantity

$$\tilde{\mathcal{W}}_p(\alpha, \beta)^p \coloneqq \inf_{T:\mathcal{X}\to\mathcal{X}} \left\{ \mathcal{E}_\alpha(T) \; ; \; T_\sharp \alpha = \beta \right\}. \tag{2.6}$$

If the constraint set is empty, then $\tilde{\mathcal{W}}_p(\alpha, \beta) = +\infty$.

**Proposition 2.17** (Directed Monge distance)**.** *Assume that $\mathcal{X} = \mathcal{Y}$ is a metric space. 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 metric spaces, $\alpha = \beta$.

We prove the triangle inequality. If $\tilde{\mathcal{W}}_p(\alpha, \beta) = +\infty$ while both $\tilde{\mathcal{W}}_p(\alpha, \gamma)$ and $\tilde{\mathcal{W}}_p(\gamma, \beta)$ were finite, there would be maps $S_\sharp \alpha = \gamma$ and $T_\sharp \gamma = \beta$, hence $(T \circ S)_\sharp \alpha = \beta$, a contradiction. Thus the inequality is automatic whenever the left-hand side is infinite. In the finite case, fix $\varepsilon > 0$ and choose $\varepsilon$-minimizers $S_\sharp \alpha = \gamma$ and $T_\sharp \gamma = \beta$ such that

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

The composed map is feasible from $\alpha$ to $\beta$, 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 [43, 44] 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 2.18** (Brenier)**.** *Let $\alpha, \beta \in \mathcal{M}_+^1(\mathbb{R}^d)$ have finite second moments, and assume that $\alpha$ 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 $\alpha$-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 4. Kantorovich duality for the quadratic cost gives optimal potentials $(f, g)$ with equality $f(x) + g(y) = \|x - y\|^2$ on the support of any optimal plan. After subtracting the harmless quadratic terms, this equality is equivalent to the Fenchel equality $\varphi(x) + \varphi^*(y) = \langle x, y\rangle$ for a convex function $\varphi$. Hence the support of every optimal plan lies in the graph of the subdifferential $\partial\varphi$. Since $\alpha$ has a density and convex functions are differentiable Lebesgue-almost everywhere, $\partial\varphi(x)$ is a singleton for $\alpha$-almost every $x$. The plan is therefore concentrated on the graph of $T = \nabla\varphi$, which proves that the relaxed optimizer is induced by a Monge map. Any two optimal plans are concentrated on the same single-valued graph $\alpha$-almost everywhere, which gives uniqueness of the map. This is the standard route behind Brenier's polar factorization theorem; related existence and uniqueness results for more general strictly convex costs are developed for instance in [99, 48, 226]. □

**Definition 2.19** (Measures not charging hypersurfaces)**.** A Borel measure $\alpha$ 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 2.20 (A sharp source hypothesis).** The absolute-continuity assumption in Brenier's theorem can be weakened: for the quadratic cost, it is enough that $\alpha$ does not charge hypersurfaces [99, 226, 202]. 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 $\alpha$-mass, and the optimal relaxed plan may need to split mass, as in Example 2.15.

**Remark 2.21 (Beyond the quadratic Euclidean cost).** Brenier's theorem is the cleanest statement because the squared Euclidean cost turns optimal maps into gradients of convex functions. For costs $c(x, y) = \|x - y\|^p$ with $p > 1$, or more generally costs $c(x, y) = h(x - y)$ with $h$ smooth and strictly convex, the same strategy gives a unique optimal map under absolute continuity of the source, but the map is written in terms of a $c$-convex potential. On a Riemannian manifold, the local analogue for the cost $d_M(x, y)^2/2$ uses the exponential map

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

where $\varphi$ is $c$-convex. The main additional issues are the cut locus and regularity of geodesics, which is why the Euclidean statement is usually presented first. These extensions are part of McCann's displacement convexity theory and the general theory of optimal maps on manifolds [157, 226].

Brenier's theorem should be understood through the analogy between convex gradients and increasing functions. The gradient of a convex function is a monotone field:

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

**Remark 2.22 (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

$$R_\theta = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}.$$

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 $\theta \neq 0$, $R_\theta$ 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 $d \geqslant 2$.

**Polar factorization.** Brenier's theorem does more than solve one transport problem: it provides a canonical way to extract the "monotone part" of an arbitrary 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 $\nu = u_\sharp\lambda$ records where the mass ends up, but it forgets how the points of $\Omega$ were labelled before being sent there. Brenier's polar factorization [43, 44] separates these two effects.

The map first applies a measure-preserving rearrangement $s$ of the source, which changes labels but not mass, and then applies the unique convex-gradient map $\nabla\varphi$ sending 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. This is useful because it isolates the part of a map that carries genuine Wasserstein displacement from the volume-preserving noise of a parametrization. It is also a bridge to fluid mechanics, where measure-preserving maps describe incompressible relabellings, and to data analysis, where one often wants to compare maps modulo such relabellings.

**Proposition 2.23** (Polar factorization)**.** *Let $\Omega \subset \mathbb{R}^d$ be endowed with normalized Lebesgue measure $\lambda$, and let $u \in L^2(\Omega;\mathbb{R}^d)$. Assume that the law $\nu = u_\sharp\lambda$ has finite second moment. Then there exist a measure-preserving map $s : \Omega \to \Omega$ and a convex function $\varphi$ such that*

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

*The Brenier factor $\nabla\varphi$ is unique $\lambda$-almost everywhere.*

*Proof.* By Brenier's theorem there is a unique gradient of a convex function $T = \nabla\varphi$ transporting $\lambda$ to $\nu$. The maps $u$ and $T$ have the same image law. The rearrangement theorem for non-atomic probability spaces gives a measure-preserving map $s$ such that $u = T \circ s$; equivalently, $s$ chooses, with the correct conditional probabilities, preimages of $u(x)$ under $T$. Uniqueness of the Brenier factor follows from Theorem 2.18. □

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

$$A = SO, \qquad O = S^\dagger A,$$

with $O$ orthogonal when $A$ is invertible, or a partial isometry in the singular case. The factor $Sx$ is the convex-gradient transport part, while $Ox$ preserves the standard Gaussian law. This finite-dimensional picture is often the fastest way to remember the general theorem: polar factorization is matrix polar decomposition with matrices replaced by maps and orthogonal transformations replaced by measure-preserving rearrangements.

**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 2.24** (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 [157] clarifies the geometric meaning of interpolating along Brenier maps; this is the language of Wasserstein geodesics used later for barycenters in Section 10.1 and for gradient flows in Chapter 13. 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 3.5. The following proposition makes the constant-speed property precise for the directed Monge value.

**Proposition 2.25** (Directed Monge displacement geodesics)**.** *Let $p \geqslant 1$, let $\alpha, \beta \in \mathcal{M}_+^1(\mathbb{R}^d)$ have finite $p$-th moments, and let $T$ be an optimal map for $\vec{\mathcal{W}}_p(\alpha,\beta)$ 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 $\alpha$-measure Borel set, so that $T_t^{-1}$ is defined $\alpha_t$-almost everywhere. Then, for $0 \leqslant s \leqslant t \leqslant 1$,*

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

*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 2.18.*

*Proof.* The case $s = t$ is trivial, so assume $s < t$. Since $s < 1$, the inverse $T_s^{-1}$ is defined $\alpha_s$-almost everywhere along the transported particles. Define

$$S_{s,t} := T_t \circ T_s^{-1} \qquad \alpha_s\text{-a.e.}$$

Then $(S_{s,t})_\sharp \alpha_s = \alpha_t$, and, using the optimality of $T$,

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

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 $\alpha_t$ to $\beta$ 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 2.17,

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

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 $\varphi$ 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. □

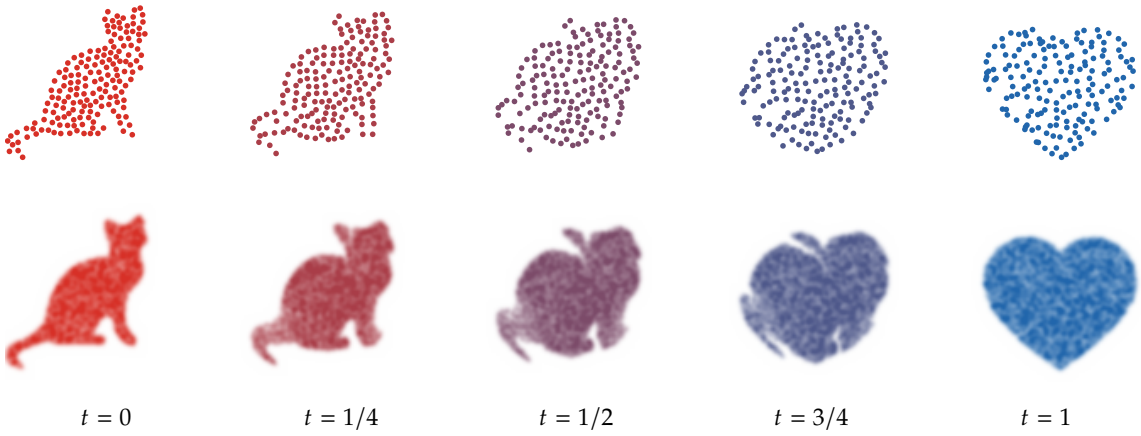



*Figure 2.2:* McCann displacement interpolation between a cat silhouette and a heart silhouette. The first row displays a small farthest-point subset of transported particles along $T_t(x) = (1-t)x + tT(x)$. The second row renders kernel-smoothed densities from a denser transported cloud as color images: 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 2.26** (Caffarelli regularity)**.** *Let $\Omega, \Lambda \subset \mathbb{R}^d$ be bounded uniformly convex domains with $C^2$ boundaries. Let $\alpha = \rho(x)\mathrm{d}x$ be supported on $\Omega$ and $\beta = \eta(y)\mathrm{d}y$ be supported on $\Lambda$, with $0 < m \leqslant \rho, \eta \leqslant M < +\infty$. If $\rho, \eta \in C^\alpha$ for some $\alpha \in (0,1)$, then the Brenier potential $\varphi$ transporting $\alpha$ to $\beta$ is strictly convex and satisfies $\varphi \in C^{2,\alpha}_{\mathrm{loc}}(\Omega)$; in particular $\nabla\varphi$ is locally Hölder. Under the corresponding boundary compatibility and smoothness assumptions, the regularity holds up to the boundary.*

*Proof.* The potential solves the Monge–Ampère equation

$$\det(\nabla^2\varphi(x)) = \frac{\rho(x)}{\eta(\nabla\varphi(x))}$$

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 $\varphi$; the boundary statement follows from the boundary regularity theory under uniform convexity and compatibility assumptions [47, 226]. □

**Remark 2.27 (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 smooth densities, 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 convexity constraint forces $\det(\nabla^2\varphi(x)) \geqslant 0$ and is necessary for this fully nonlinear elliptic equation to be well posed. Numerical schemes for this PDE connect OT with fully nonlinear elliptic solvers; see for instance [21, 94].

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 2.28** (Linearization of the Monge–Ampère equation)**.** *Let $\rho_\varepsilon = \rho_0 + \varepsilon r + o(\varepsilon)$ be a smooth perturbation of a positive reference density $\rho_0$ on a smooth bounded domain, with $\int r = 0$. If $T_\varepsilon(x) = x + \varepsilon\nabla u(x) + o(\varepsilon)$ transports $\rho_0$ to $\rho_\varepsilon$, then, to first order,*

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

*In particular, when $\rho_0$ is constant, the linearized equation is $-\Delta u = r/\rho_0$.*

*Proof.* The change-of-variables equation for $T_\varepsilon$ 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

$$\det(I + \varepsilon\nabla^2 u) = 1 + \varepsilon\Delta u + o(\varepsilon)$$

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 One-Dimensional Transport and Quantiles

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 2.29** (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

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

**Proposition 2.30** (Quantile push-forward)**.** *One has* $(C_\alpha^{-1})_\sharp \mathrm{Leb}_{[0,1]} = \alpha$*. If* $\alpha$ *has no atoms, then* $(C_\alpha)_\sharp \alpha = \mathrm{Leb}_{[0,1]}$.

*Proof.* For simplicity, assume first that $\alpha$ has a strictly positive density, so that $C_\alpha$ 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 $\alpha$ has no atoms, the probability integral transform gives $(C_\alpha)_\sharp \alpha = \mathrm{Leb}_{[0,1]}$. □

If $\alpha$ has no atoms, the map

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

satisfies $T_\sharp \alpha = \beta$.

For the cost $c(x, y) = |x - y|^2$, this map is nondecreasing, hence the derivative of a convex function in dimension one. Brenier's theorem therefore identifies it as the optimal Monge map whenever $\alpha$ is atomless. With generalized inverses, the same quantile construction gives the optimal Kantorovich coupling for arbitrary measures, and it is also optimal for costs of the form $h(|x - y|)$ with $h$ convex and nondecreasing.

**Proposition 2.31** (Monotone rearrangement on the line)**.** *Let* $\alpha, \beta \in \mathcal{M}_+^1(\mathbb{R})$ *have finite* $p$*-th moments, with* $p \geqslant 1$*. The coupling*

$$\pi^\star = (C_\alpha^{-1}, C_\beta^{-1})_\sharp \mathrm{Leb}_{[0,1]}$$

*minimizes* $\int |x - y|^p \mathrm{d}\pi(x, y)$ *among couplings. If* $\alpha$ *has no atoms, it is induced by the monotone Monge map* (2.10).

*Proof.* The displayed measure is a coupling by Proposition 2.30. Its support is monotone: for $s < t$, both quantile functions satisfy $C_\alpha^{-1}(s) \leqslant C_\alpha^{-1}(t)$ and $C_\beta^{-1}(s) \leqslant C_\beta^{-1}(t)$. If a coupling had two crossed pairs $x < x'$ and $y > y'$ with positive mass, exchanging the targets decreases the cost for strictly convex powers and does not increase it for $p = 1$, by the two-point inequality used in Proposition 1.1. Eliminating crossings yields a monotone optimal coupling, and the monotone coupling with prescribed marginals is exactly the quantile coupling. If $\alpha$ has no atoms, $(C_\alpha)_\sharp \alpha = \mathrm{Leb}_{[0,1]}$, so the coupling is generated by (2.10). □

**Remark 2.32 (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 $\alpha$- and $\beta$-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 2.33 (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 $T_1(x) = A_1x$ and $T_2(x) = A_2x$ with $A_1, A_2$ symmetric positive definite, then $T_2 \circ T_1$ has matrix $A_2A_1$. It is a gradient field only when this product is symmetric, equivalently $A_1A_2 = A_2A_1$. 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 1.1.

**Algorithm 2.1** Quantile rearrangement and one-dimensional geodesic

**Input:** One-dimensional probability measures $\alpha, \beta$; time $t \in [0,1]$.
**Output:** Quantile coupling, Monge map when defined, and geodesic point $\alpha_t$.
**Compute** $C_\alpha$, $C_\beta$ and generalized inverses.
**Couple** equal quantile levels: $X = C_\alpha^{-1}(r), \qquad Y = C_\beta^{-1}(r), \qquad r \in (0,1)$.
**If** $\alpha$ has no atoms **then**:
  **Set** $T(x) = C_\beta^{-1}(C_\alpha(x))$.

**Interpolate** quantiles: $Q_t(r) = (1-t)C_\alpha^{-1}(r) + tC_\beta^{-1}(r), \qquad \alpha_t = (Q_t)_\sharp \mathrm{Leb}_{[0,1]}$. **Return** $(X, Y)$, $T$ if defined, and $\alpha_t$.

**Proposition 2.34** (One-dimensional Wasserstein formulas)**.** *Let $\alpha, \beta \in \mathcal{M}_+^1(\mathbb{R})$ have finite $p$-th moments. For every $p \geqslant 1$,*

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

*For $p = 1$, this is equivalently*

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

$$= \int_{\mathbb{R}} \left|\int_{-\infty}^x \mathrm{d}(\alpha - \beta)\right| \mathrm{d}x. \tag{2.13}$$

*Proof.* The first formula follows from Proposition 2.31: the optimal coupling is obtained by taking the same quantile level $r$ for both measures. For $p = 1$, use the layer-cake identity. If $q_\alpha$ and $q_\beta$ 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 $x$. □

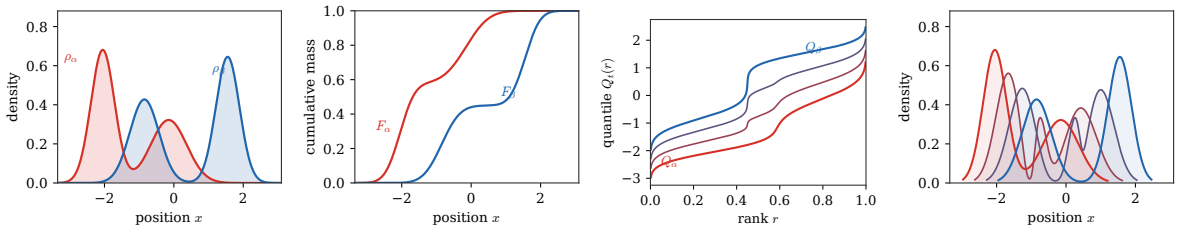
Formula (2.11) means that the map $\alpha \mapsto C_\alpha^{-1}$ embeds one-dimensional Wasserstein geometry isometrically into a linear $L^p$ space. For $p = 2$, the Wasserstein distance on probability measures over the real line is therefore Hilbertian. This makes one-dimensional OT much simpler than higher-dimensional OT, where the Wasserstein geometry is not globally Hilbertian.

The last panel of Figure 2.3 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.

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. Other classical one-dimensional distances are obtained by replacing the $L^1$ norm of cumulative functions with $L^2$ or $L^\infty$ 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.




densities cumulative functions quantiles displacement geodesic

*Figure 2.3:* 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.

**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 [133, 196]. 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 2.35** (Knothe–Rosenblatt triangular rearrangement)**.** *Let $\alpha, \beta \in \mathcal{M}_+^1(\mathbb{R}^d)$. Assume, for simplicity, that the first marginal of $\alpha$ and the one-dimensional conditional laws of $\alpha$ used below are atomless, and that regular conditional distributions are fixed. 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 $\alpha$-almost every value of $(x_1, \dots, x_{k-1})$.*

*Proof.* The construction is recursive. For $k = 1$, let $T_1$ be the monotone rearrangement between the first marginal of $\alpha$ and the first marginal of $\beta$. 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 $\beta$. 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.

**Algorithm 2.2** Knothe–Rosenblatt triangular rearrangement

**Input:** Probability measures $\alpha, \beta$ on $\mathbb{R}^d$ with conditional laws.
**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** $T(x) = (T_1(x_1), T_2(x_1, x_2), \dots, T_d(x_1, \dots, x_d))$.

Figure 2.4 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.

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

*Figure 2.4:* Triangular rearrangement between the same cat and heart densities as in Figure 2.2. 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 [52].

**Proposition 2.36** (Anisotropic Brenier maps converge to Knothe–Rosenblatt)**.** *Let $\alpha, \beta$ be compactly supported probability measures on $\mathbb{R}^d$ with densities bounded above and below on their rectangular supports, and assume that the conditional laws entering Proposition 2.35 are atomless. For $\varepsilon > 0$, set*

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

*Let $T_\varepsilon$ be the Monge map from $\alpha$ to $\beta$ for the cost $c_\varepsilon$, and let $T_{\mathrm{KR}}$ 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 give the standard lexicographic argument, which is the variational core of [52]. Let $\pi_\varepsilon = (\mathrm{Id}, T_\varepsilon)_\sharp \alpha$. Since the supports are compact, a subsequence converges weakly to some coupling $\pi$ between $\alpha$ and $\beta$. The optimality of $\pi_\varepsilon$ for

$$F_\varepsilon(\gamma) = \sum_{k=1}^{d} \varepsilon^{k-1} \int |x_k - y_k|^2 \, \mathrm{d}\gamma(x, y)$$

first implies, by letting $\varepsilon \to 0$, that $\pi$ minimizes the one-dimensional quadratic cost in the first coordinate among all couplings. Since the first marginal of $\alpha$ is atomless, the one-dimensional minimizer is the monotone rearrangement, so $y_1 = T_1(x_1)$ under $\pi$.

Now restrict attention to couplings satisfying this first-coordinate constraint. Subtract the common minimal value of the first-coordinate cost, divide the optimality inequality by $\varepsilon$, and let $\varepsilon \to 0$. The limit coupling must minimize the second-coordinate quadratic cost among all couplings that already realize the first monotone rearrangement. Disintegrating with respect to $(x_1, y_1)$ reduces this constrained problem to the one-dimensional monotone rearrangement between the conditional laws of $x_2$ and $y_2$. Hence $y_2 = T_2(x_1, x_2)$ under $\pi$.

Repeating the same subtraction-and-rescaling argument gives, for every $k$, the conditional monotone rearrangement $y_k = T_k(x_1, \dots, x_k)$. Thus every weak limit of $(\pi_\varepsilon)_\varepsilon$ is concentrated on the graph of $T_{\mathrm{KR}}$. This graph coupling is unique, so the whole family $\pi_\varepsilon$ converges weakly to $(\mathrm{Id}, T_{\mathrm{KR}})_\sharp \alpha$.

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 $T_{\mathrm{KR}}$ 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 $\alpha$-probability. Since all maps take values in a common compact set, this convergence is uniformly integrable and hence holds in $L^2(\alpha)$. □

## 2.6 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

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

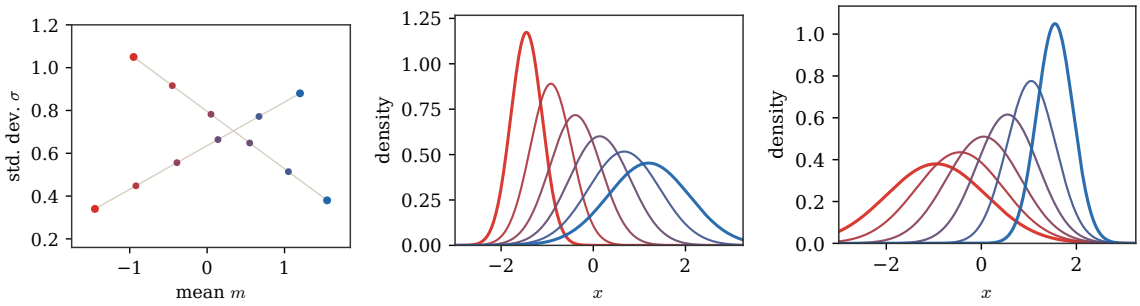
satisfies $T_\sharp \alpha = \beta$. It is the derivative of the convex function

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

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 half-plane $(m, \sigma) \in \mathbb{R} \times \mathbb{R}_+$. This contrasts with KL geometry, where singular Gaussians are infinitely distant.




$(m, \sigma)$ half-plane first geodesic second geodesic

*Figure 2.5:* One-dimensional Gaussian $\mathcal{W}_2$ geodesics. In the coordinates $(m, \sigma)$, the geodesics are Euclidean segments in the upper half-plane. The two density panels show the corresponding Gaussian densities along the two segments, with colors interpolating from the red endpoint to the blue endpoint.

**Multivariate Gaussians.** If

$$\alpha = \mathcal{N}(\mathbf{m}_\alpha, \mathbf{\Sigma}_\alpha), \qquad \beta = \mathcal{N}(\mathbf{m}_\beta, \mathbf{\Sigma}_\beta), \qquad T(x) = \mathbf{m}_\beta + A(x - \mathbf{m}_\alpha), \tag{2.14}$$

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

**Proposition 2.37** (Affine push-forward of Gaussians)**.** *One has $T_\sharp \alpha = \beta$ if and only if*

$$A\mathbf{\Sigma}_\alpha A^\top = \mathbf{\Sigma}_\beta. \tag{2.15}$$

*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

$$\mathbb{E}(Y) = \mathbf{m}_\beta + A\mathbb{E}(X - \mathbf{m}_\alpha) = \mathbf{m}_\beta,$$

and

$$\mathbb{E}((Y - \mathbf{m}_\beta)(Y - \mathbf{m}_\beta)^\top) = A\mathbb{E}((X - \mathbf{m}_\alpha)(X - \mathbf{m}_\alpha)^\top)A^\top = A\mathbf{\Sigma}_\alpha A^\top.$$

Thus $A\mathbf{\Sigma}_\alpha A^\top = \mathbf{\Sigma}_\beta$ is necessary and sufficient for $Y$ to have the same mean and covariance as $\beta$. □

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

**Definition 2.38** (Bures metric)**.** For positive semidefinite covariance matrices $\Sigma$ and $\Lambda$, the Bures metric is

$$\mathcal{B}(\Sigma,\Lambda)^2 := \operatorname{tr}\left(\Sigma+\Lambda-2(\Sigma^{1/2}\Lambda\Sigma^{1/2})^{1/2}\right). \tag{2.16}$$

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

**Proposition 2.39** (Gaussian $\mathcal{W}_2$ formula and Bures covariance term)**.** *Assume that $\mathbf{\Sigma}_\alpha$ and $\mathbf{\Sigma}_\beta$ are positive definite. The unique symmetric positive-definite solution of*

$$A\mathbf{\Sigma}_\alpha A = \mathbf{\Sigma}_\beta$$

*is*

$$A = \mathbf{\Sigma}_\alpha^{-1/2}\left(\mathbf{\Sigma}_\alpha^{1/2}\mathbf{\Sigma}_\beta\mathbf{\Sigma}_\alpha^{1/2}\right)^{1/2}\mathbf{\Sigma}_\alpha^{-1/2}. \tag{2.17}$$

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

$$\mathcal{W}_2(\alpha,\beta)^2 = \|\mathbf{m}_\alpha-\mathbf{m}_\beta\|^2+\mathcal{B}(\mathbf{\Sigma}_\alpha,\mathbf{\Sigma}_\beta)^2, \tag{2.18}$$

*where $\mathcal{B}$ is the Bures metric of Definition 2.38.*

*Proof.* Multiplying $A\mathbf{\Sigma}_\alpha A=\mathbf{\Sigma}_\beta$ on the left and right by $\mathbf{\Sigma}_\alpha^{1/2}$ gives

$$(\mathbf{\Sigma}_\alpha^{1/2}A\mathbf{\Sigma}_\alpha^{1/2})^2 = \mathbf{\Sigma}_\alpha^{1/2}\mathbf{\Sigma}_\beta\mathbf{\Sigma}_\alpha^{1/2}.$$

Since $A$ is symmetric positive, $\mathbf{\Sigma}_\alpha^{1/2}A\mathbf{\Sigma}_\alpha^{1/2}$ is symmetric positive and is therefore the unique positive square root of the right-hand side. Conversely, the matrix in (2.17) is symmetric positive and satisfies the covariance equation.

By Proposition 2.37, this affine map pushes $\alpha$ to $\beta$. It is the gradient of a convex quadratic potential, so Brenier's theorem implies optimality. If $X\sim\alpha$, then

$$\begin{aligned}
\mathbb{E}\|X-T(X)\|^2 &= \|\mathbf{m}_\alpha-\mathbf{m}_\beta\|^2+\mathbb{E}\|(I-A)(X-\mathbf{m}_\alpha)\|^2\\
&= \|\mathbf{m}_\alpha-\mathbf{m}_\beta\|^2+\operatorname{tr}((I-A)\mathbf{\Sigma}_\alpha(I-A)^\top)\\
&= \|\mathbf{m}_\alpha-\mathbf{m}_\beta\|^2+\operatorname{tr}(\mathbf{\Sigma}_\alpha)+\operatorname{tr}(A\mathbf{\Sigma}_\alpha A)-2\operatorname{tr}(A\mathbf{\Sigma}_\alpha)\\
&= \|\mathbf{m}_\alpha-\mathbf{m}_\beta\|^2+\operatorname{tr}(\mathbf{\Sigma}_\alpha)+\operatorname{tr}(\mathbf{\Sigma}_\beta)-2\operatorname{tr}((\mathbf{\Sigma}_\alpha^{1/2}\mathbf{\Sigma}_\beta\mathbf{\Sigma}_\alpha^{1/2})^{1/2}),
\end{aligned}$$

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

The covariance term $\mathcal{B}$ is the Bures–Wasserstein metric on positive semidefinite matrices [46, 102, 29]. It separates the Euclidean displacement of the mean from the intrinsic transport geometry of covariance ellipsoids.

**Proposition 2.40** (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 $t\in[0,1]$,*

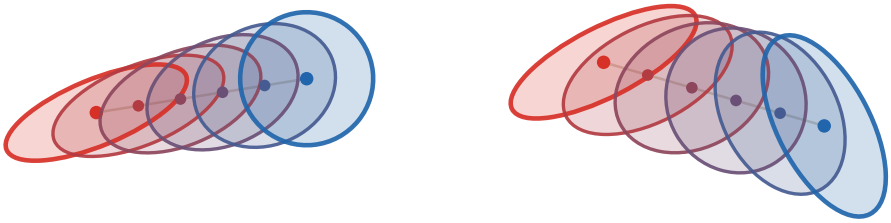
$$\mathcal{B}^2((1-t)\Sigma_0+t\Sigma_1,(1-t)\Lambda_0+t\Lambda_1)\leqslant(1-t)\mathcal{B}^2(\Sigma_0,\Lambda_0)+t\mathcal{B}^2(\Sigma_1,\Lambda_1).$$

*Proof.* The key identity is the Procrustes representation

$$\mathcal{B}^2(\Sigma,\Lambda)=\min_{Q^\top Q=I}\|\Sigma^{1/2}-\Lambda^{1/2}Q\|_F^2.$$

Indeed, expanding the square gives $\operatorname{tr}\Sigma+\operatorname{tr}\Lambda-2\max_{Q^\top Q=I}\operatorname{tr}(\Sigma^{1/2}Q^\top\Lambda^{1/2})$, and the orthogonal Procrustes formula identifies the maximum with $\operatorname{tr}((\Sigma^{1/2}\Lambda\Sigma^{1/2})^{1/2})$. Symmetry, positivity and separation follow immediately from this representation. The triangle inequality follows by choosing two almost optimal orthogonal matrices $Q_1,Q_2$ and writing

$$\|\Sigma^{1/2}-\Gamma^{1/2}Q_2Q_1\|_F\leqslant\|\Sigma^{1/2}-\Lambda^{1/2}Q_1\|_F+\|\Lambda^{1/2}-\Gamma^{1/2}Q_2\|_F.$$



anisotropic to isotropic rotated anisotropies

*Figure 2.6:* Two-dimensional Gaussian $\mathcal{W}_2$ geodesics. Means move linearly, while covariance ellipses follow the Bures–Wasserstein interpolation. The left panel contracts an anisotropic Gaussian toward an isotropic one; the right panel interpolates between two strongly oriented anisotropic covariances.

Letting the two choices become optimal proves the metric property.

For convexity, use the equivalent factor formulation

$$\mathcal{B}^2(\Sigma, \Lambda) = \min_{UU^\top = \Sigma,\, VV^\top = \Lambda} \|U - V\|_F^2.$$

Choose nearly optimal factors $(U_0, V_0)$ and $(U_1, V_1)$ for the two pairs, and define block factors

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

Then $U_t U_t^\top = (1-t)\Sigma_0 + t\Sigma_1$ and $V_t V_t^\top = (1-t)\Lambda_0 + t\Lambda_1$, while

$$\|U_t - V_t\|_F^2 = (1-t)\|U_0 - V_0\|_F^2 + t\|U_1 - V_1\|_F^2.$$

Taking the infimum over the initial factors proves joint convexity. □

> **Remark 2.41 (Diagonal covariances and Hellinger geometry).** If $\mathbf{\Sigma}_\alpha = \operatorname{diag}(r_i)_i$ and $\mathbf{\Sigma}_\beta = \operatorname{diag}(s_i)_i$ are diagonal, the Bures metric reduces to the Euclidean distance between square roots,
>
> $$\mathcal{B}(\mathbf{\Sigma}_\alpha, \mathbf{\Sigma}_\beta) = \|\sqrt{r} - \sqrt{s}\|_2.$$
>
> This is the finite-dimensional Hellinger geometry on the nonnegative covariance coordinates: variances are compared after the amplitude change of variables $r_i \mapsto \sqrt{r_i}$.

CHAPTER 3

# Kantorovich Relaxation

Kantorovich's relaxation is the decisive move that turns transport into convex optimization. This chapter explains how deterministic maps are replaced by couplings, why this fixes infeasibility and symmetry, and how it produces the Wasserstein distances. Historically, this linear-programming viewpoint grew from Kantorovich's economic planning work [128] and is now the standard foundation of OT [225, 226, 191].

## 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.

Monge's discrete matching problem is problematic because it cannot be applied when $n \neq m$. A faithful formulation must keep track of masses $(a_i, b_j)$. 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 $\alpha$ and $\beta$: 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 [128] 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 $\sigma$ or a map $T$, by a coupling matrix $P \in \mathbb{R}_+^{n \times m}$, where $P_{i,j}$ describes the amount of mass flowing from $x_i$ to $y_j$ in the formalism of discrete measures

$$\alpha = \sum_i a_i \delta_{x_i}, \qquad \beta = \sum_j b_j \delta_{y_j}.$$

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

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

Equivalently,

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

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 3.2** (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 3.3** (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

$$R\mathbb{1}_m = (1 + t)a - ta = a, \qquad R^\top \mathbb{1}_n = (1 + t)b - tb = b,$$

so $R \in U(a, b)$. Moreover,

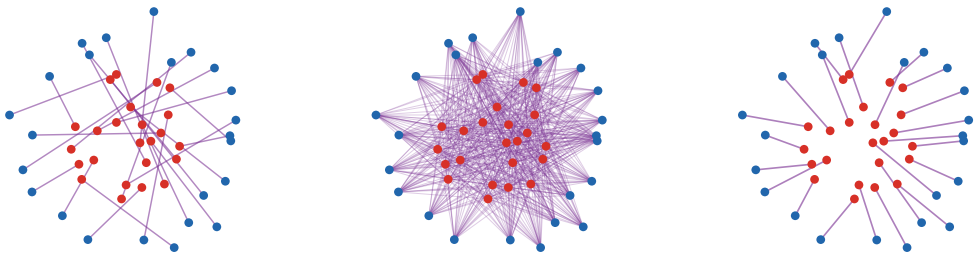
$$a \otimes b = \frac{1}{1 + t} R + \frac{t}{1 + t} Q.$$

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 3.4 shows below that sparse optimal plans with at most $n + m - 1$ positive entries always exist.



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.

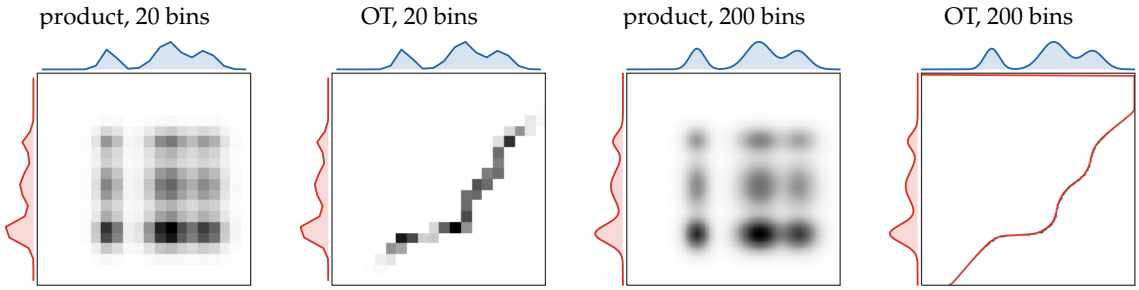



*Figure 3.2:* Coupling matrices with their prescribed marginals. The central grayscale image displays $P_{ij}$; the red curve on the left is the source marginal a, and the blue curve on top is the target marginal b. The independent product plan is diffuse, whereas the one-dimensional optimal plan concentrates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

Kantorovich, aiming for economic planning, made a strong simplifying assumption: the cost of transportation should be linear in the amount of transported mass. Under this assumption, denoting $C_{i,j}$ the cost of moving a unit amount of mass from $x_i$ to $y_j$, the discrete Kantorovich problem reads

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

This is a linear program, and its solutions need not be unique.

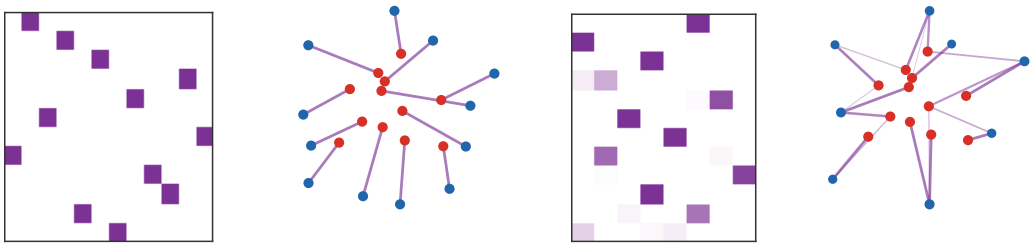



*Figure 3.3:* From permutation matrices to splitting couplings. When the two empirical measures have the same number of atoms and uniform weights, an optimal plan can be a permutation matrix. Once target masses are nonuniform, the admissible set $U(a, b)$ also contains plans in which one source sends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

**Proposition 3.4** (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 compact, so a linear objective attains its minimum at an extreme point. The row and column marginal equations have rank $n + m - 1$: the only linear redundancy is that both sets of constraints impose the same total mass. The support-graph argument below is the combinatorial form of this basic-feasible-variable count.

Let P be an extreme point and let $E = \{(i, j) : P_{ij} > 0\}$ be its support graph on the bipartite vertex set $\{1, \dots, n\} \cup \{1, \dots, m\}$. If this graph contains a cycle, orient the cycle and put alternating signs $+1, -1$ on its edges, obtaining a nonzero matrix $H$ supported on $E$ with zero row and column sums. For sufficiently small $t > 0$, both $P + tH$ and $P - tH$ are nonnegative couplings, and P is their midpoint, contradicting extremality. Thus the support graph is a forest. Since a forest on $n + m$ vertices has at most $n + m - 1$ edges, the claim follows. □

**Proposition 3.5** (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* $r_i = \mathrm{a}_i$ *and* $s_j = \mathrm{b}_j$*, skip zero residuals, set*

$$\mathrm{P}_{ij} = \min(r_i, s_j),$$

*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 $\sum_i \mathrm{a}_i = \sum_j \mathrm{b}_j$, 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 $\mathrm{a} \in \Sigma_n$ and target weights $\mathrm{b} \in \Sigma_m$.
**Output:** Sparse feasible coupling $\mathrm{P} \in \mathrm{U}(\mathrm{a}, \mathrm{b})$.
**Initialize:** Set $\mathrm{P} = 0$, $r = \mathrm{a}$, $s = \mathrm{b}$, and $(i, j) = (1, 1)$.
**While** $i \leqslant n$ and $j \leqslant m$ **do**:
  $\eta = \min(r_i, s_j)$, $\quad \mathrm{P}_{ij} \leftarrow \eta$.
  **Update residuals:** $r_i \leftarrow r_i - \eta$, $\quad s_j \leftarrow s_j - \eta$.
  **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 3.6** (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. Conversely, any feasible plan with a crossing pair of positive entries can be improved by moving a small mass $\eta$ from $(i, j)$ and $(i', j')$ to $(i, j')$ and $(i', j)$. The two marginals are unchanged, and convexity of $h$ gives

$$h(x_i - y_j) + h(x_{i'} - y_{j'}) \geqslant h(x_i - y_{j'}) + h(x_{i'} - y_j)$$

for $i < i'$ and $j' < j$, with strict inequality for strictly convex $h$ and distinct points. Repeating this uncrossing procedure until no crossing remains yields a monotone optimal plan. There is only one monotone feasible plan with the prescribed sorted marginals, namely the sweep plan: it pairs the leftmost remaining source mass with the leftmost remaining target mass at every step. Sorting costs $O(n \log n + m \log m)$ and the sweep uses at most $n + m - 1$ assignments. □

**Algorithm 3.2** Weighted one-dimensional sweep

**Input:** One-dimensional atoms $(x_i, a_i)$ and $(y_j, b_j)$; convex cost $h(x-y)$.
**Output:** Monotone optimal coupling P.
**Sort** atoms: $x_1 \leqslant \cdots \leqslant x_n$, $\quad y_1 \leqslant \cdots \leqslant y_m$.
**Set** P to the output of Algorithm 3.1 applied to the sorted weights $(a_i)_i$ and $(b_j)_j$. **Return** P.

**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 3.7** (Permutation matrices)**.** For a permutation $\sigma \in \mathrm{Perm}(n)$, its permutation matrix $P_\sigma$ 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 $P_\sigma/n$. 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 3.8** (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 $\mathrm{U}(\mathbb{1}_n/n, \mathbb{1}_n/n) = \mathcal{B}_n/n$, 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 [31, 228], which explains why this relaxation is tight for uniform matching.

**Definition 3.9** (Extreme points)**.** For a compact 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 3.10** (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 3.11 (Unbounded convex sets may have no extreme point).** Compactness cannot be dropped from Proposition 3.10. For instance, the closed convex set $\{(x, y) \in \mathbb{R}^2_+ \,;\, xy \geqslant 1\}$ is unbounded and has no extreme point.

**Proposition 3.12** (Linear programs have extreme minimizers)**.** *Let $C$ be non-empty and compact. For every linear form $\ell$,*

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

*Proof.* The set $S = \operatorname{argmin}_{x \in C} \ell(x)$ is non-empty, compact and convex. By Proposition 3.10, 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 $\ell$ on $C$, hence $y, z \in S$. Since $x$ is extreme in $S$, $y = z = x$. Thus $x$ is extreme in $C$. □

**Theorem 3.13** (Birkhoff–von Neumann)**.** *The extreme points of $\mathcal{B}_n$ are exactly the permutation matrices.*

*Proof.* We first prove that permutation matrices are extreme. Let $P_\sigma \in \mathcal{P}_n^{\mathrm{perm}}$ and assume that

$$P_\sigma = \frac{Q + R}{2} \qquad \text{with} \qquad Q, R \in \mathcal{B}_n.$$

Every bistochastic matrix has entries in $[0, 1]$. Since the only extreme points of $[0, 1]$ are 0 and 1, each entry of $P_\sigma$ fixes the corresponding entries of $Q$ and $R$: if $(P_\sigma)_{ij} = 0$, then $Q_{ij} = R_{ij} = 0$, while if $(P_\sigma)_{ij} = 1$, then $Q_{ij} = R_{ij} = 1$. Hence $Q = R = P_\sigma$, so $P_\sigma$ is extreme.

We now prove the converse by contrapositive. Pick $P \in \mathcal{B}_n \setminus \mathcal{P}_n^{\mathrm{perm}}$. Since an integral bistochastic matrix is necessarily a permutation matrix, $P$ has at least one fractional entry. We shall split

$$P = \frac{Q + R}{2}$$

with $Q, R \in \mathcal{B}_n$ and $Q \neq R$, 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

$$0 < P_{ij} < 1.$$

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 $P_{i,j_1} \in (0, 1)$, a positive amount $1 - P_{i,j_1}$ 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

$$(i_1, j_1, i_2, j_2, \ldots, i_p, j_p), \qquad i_{p+1} = i_1,$$

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,

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

Define

$$\varepsilon \coloneqq \min_{1 \leqslant s \leqslant p} \left\{ P_{i_s, j_s},\, P_{i_{s+1}, j_s},\, 1 - P_{i_s, j_s},\, 1 - P_{i_{s+1}, j_s} \right\}.$$

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

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

We now perform the standard alternating-cycle perturbation:

$$Q_{ij} \coloneqq \begin{cases} P_{ij}, & (i,j) \notin A \cup B, \\ P_{ij} + \varepsilon/2, & (i,j) \in A, \\ P_{ij} - \varepsilon/2, & (i,j) \in B, \end{cases} \qquad R_{ij} \coloneqq \begin{cases} P_{ij}, & (i,j) \notin A \cup B, \\ P_{ij} - \varepsilon/2, & (i,j) \in A, \\ P_{ij} + \varepsilon/2, & (i,j) \in B. \end{cases}$$

By the definition of $\varepsilon$, 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

$$Q\mathbb{1}_n = R\mathbb{1}_n = \mathbb{1}_n, \qquad Q^\top \mathbb{1}_n = R^\top \mathbb{1}_n = \mathbb{1}_n,$$

so $Q, R \in \mathcal{B}_n$. Finally, $Q \neq R$ because $\varepsilon > 0$ and the cycle is non-empty, while by construction $P = (Q+R)/2$. Thus $P$ is not extreme. Consequently every extreme point of $\mathcal{B}_n$ is integral, and every integral bistochastic matrix is a permutation matrix. □

The same combinatorial idea gives the constructive decomposition used to express a bistochastic matrix as a convex combination of permutations.

**Algorithm 3.3** Birkhoff–von Neumann decomposition

**Input:** Bistochastic matrix $P \in \mathcal{B}_n$.
**Output:** Decomposition $P = \sum_r \lambda_r P_{\sigma_r}$.
**Initialize:** Set $R = P$ and $\mathcal{L} = \emptyset$.
**While** $R \neq 0$ **do**:
  **Build** bipartite graph $G_R = \{(i,j) : R_{ij} > 0\}$.
  **Set** $\sigma$ to the lexicographically first perfect matching of $G_R$.
  **Set** $\lambda = \min_i R_{i,\sigma(i)}$.
  **Append** $(\lambda, \sigma)$ to $\mathcal{L}$.
  **Update** $R \leftarrow R - \lambda P_\sigma$.
**Return** $P = \sum_{(\lambda_r,\sigma_r)\in\mathcal{L}} \lambda_r P_{\sigma_r}, \qquad \sum_r \lambda_r = 1.$

**Corollary 3.14** (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 $P_\sigma/n$. The associated permutation $\sigma$ solves the assignment problem of Section 1.1.*

*Proof.* The feasible set is $\mathcal{B}_n/n$. By Proposition 3.12, the linear objective has an optimal extreme point. Since scaling preserves extreme points and Theorem 3.13 identifies the extreme points of $\mathcal{B}_n$, this optimizer is $P_\sigma/n$ for some permutation $\sigma$. Its cost is exactly $n^{-1}\sum_i C_{i,\sigma(i)}$, so $\sigma$ 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 3.15 (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.

## 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 [75]. 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 [27]. 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 [175]. 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. For the transport polytope, the logarithmic-barrier version is

$$\mathrm{P}_\varepsilon \coloneqq \operatorname*{argmin}_{\substack{\mathrm{P}\mathbb{1}_m=\mathrm{a},\ \mathrm{P}^\top\mathbb{1}_n=\mathrm{b}\\ \mathrm{P}_{ij}>0}} \langle \mathrm{C},\, \mathrm{P}\rangle - \varepsilon \sum_{i,j} \log \mathrm{P}_{ij}, \tag{3.3}$$

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 $\operatorname{diag}(\varepsilon/\mathrm{P}_{ij}^2)$, which gives robust polynomial complexity but can be expensive for dense couplings [172].

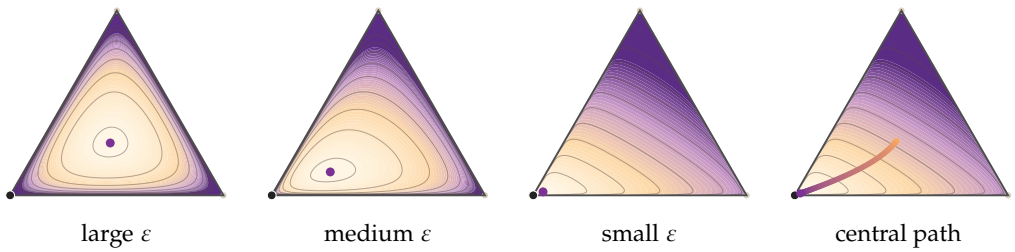



*Figure 3.4:* Logarithmic-barrier central path for a two-dimensional equilateral triangular slice of a linear program. The feasible triangle is constrained by positive slacks, and the displayed objective is $\ell^\top z - \varepsilon \sum_i \log(b_i - \langle a_i, z\rangle)$. Large $\varepsilon$ selects a central interior point; decreasing $\varepsilon$ moves the minimizer toward the optimal vertex while never touching the boundary. This should be contrasted with entropic OT, where the entropy temperature $\varepsilon$ 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 $\varepsilon$ and following a central path. Sinkhorn fixes an entropic temperature $\varepsilon$ and solves a different, KL-regularized OT problem by alternating diagonal scalings. The parameter $\varepsilon$ 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 3.16** (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 $\pi$ 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 3.16, 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 3.17** (Couplings)**.** Given $\alpha \in \mathcal{M}_+^1(\mathcal{X})$ and $\beta \in \mathcal{M}_+^1(\mathcal{Y})$, the set of couplings between $\alpha$ and $\beta$ is

$$\mathcal{U}(\alpha, \beta) := \left\{\pi \in \mathcal{M}_+^1(\mathcal{X} \times \mathcal{Y}) \; ; \; \pi_1 = \alpha \quad \text{and} \quad \pi_2 = \beta\right\}. \tag{3.4}$$

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

**Remark 3.18 (Probabilistic interpretation of couplings).** If $X \sim \alpha$ and $Y \sim \beta$, then $\pi \in \mathcal{U}(\alpha, \beta)$ means that $\pi$ is the law of a pair $(X, Y)$ whose coordinates have laws $\alpha$ and $\beta$. 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 $\pi = \sum_{i,j} P_{ij} \delta_{(x_i, y_j)}$ 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 3.19** (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 $\alpha \otimes \beta \in \mathcal{U}(\alpha, \beta)$.

**Proposition 3.20** (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,*

$$c(x, y) = u(x) + v(y).$$

*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

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

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 $\alpha_i$ and $\beta_i$ the normalized restrictions of $\alpha$ to $U_i$ and of $\beta$ 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. 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 $\alpha$ has a density and $\beta$ is discrete. In this case couplings are singular measures supported on a union of graphs or cells inside $\mathcal{X} \times \mathcal{Y}$.

**Continuous Kantorovich problem.** The discrete Kantorovich problem (3.2) becomes, for arbitrary measures, the minimization of the average cost over all couplings,

$$\mathcal{L}_c(\alpha, \beta) \coloneqq \inf_{\pi \in \mathcal{U}(\alpha,\beta)} \int_{\mathcal{X}\times\mathcal{Y}} c(x, y)\mathrm{d}\pi(x, y). \tag{3.5}$$

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

**Remark 3.21 (Probabilistic interpretation of Kantorovich's problem).** The same problem can 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.

**Proposition 3.22** (Existence on compact spaces)**.** *Assume that $\mathcal{X}$ and $\mathcal{Y}$ are compact metric spaces and that $c \in C(\mathcal{X}\times\mathcal{Y})$. Then the Kantorovich problem (3.5) admits at least one minimizer.*

*Proof.* The constraint set is non-empty because it contains the product coupling $\alpha\otimes\beta$. It is closed for weak convergence of measures because the marginal constraints are preserved under weak convergence. Since $\mathcal{X}\times\mathcal{Y}$ is compact, the set of probability measures on it is compact for the weak topology, and therefore $\mathcal{U}(\alpha,\beta)$ is compact. Finally, the functional $\pi \mapsto \int c\mathrm{d}\pi$ is weakly continuous because $c$ is continuous and bounded. The minimum is thus attained. □

On non-compact domains, one needs coercivity and moment conditions. For the Wasserstein cost $c(x,y) = d(x,y)^p$ on a Polish metric space, the natural domain is

$$\mathcal{P}_p(\mathcal{X}) := \left\{\mu \in \mathcal{M}_+^1(\mathcal{X}) \,;\, \int d(x,x_0)^p \mathrm{d}\mu(x) < +\infty\right\},$$

for one, and hence every, reference point $x_0$. If $\alpha,\beta \in \mathcal{P}_p(\mathcal{X})$, then the product coupling has finite $p$-cost up to the triangle inequality, so the Kantorovich value is finite. Existence of minimizers holds under standard lower-semicontinuity assumptions on $c$, using tightness of finite-moment sublevel sets [226, 202].

**Monge–Kantorovich equivalence.** The proof of Brenier's theorem 2.18 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 3.23** (Monge–Kantorovich equivalence under Brenier)**.** *Assume that $\alpha$ is absolutely continuous with respect to Lebesgue measure and that $c(x,y) = \|x-y\|^2$. 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 $\varphi$. When $\alpha$ has a density, $\varphi$ is differentiable $\alpha$-almost everywhere, so $\partial\varphi(x) = \{\nabla\varphi(x)\}$ for $\alpha$-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 3.24 (Nonsmooth potentials and splitting).** If $\alpha$ does not have a density, then $\varphi$ may be non-smooth on a set charged by $\alpha$, and non-smooth points can lead to mass splitting. For instance, moving $\delta_0$ 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 3.14 and the general splitting case.

**Remark 3.25 (Probabilistic form of tightness).** If $(X,Y)$ has the optimal Kantorovich law under the assumptions of Corollary 3.23, 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.

## 3.4 $c$-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.** To formalize this, one needs a precise notion of support, i.e. the closed set that carries the mass of the coupling.

**Definition 3.26** (Support)**.** For a Radon measure $\pi$ on $\mathcal{X} \times \mathcal{Y}$,

$$\operatorname{supp}(\pi) := \left\{(x, y)\ ;\ \pi(U \times V) > 0 \text{ for every open } U \ni x,\ V \ni y\right\}.$$

**Definition 3.27** ($c$-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 $\sigma$ 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 3.14, 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 [195], says that the same finite-exchange condition holds for any optimal coupling, not only for finite uniform matchings.

**Theorem 3.28** (Optimal plans are $c$-cyclically monotone)**.** *Assume $c$ is continuous. For any optimal plan $\pi$ solving the Kantorovich problem (3.5), $\operatorname{supp}(\pi)$ 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 $\sigma$ 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 $\pi(U_i \times V_i) > 0$. Because there are only finitely many rectangles, one can choose $\lambda > 0$ small enough that the scaled restrictions

$$\pi_i = \lambda \frac{\pi|_{U_i \times V_i}}{\pi(U_i \times V_i)}$$

have common mass $\lambda$ and satisfy $\sum_i \pi_i \leqslant \pi$. Let $\alpha_i = (P_{\mathcal{X}})_\sharp \pi_i$ and $\beta_i = (P_{\mathcal{Y}})_\sharp \pi_i$. Define

$$\tilde{\pi} = \pi - \sum_i \pi_i + \sum_i \frac{\alpha_i \otimes \beta_{\sigma(i)}}{\lambda}.$$

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 $\sigma$ 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 $\pi$. □

**Monotonicity.** Assume the optimal plan is induced by a measurable map $T : \mathcal{X} \to \mathcal{Y}$, i.e. $\pi = (\mathrm{Id}, T)_\sharp \alpha$. For any $k$ points $x_1, \dots, x_k$ in the domain, 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, for any $x, y$,

$$\langle T(x) - T(y),\, x - y \rangle \geqslant 0,$$

so $T$ is a monotone vector field. Brenier's theorem adds that, when $\alpha$ is absolutely continuous, $T = \nabla\varphi$ for a convex potential $\varphi$. 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 $c(x, y) = |x - y|^p$ for any $p \geqslant 1$, the two-point inequality becomes

$$|x - T(x)|^p + |y - T(y)|^p \leqslant |x - T(y)|^p + |y - T(x)|^p,$$

which is equivalent to $T(x) \leqslant T(y)$ whenever $x < y$. Thus $T$ must be nondecreasing, recovering the classical monotone rearrangement.

## 3.5 Metric Properties: Wasserstein Distances

The final part of the section proves that OT costs are genuine distances when the ground cost comes from a metric. It also compares Wasserstein convergence with total variation and explains why OT is weak enough to move Dirac masses continuously.

**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 3.29** (Discrete gluing lemma). *Given* $(\mathrm{a}, \mathrm{b}, \mathrm{c}) \in \Sigma_n \times \Sigma_p \times \Sigma_m$, *let* $\mathrm{P} \in \mathrm{U}(\mathrm{a}, \mathrm{b})$ *and* $\mathrm{Q} \in \mathrm{U}(\mathrm{b}, \mathrm{c})$. *Then there exists a 3-D tensor coupling* $\mathrm{S} \in \mathbb{R}_+^{n \times p \times m}$ *such that the 2-D marginals satisfy*

$$\sum_k \mathrm{S}_{i,j,k} = \mathrm{P}_{i,j} \quad \text{and} \quad \sum_i \mathrm{S}_{i,j,k} = \mathrm{Q}_{j,k}.$$

*Consequently the marginal between the first and third variables,*

$$\mathrm{R}_{i,k} := \sum_j \mathrm{S}_{i,j,k},$$

*belongs to* $\mathrm{U}(\mathrm{a}, \mathrm{c})$. *For the canonical construction below, this glued coupling is the twisted matrix product*

$$\mathrm{R} = \mathrm{P}\,\mathrm{diag}(1/\mathrm{b})\mathrm{Q}, \qquad \mathrm{R}_{i,k} = \sum_{j : \mathrm{b}_j > 0} \frac{\mathrm{P}_{i,j}\mathrm{Q}_{j,k}}{\mathrm{b}_j}.$$

*In the matrix notation,* $1/\mathrm{b}_j$ *is understood as* $0$ *when* $\mathrm{b}_j = 0$. *Figure 3.5 displays this construction in matrix form.*

*Proof.* One verifies that

$$S_{i,j,k} = \begin{cases} \frac{P_{i,j}Q_{j,k}}{b_j} & \text{if} \quad b_j \neq 0 \\ 0 \text{ otherwise} \end{cases} \tag{3.6}$$

is acceptable. 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

$$\sum_k R_{i,k} = \sum_j P_{i,j} = a_i, \qquad \sum_i R_{i,k} = \sum_j Q_{j,k} = c_k,$$

so $R \in U(a, c)$. □

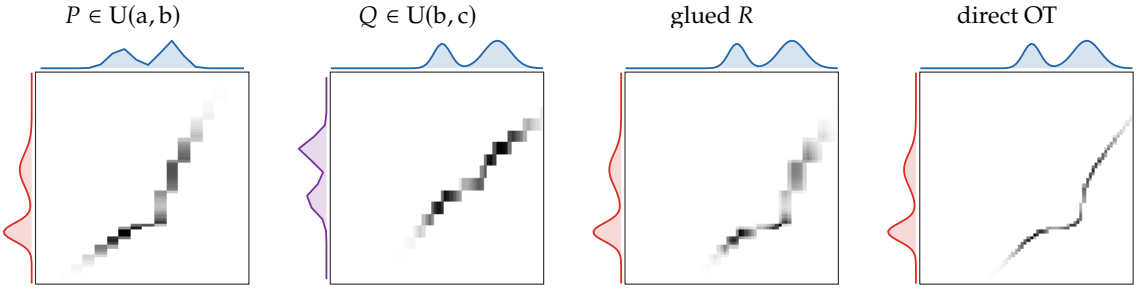



*Figure 3.5:* Discrete gluing lemma in matrix form. The first two panels are optimal one-dimensional couplings through an intermediate marginal b. The third panel shows the induced marginal $R = P\,\mathrm{diag}(1/b)Q$ between a and c; it is feasible and is the coupling used in the triangle-inequality 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 $p$th power of a distance matrix, the discrete Kantorovich value becomes a metric on histograms.

**Definition 3.30** (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

$$W_p(a, b) := L_{D^p}(a, b)^{1/p}. \tag{3.7}$$

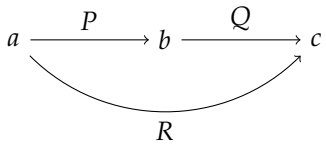
It depends on the chosen ground distance D.

**Proposition 3.31** (Metric property of discrete Wasserstein distance)**.** *For every distance matrix* D *on* $[\![n]\!]$*, Definition 3.30 defines a distance on* $\Sigma_n$*:* $W_p$ *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.* For symmetry, since $D^p$ is symmetric, we use the fact that if $P \in U(a, b)$ is optimal for $W_p(a, b)$, then $P^\top \in U(b, a)$ is optimal for $W_p(b, a)$. For definiteness, since $C = D^p$ has a null diagonal, $W_p(a, b) = 0$ is achieved by the diagonal coupling $P^\star = \mathrm{diag}(a) = \mathrm{diag}(b)$ when $a = b$; by positivity of all off-diagonal elements of $D^p$, $W_p(a, b) > 0$ whenever $a \neq b$ because any admissible coupling then has a nonzero element outside the diagonal.

To prove the triangle inequality in this discrete setting, we consider $a, b, c \in \Sigma_n$, and let P and Q be two optimal solutions of the transport problems between a and b, and b and c respectively. We use the gluing Lemma 3.29 which defines $S \in \mathbb{R}_+^{n^3}$ with marginals $\sum_k S_{\cdot,\cdot,k} = P$ and $\sum_i S_{i,\cdot,\cdot} = Q$. We define $R = \sum_j S_{\cdot,j,\cdot}$, which is an element of $U(a, c)$.




Note that if one assumes $b > 0$ then $R = P \operatorname{diag}(1/b) Q$.

The triangle inequality follows from

$$\begin{aligned}
W_p(a,c) &= \Big( \min_{\tilde{R} \in U(a,c)} \langle \tilde{R}, D^p \rangle \Big)^{1/p} \leqslant \langle R, D^p \rangle^{1/p} \\
&= \Big( \sum_{i,k} D_{ik}^p \sum_j S_{i,j,k} \Big)^{1/p} \leqslant \Big( \sum_{i,j,k} \Big( D_{ij} + D_{j,k} \Big)^p S_{i,j,k} \Big)^{1/p} \\
&\leqslant \Big( \sum_{i,j,k} D_{ij}^p S_{i,j,k} \Big)^{1/p} + \Big( \sum_{i,j,k} D_{j,k}^p S_{i,j,k} \Big)^{1/p} \\
&= \Big( \sum_{i,j} D_{i,j}^p \sum_k S_{i,j,k} \Big)^{1/p} + \Big( \sum_{j,k} D_{j,k}^p \sum_i S_{i,j,k} \Big)^{1/p} \\
&= \Big( \sum_{i,j} D_{i,j}^p P_{i,j} \Big)^{1/p} + \Big( \sum_{j,k} D_{j,k}^p Q_{j,k} \Big)^{1/p} = W_p(a,b) + W_p(b,c).
\end{aligned}$$

The first inequality follows from the feasibility of R, the second is the usual triangle inequality for elements in D, and the third comes from Minkowski's inequality. □

**Continuous gluing.** Proposition 3.31 generalizes from histogram to arbitrary measures that need not be discrete. For this, one needs the following general gluing lemma.

**Lemma 3.32** (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 $(\mathcal{X}, \mathcal{Y}, \mathcal{Z})$ are Polish spaces in the sense of Definition 2.3. Given $\pi \in \mathcal{U}(\alpha, \beta)$ and $\xi \in \mathcal{U}(\beta, \gamma)$, then there exists a tensor coupling measure $\sigma \in \mathcal{M}_+(\mathcal{X} \times \mathcal{Y} \times \mathcal{Z})$ such that*

$$(P_{\mathcal{X},\mathcal{Y}})_\sharp \sigma = \pi \quad \text{and} \quad (P_{\mathcal{Y},\mathcal{Z}})_\sharp \sigma = \xi$$

*where we denoted the projector $P_{\mathcal{X},\mathcal{Y}}(x, y, z) = (x, y)$ and $P_{\mathcal{Y},\mathcal{Z}}(x, y, z) = (y, z)$.*

*Proof.* The proof of this fundamental result is involved since it requires using the disintegration of measure (which corresponds to conditional probabilities). The disintegration of measures is applicable because the spaces are Polish. We disintegrate $\pi$ and $\xi$ against $\beta$ to obtain two families $(\pi_y)_{y \in \mathcal{Y}}$ and $(\xi_y)_{y \in \mathcal{Y}}$ of probability distributions on $\mathcal{X}$ and $\mathcal{Z}$. These families are defined by the fact that

$$\forall\, h \in C(\mathcal{X} \times \mathcal{Y}), \quad \int_{\mathcal{Y}} \Big( \int_{\mathcal{X}} h(x, y) \mathrm{d}\pi_y(x) \Big) \mathrm{d}\beta(y) = \int h(x, y) \mathrm{d}\pi(x, y).$$

and similarly for $\xi$. When $\beta = \sum_j b_j \delta_{y_j}$ and $\pi = \sum_{i,j} P_{i,j} \delta_{(x_i, y_j)}$, then this conditional distribution is defined on the support of $\beta$ as $\pi_{y_j} = \sum_i \frac{P_{i,j}}{b_j} \delta_{x_i}$ (and similarly for $\xi$). The glued measure is then defined by the conditional-product formula

$$\forall\, g \in C(\mathcal{X} \times \mathcal{Y} \times \mathcal{Z}), \quad \int g(x, y, z) \mathrm{d}\sigma(x, y, z) = \int g(x, y, z) \mathrm{d}\pi_y(x) \mathrm{d}\xi_y(z) \mathrm{d}\beta(y).$$

For discrete measures, this matches the definition (3.6), since $\sigma = \sum_{i,j,k} S_{i,j,k} \delta_{x_i, y_j, z_k}$ where

$$S_{i,j,k} = \frac{P_{i,j}}{b_j} \frac{Q_{j,k}}{b_j} b_j.$$

□

Using this gluing lemma, we can now construct the Wasserstein distance in the general setting of arbitrary distributions on a Polish space.

**Definition 3.33** (Wasserstein distance)**.** Let $(\mathcal{X}, d)$ be a metric space and $p \geqslant 1$. For $\alpha, \beta \in \mathcal{P}_p(\mathcal{X})$, the $p$-Wasserstein distance is

$$\mathcal{W}_p(\alpha,\beta) \coloneqq \mathcal{L}_{d^p}(\alpha,\beta)^{1/p} = \left(\inf_{\pi\in\mathcal{U}(\alpha,\beta)} \int_{\mathcal{X}\times\mathcal{X}} d(x,y)^p \mathrm{d}\pi(x,y)\right)^{1/p}. \tag{3.8}$$

It depends on the ground distance $d$.

**Proposition 3.34** (Metric property of the Wasserstein distance)**.** *Definition 3.33 defines a distance: $\mathcal{W}_p$ is symmetric, positive, $\mathcal{W}_p(\alpha,\beta) = 0$ if and only if $\alpha = \beta$, and it satisfies the triangle inequality*

$$\forall\,(\alpha,\beta,\gamma) \in \mathcal{P}_p(\mathcal{X})^3, \quad \mathcal{W}_p(\alpha,\gamma) \leqslant \mathcal{W}_p(\alpha,\beta) + \mathcal{W}_p(\beta,\gamma).$$

*Proof.* The symmetry follows from the fact that since $d$ is symmetric, if $\pi(x,y)$ is optimal for $\mathcal{L}_{d^p}(\alpha,\beta)$, then $\pi(y,x) \in \mathcal{U}(\beta,\alpha)$ is optimal for $\mathcal{L}_{d^p}(\beta,\alpha)$. If $\mathcal{L}_{d^p}(\alpha,\beta) = 0$, then necessarily an optimal coupling $\pi$ is supported on the diagonal $\Delta \coloneqq \{(x,x)\}_x \subset \mathcal{X}^2$. We denote $\lambda(x)$ the corresponding measure on the diagonal, i.e. such that $\int h(x,y)\mathrm{d}\pi(x,y) = \int h(x,x)\mathrm{d}\lambda(x)$. Then since $\pi \in \mathcal{U}(\alpha,\beta)$ necessarily $\lambda = \alpha$ and $\lambda = \beta$ so that $\alpha = \beta$.

For the triangle inequality, we consider optimal couplings $\pi \in \mathcal{U}(\alpha,\beta)$ and $\xi \in \mathcal{U}(\beta,\gamma)$ and we glue them according to the Lemma 3.32. We define the composition of the two couplings $(\pi,\xi)$ as $\rho \coloneqq (P_{\mathcal{X},\mathcal{Z}})_\sharp\sigma$. Note that if $\pi$ and $\xi$ are couplings induced by two Monge maps $T_{\mathcal{X}}(x)$ and $T_{\mathcal{Y}}(y)$, then $\rho$ is itself induced by the Monge map $T_{\mathcal{Y}} \circ T_{\mathcal{X}}$, so that this notion of composition of coupling generalizes the composition of maps. The triangular inequality follows from

$$\begin{aligned}
\mathcal{W}_p(\alpha,\gamma) &\leqslant \left(\int_{\mathcal{X}\times\mathcal{Z}} d(x,z)^p \mathrm{d}\rho(x,z)\right)^{1/p} = \left(\int_{\mathcal{X}\times\mathcal{Y}\times\mathcal{Z}} d(x,z)^p \mathrm{d}\sigma(x,y,z)\right)^{1/p} \\
&\leqslant \left(\int_{\mathcal{X}\times\mathcal{Y}\times\mathcal{Z}} (d(x,y)+d(y,z))^p \mathrm{d}\sigma(x,y,z)\right)^{1/p} \\
&\leqslant \left(\int_{\mathcal{X}\times\mathcal{Y}\times\mathcal{Z}} d(x,y)^p \mathrm{d}\sigma(x,y,z)\right)^{1/p} + \left(\int_{\mathcal{X}\times\mathcal{Y}\times\mathcal{Z}} d(y,z)^p \mathrm{d}\sigma(x,y,z)\right)^{1/p} \\
&= \left(\int_{\mathcal{X}\times\mathcal{Y}} d(x,y)^p \mathrm{d}\pi(x,y)\right)^{1/p} + \left(\int_{\mathcal{Y}\times\mathcal{Z}} d(y,z)^p \mathrm{d}\xi(y,z)\right)^{1/p} = \mathcal{W}_p(\alpha,\beta) + \mathcal{W}_p(\beta,\gamma).
\end{aligned}$$

□

**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 3.35** ($\mathcal{W}_2$ 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

$$e_t(x,y) \coloneqq (1-t)x + ty, \qquad \alpha_t \coloneqq (e_t)_\sharp\pi^\star.$$

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 $\mathrm{P}_{ij}$ 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 3.36** (Optimal-plan interpolation is a $\mathcal{W}_2$ geodesic)**.** *Let $(\alpha_t)_{t\in[0,1]}$ be defined by Definition 3.35. 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 $\mathcal{W}_2$.*

**Algorithm 3.4** Displacement interpolation from a transport plan

**Input:** Measures $\alpha, \beta$ on $\mathbb{R}^d$, time $t \in [0,1]$.
**Output:** Displacement interpolant $\alpha_t$.
**Let** $\pi^\star$ be any minimizer of the quadratic Kantorovich problem.
**Set** interpolation map: $e_t(x,y) = (1-t)x + ty$.
**Push forward:** $\alpha_t = (e_t)_\sharp \pi^\star$.
**If** $\pi^\star = \sum_{i,j} P^\star_{ij} \delta_{(x_i,y_j)}$ **then**:
  **Compute** $\alpha_t = \sum_{i,j} P^\star_{ij} \delta_{(1-t)x_i + t y_j}$.

**Return** $\alpha_t$.

*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 3.34, 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. □

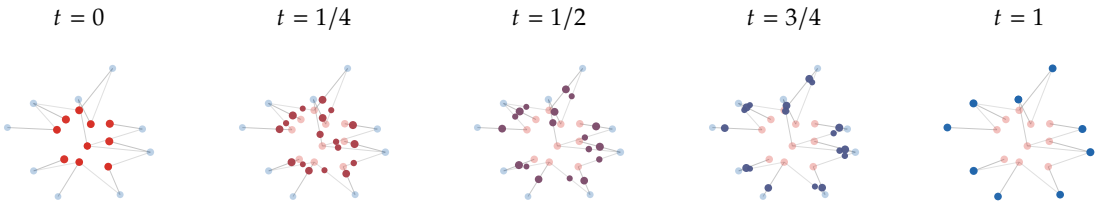



*Figure 3.6:* McCann interpolation induced by a non-deterministic optimal transport plan. In every panel, the red and blue endpoint measures are shown with low opacity, thin gray segments display the support $\mathrm{P}_{ij} > \mathrm{tol}$ of the coupling, and the moving atoms are colored from red to blue along the interpolation.

**General geodesic spaces.** For Dirac masses in Euclidean space, the $\mathcal{W}_2$ geodesic from $\delta_x$ to $\delta_y$ 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 $\eta$ 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 [7, 226, 202].

**Comparison with Monge.** This distance $\mathcal{W}_p$ defined through the Kantorovich problem (3.8) should be contrasted with the directed distance $\tilde{\mathcal{W}}$ 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 $(\mathrm{Id}, T)_\sharp \alpha$; in this sense the Kantorovich problem is the convex relaxation of Monge's problem, with much better stability properties.

## 3.6 Metric Properties: 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 3.37** (Equivalence of Wasserstein distances on compact spaces)**.** *One has for $p \leqslant q$*

$$\mathcal{W}_p(\alpha,\beta) \leqslant \mathcal{W}_q(\alpha,\beta) \leqslant \mathrm{diam}(\mathcal{X})^{\frac{q-p}{q}}\, \mathcal{W}_p(\alpha,\beta)^{\frac{p}{q}}$$

*where* $\mathrm{diam}(\mathcal{X}) \coloneqq \sup_{x,y} d(x,y)$.

*Proof.* The left inequality follows from Jensen inequality, $\varphi(\int c(x,y)\mathrm{d}\pi(x,y)) \leqslant \int \varphi(c(x,y))\mathrm{d}\pi(x,y)$, applied to any probability distribution $\pi$ and to the convex function $\varphi(r) = r^{q/p}$ with $c(x,y) = d(x,y)^p$, so that one gets

$$\left(\int d(x,y)^p \mathrm{d}\pi(x,y)\right)^{\frac{q}{p}} \leqslant \int d(x,y)^q \mathrm{d}\pi(x,y).$$

The right inequality follows from

$$d(x,y)^q \leqslant \mathrm{diam}(\mathcal{X})^{q-p} d(x,y)^p.$$

□

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 [226, 202, 6], while empirical rates are quantified in [82, 92, 232, 37, 38].

**Definition 3.38** (Weak* topology)**.** $(\alpha_k)_k$ converges weakly* to $\alpha$ in $\mathcal{M}_+^1(\mathcal{X})$ (denoted $\alpha_k \rightharpoonup \alpha$) if and only if for any bounded continuous function $f \in C_b(\mathcal{X})$, $\int_{\mathcal{X}} f\mathrm{d}\alpha_k \to \int_{\mathcal{X}} f\mathrm{d}\alpha$. On compact spaces, $C_b(\mathcal{X}) = C(\mathcal{X})$, which is why the boundedness is often left implicit there.

**Remark 3.39 (A Riemann-sum weak limit).** On $\mathcal{X} = \mathbb{R}$, the empirical measures on a regular grid satisfy

$$\frac{1}{n}\sum_{k=1}^{n} \delta_{k/n} \rightharpoonup \mathcal{U}_{[0,1]}.$$

Indeed, for every continuous bounded function $f$,

$$\frac{1}{n}\sum_{k=1}^{n} f(k/n) \longrightarrow \int_0^1 f(x)\mathrm{d}x,$$

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 3.40 (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 $\alpha_n$ 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.

In terms of random vectors, if $X_n \sim \alpha_n$ and $X \sim \alpha$ (not necessarily defined on the same probability space), weak convergence corresponds to convergence in law of $X_n$ toward $X$.

**Remark 3.41 (Modes of convergence for random variables).** Convergence of laws should be distinguished from stronger notions of convergence for random variables. If $X_n$ and $X$ are defined on a common probability space, then $X_n \to X$ 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, and convergence in probability implies convergence in law. Convergence in law is exactly weak* convergence of the probability measures $(X_n)_\sharp \mathbb{P} \rightharpoonup X_\sharp \mathbb{P}$, and does not require all variables to live on the same probability space. Strong convergence of measures, for instance convergence in total variation, is different and usually much stronger: it controls the mass assigned to all measurable sets, not only averages against continuous test functions. In particular, total variation convergence implies weak convergence, but the converse fails for empirical approximations of continuous laws.

**Remark 3.42 (Central limit theorem).** The central limit theorem states that if $(X_1, \dots, X_n)$ 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 rescaled average $Z_n := \frac{1}{\sqrt{n}} \sum_{i=1}^n X_i$ converges in law toward a Gaussian $\mathcal{N}(0, \mathrm{Id})$. This means that the measure $\alpha_n$ representing the law of $Z_n$ converges weakly toward the measure $\alpha$ of the centered normalized Gaussian.

The total variation norm was introduced in Definition 2.6 and Proposition 2.7. 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 $\ell^1$ 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 3.43** (Total variation as Wasserstein for the discrete metric)**.** *Denoting $d$ the 0/1 distance such that $d(x, x) = 0$ and $d(x, y) = 1$ if $x \neq y$, then*

$$\mathcal{W}_p(\alpha, \beta)^p = \frac{1}{2}\|\alpha - \beta\|_{\mathrm{TV}}.$$

*Proof.* For the sake of simplicity, we do the proof for discrete measures with weights $(\mathrm{a}, \mathrm{b})$ and without loss of generality assume they have the same support $(x_i)_i$ and we denote $\mathrm{D} := (d(x_i, x_j))_{i,j}$ which is 0 on the diagonal and one outside. Also since $d^p = d$ we consider $p = 1$. We denote $\mathrm{c}_i = \min(\mathrm{a}_i, \mathrm{b}_i)$. By conservation of mass, for every $\mathrm{P} \in \mathrm{U}(\mathrm{a}, \mathrm{b})$, $\mathrm{P}_{i,i} \leqslant \mathrm{c}_i$, thus

$$\langle \mathrm{P}, \mathrm{D} \rangle = \sum_{i \neq j} \mathrm{P}_{i,j} = 1 - \sum_i \mathrm{P}_{i,i} \geqslant 1 - \sum_i \mathrm{c}_i.$$

We need to show that this bound is tight, namely to construct $\hat{P} \in \mathrm{U}(\mathrm{a}, \mathrm{b})$ such that $\mathrm{diag}(\hat{P}) = \mathrm{c}$. Let

$$\bar{\mathrm{a}} := \mathrm{a} - \mathrm{c} = (\mathrm{a} - \mathrm{b})_+ \geqslant 0 \quad \text{and} \quad \bar{\mathrm{b}} := \mathrm{b} - \mathrm{c} = (\mathrm{b} - \mathrm{a})_+ \geqslant 0$$

If $\bar{\mathrm{a}} = \bar{\mathrm{b}} = 0$, then $\mathrm{a} = \mathrm{b}$ and the diagonal coupling is optimal. Otherwise, one has

$$\frac{\bar{\mathrm{a}} \otimes \bar{\mathrm{b}}}{\langle \bar{\mathrm{a}}, \mathbb{1} \rangle} \in \mathrm{U}(\bar{\mathrm{a}}, \bar{\mathrm{b}})$$

and we remark that $\langle \bar{\mathrm{a}}, \mathbb{1} \rangle = \langle \bar{\mathrm{b}}, \mathbb{1} \rangle = 1 - \langle \mathrm{c}, \mathbb{1} \rangle$. Thus denoting

$$\hat{\mathrm{P}} := \mathrm{diag}(\mathrm{c}) + \frac{\bar{\mathrm{a}} \otimes \bar{\mathrm{b}}}{\langle \bar{\mathrm{a}}, \mathbb{1} \rangle} \geqslant 0$$

satisfies

$$\hat{\mathrm{P}}\mathbb{1} = \mathrm{c} + \bar{\mathrm{a}} = \mathrm{a} \quad \text{and} \quad \hat{\mathrm{P}}^\top \mathbb{1} = \mathrm{c} + \bar{\mathrm{b}} = \mathrm{b}$$

so that $\hat{P} \in \mathrm{U}(\mathrm{a}, \mathrm{b})$ is a coupling so that $\mathrm{diag}(\hat{P}) = \mathrm{diag}(\mathrm{c})$ since $\mathrm{diag}(\bar{\mathrm{a}} \otimes \bar{\mathrm{b}}) = 0$. We thus conclude that

$$\mathrm{W}_1(\mathrm{a}, \mathrm{b}) = \langle \mathrm{D}, \hat{\mathrm{P}} \rangle = \sum_{i,j} \frac{\bar{\mathrm{a}}_i \bar{\mathrm{b}}_j}{\langle \bar{\mathrm{a}}, \mathbb{1} \rangle} = \sum_i \bar{\mathrm{a}}_i = \sum_i \bar{\mathrm{b}}_i = \frac{1}{2} \sum_i (\bar{\mathrm{a}}_i + \bar{\mathrm{b}}_i) = \frac{1}{2}\|\mathrm{a} - \mathrm{b}\|_{\mathrm{TV}}.$$

□

As explained in Remark 3.40, 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 $x_n \neq x$,

$$\|\delta_{x_n} - \delta_x\|_{\mathrm{TV}} = 2 \quad \text{and} \quad \mathcal{W}_p(\delta_{x_n}, \delta_x) = d(x_n, x).$$

This shows that for the strong topology, Diracs never converge, while they do converge for the Wasserstein distance. It is a powerful property of the Wasserstein distance: on compact spaces, it metrizes weak convergence.

**Proposition 3.44** (Wasserstein metrizes weak convergence on compact spaces)**.** *If $\mathcal{X}$ is compact, $\alpha_k \rightharpoonup \alpha$ if and only if $\mathcal{W}_p(\alpha_k, \alpha) \to 0$.*

*Proof.* For $p = 1$, this is the Kantorovich–Rubinstein metrization theorem: by duality, $\mathcal{W}_1$ is the supremum over 1-Lipschitz test functions, and on a compact metric space this class is compact modulo constants by Arzelà–Ascoli. Thus convergence in $\mathcal{W}_1$ is equivalent to weak convergence. Proposition 3.37 then shows that all Wasserstein distances $\mathcal{W}_p$ induce the same convergent sequences on compact spaces. Hence weak convergence is equivalent to convergence in $\mathcal{W}_p$ for every $p \geqslant 1$. □

On non-compact spaces, one needs also to impose convergence of the $p$-th moments. More precisely, on a Polish metric space, $\mathcal{W}_p(\alpha_k, \alpha) \to 0$ if and only if $\alpha_k \rightharpoonup \alpha$ and, for some reference point $x_0$,

$$\int d(x, x_0)^p \mathrm{d}\alpha_k(x) \longrightarrow \int d(x, x_0)^p \mathrm{d}\alpha(x).$$

On a discrete space, the strong and weak topologies coincide, and the following proposition relates the TV and Wasserstein distances.

**Proposition 3.45** (Comparison with total variation on discrete spaces)**.** *One has*

$$\frac{d_{\min}}{2}\|\alpha - \beta\|_{\mathrm{TV}} \leqslant \mathcal{W}_1(\alpha, \beta) \leqslant \frac{d_{\max}}{2}\|\alpha - \beta\|_{\mathrm{TV}} \quad \textit{where} \quad \begin{cases} d_{\min} := \inf_{x \neq y} d(x, y) \\ d_{\max} := \sup_{x, y} d(x, y) \end{cases}$$

*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 $\pi$ gives the result using Proposition 3.43. □

This bound is sharp, as this can be observed by taking $\alpha = \delta_x$ and $\beta = \delta_y$, in which case the bound simply reads if $x \neq y$

$$d_{\min} \leqslant d(x, y) \leqslant d_{\max}.$$

This shows that the ratio between the two distances can blow up as $d_{\max}/d_{\min}$ increases. On non-discrete spaces, if $d_{\min} = 0$, then the two distances are not equivalent, in line with the fact that the strong and weak topologies do not coincide.

## 3.7 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 3.46** (Wasserstein spaces as ground spaces)**.** *If $(\mathcal{X}, d)$ is a Polish metric space, then $\mathcal{P}_p(\mathcal{X})$ endowed with $\mathcal{W}_p$ is Polish. If $\mathcal{X}$ is compact, then $\mathcal{P}(\mathcal{X})$ is compact for the Wasserstein topology, and the construction can be iterated to form $\mathcal{P}(\mathcal{P}(\mathcal{X}))$, $\mathcal{P}(\mathcal{P}(\mathcal{P}(\mathcal{X})))$, and so on.*

*Proof.* This is a standard structural theorem for Wasserstein spaces [226, 202, 7]. Completeness follows by representing a $\mathcal{W}_p$-Cauchy sequence through almost optimally glued couplings, which gives a Cauchy random sequence whose law is the desired limit; separability follows by approximating measures with finitely supported measures on a countable dense subset and rational weights. If $\mathcal{X}$ is compact, Prokhorov compactness gives compactness of $\mathcal{P}(\mathcal{X})$ for weak convergence, and Proposition 3.44 identifies this topology with any Wasserstein topology. □

We denote elements of $\mathcal{P}_2(\mathcal{X})$ by $\alpha, \beta, \ldots$. Elements of $\mathcal{P}_2(\mathcal{P}_2(\mathcal{X}))$ 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 family $(\alpha_\zeta)_{\zeta\in Z}$ and a probability law $\gamma$ on the parameter space:

$$\mathfrak{A} = (\zeta \mapsto \alpha_\zeta)_\sharp \gamma. \tag{3.9}$$

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 3.47** (Collapsed, or barycentric, mixture)**.** For $\mathfrak{A} \in \mathcal{P}(\mathcal{P}_2(\mathcal{X}))$, the collapsed, or barycentric, mixture associated with $\mathfrak{A}$ is the measure $\bar{\alpha}_\mathfrak{A}$ defined by

$$\int_\mathcal{X} f(x) \mathrm{d}\bar{\alpha}_\mathfrak{A}(x) = \int_{\mathcal{P}_2(\mathcal{X})} \left( \int_\mathcal{X} f(x) \mathrm{d}\alpha(x) \right) \mathrm{d}\mathfrak{A}(\alpha), \tag{3.10}$$

for bounded continuous $f$.

In the finite case, $\bar{\alpha}_\mathfrak{A} = \sum_i a_i \alpha_{\zeta_i}$.

The Wasserstein distance on the Wasserstein space is

$$\mathbb{W}_2^2(\mathfrak{A}, \mathfrak{B}) := \inf_{\Pi \in \mathcal{U}(\mathfrak{A},\mathfrak{B})} \int_{\mathcal{P}_2(\mathcal{X})\times\mathcal{P}_2(\mathcal{X})} \mathcal{W}_2^2(\alpha, \beta) \mathrm{d}\Pi(\alpha, \beta). \tag{3.11}$$

For Gaussian mixtures, this separates 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 (3.11) uses the component cost

$$C_{ij} = \|m_i - n_j\|^2 + \mathcal{B}(\Sigma_i, \Lambda_j)^2.$$

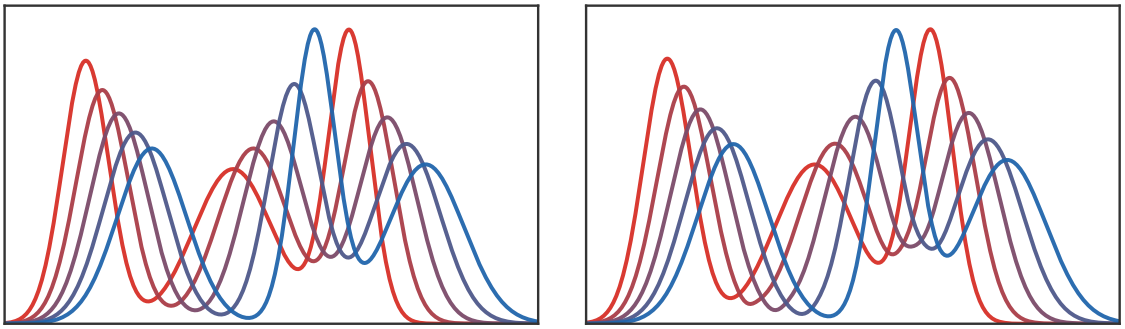
Let $\Pi^\star$ be an optimal coupling between the weights $a$ and $b$. If $A_{ij}$ denotes the Brenier linear part from $\Sigma_i$ to $\Lambda_j$, then each active component pair is interpolated by

$$m_{ij,t} = (1-t)m_i + t n_j, \qquad \Sigma_{ij,t} = \big((1-t)\mathrm{Id} + tA_{ij}\big)\Sigma_i\big((1-t)\mathrm{Id} + tA_{ij}\big),$$

and collapsing these component geodesics gives

$$\bar{\alpha}_t = \sum_{i,j} \Pi_{ij}^\star \mathcal{N}(m_{ij,t}, \Sigma_{ij,t}).$$

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.



component law on Gaussian atoms collapsed mixture law

*Figure 3.7:* Two interpolations between the same three-component one-dimensional Gaussian mixtures. On the left, each mixture is represented as a discrete law over Gaussian components, and the components are matched using the Bures-Wasserstein distance between Gaussians. On the right, the mixtures are first collapsed into ordinary one-dimensional densities and then interpolated by the true quantile formula for $\mathcal{W}_2$. The two constructions encode different geometries.

**Proposition 3.48** (Collapsing is non-expansive)**.** *Let $\mathfrak{A}, \mathfrak{B} \in \mathcal{P}_2(\mathcal{P}_2(\mathcal{X}))$, and let $\bar{\alpha}_\mathfrak{A}$ and $\bar{\beta}_\mathfrak{B}$ be the collapsed mixtures defined by (3.10). Then*

$$\mathcal{W}_2(\bar{\alpha}_\mathfrak{A}, \bar{\beta}_\mathfrak{B}) \leqslant \mathbb{W}_2(\mathfrak{A}, \mathfrak{B}).$$

*Proof.* Fix $\Pi \in \mathcal{U}(\mathfrak{A}, \mathfrak{B})$. For every $(\alpha, \beta)$ choose, by a standard measurable selection argument and up to an arbitrarily small error, a coupling $\pi_{\alpha,\beta} \in \mathcal{U}(\alpha, \beta)$ whose quadratic cost is $\mathcal{W}_2^2(\alpha, \beta)$. Integrating this Markov kernel against $\Pi$ gives a coupling $\bar{\pi}$ between $\bar{\alpha}_\mathfrak{A}$ and $\bar{\beta}_\mathfrak{B}$. Its cost satisfies

$$\int_{\mathcal{X}\times\mathcal{X}} d(x,y)^2 \mathrm{d}\bar{\pi}(x,y) \leqslant \int_{\mathcal{P}_2(\mathcal{X})^2} \mathcal{W}_2^2(\alpha,\beta)\mathrm{d}\Pi(\alpha,\beta)$$

up to the arbitrary selection error. Taking first the infimum over $\bar{\pi}$ and then over $\Pi$ proves the claim. □

The following remark records a useful way in which this iterated construction reappears later for Gromov–Wasserstein lower bounds.

> **Remark 3.49 (Local profiles as Wasserstein-over-Wasserstein laws).** Given a metric-measure space $\mathbb{X} = (\mathcal{X}, d_\mathcal{X}, \mu_\mathcal{X})$, each point defines a local distance distribution
>
> $$\alpha_x = (d_\mathcal{X}(x,\cdot))_\sharp \mu_\mathcal{X} \in \mathcal{P}(\mathbb{R}_+), \qquad \mathfrak{D}_\mathcal{X} = (x \mapsto \alpha_x)_\sharp \mu_\mathcal{X} \in \mathcal{P}(\mathcal{P}(\mathbb{R}_+)).$$
>
> The Mémoli profile lower bound in Proposition 11.11 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. Note that there exist alternative distances which also metricize weak convergence. The simplest ones are Hilbertian kernel norms, which are detailed in Section 6.1.

## 3.8 Distributional Robustness and $\mathcal{W}_\infty$

**DRO ambiguity sets.** Wasserstein distances are also used to define ambiguity sets around an empirical law. Given samples $z_i$ and $\hat{\alpha}_n = \frac{1}{n}\sum_i \delta_{z_i}$, a distributionally robust optimization (DRO) problem replaces the empirical risk $\frac{1}{n}\sum_i \ell_\theta(z_i)$ by

$$\sup_{\beta:\, \mathcal{W}_p(\beta,\hat{\alpha}_n)\leqslant\rho} \int \ell_\theta(z)\mathrm{d}\beta(z),$$

or, in Lagrangian penalized form, by $\sup_\beta \int \ell_\theta \mathrm{d}\beta - \lambda\, \mathcal{W}_p(\beta, \hat{\alpha}_n)^p$. The constrained and penalized formulations are linked by the choice of multiplier $\lambda$, but are not the same problem for an arbitrary fixed $\lambda$. Both ask for performance against distributions that can be reached by transporting the empirical mass within a budget. The radius $\rho$ is expressed in the geometry of the data space, so it can encode feature perturbations, domain shift or model misspecification [166, 33, 100].

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

$$\sup_{\beta:\mathcal{W}_p(\beta,\hat{\alpha}_n)^p\leqslant\rho^p}\int \ell_\theta \mathrm{d}\beta = \inf_{\lambda\geqslant 0}\lambda\rho^p + \frac{1}{n}\sum_{i=1}^{n}\sup_{z}\{\ell_\theta(z)-\lambda d(z,z_i)^p\}. \tag{3.12}$$

Thus the robust risk is an empirical risk in which each sample is replaced by its worst penalized perturbation. For $p = 1$ and an $L_\theta$-Lipschitz loss, the Kantorovich–Rubinstein dual gives the transparent upper bound

$$\sup_{\beta:\mathcal{W}_1(\beta,\hat{\alpha}_n)\leqslant\rho}\int \ell_\theta \mathrm{d}\beta \leqslant \frac{1}{n}\sum_{i}\ell_\theta(z_i)+\rho L_\theta,$$

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 (3.12) 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 [208, 207]. 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 [234, 235].

**Proposition 3.50** (Convexity of transport costs)**.** *For any nonnegative lower-semicontinuous cost $c$, the value*

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

*is jointly convex. In particular, for a ground metric $d$ and $p\geqslant 1$, the map $(\alpha,\beta)\mapsto \mathcal{W}_p(\alpha,\beta)^p$ is jointly convex. The distance $\mathcal{W}_1$ is jointly convex, but $\mathcal{W}_p$ itself need not be convex for $p>1$.*

*Proof.* Let $\pi_0\in\mathcal{U}(\alpha_0,\beta_0)$ and $\pi_1\in\mathcal{U}(\alpha_1,\beta_1)$ be $\eta$-optimal. Then $(1-t)\pi_0+t\pi_1$ is a coupling between $(1-t)\alpha_0+t\alpha_1$ and $(1-t)\beta_0+t\beta_1$, and its cost is the corresponding convex combination of the two costs. Letting $\eta\to 0$ proves joint convexity of $\mathcal{L}_c$. Taking $c=d^p$ gives convexity of $\mathcal{W}_p^p$. For $p=1$, this is convexity of $\mathcal{W}_1$ itself. For $p>1$, the root can destroy convexity: on the real line, $F(t):=\mathcal{W}_p((1-t)\delta_0+t\delta_1,\delta_0)=t^{1/p}$ satisfies $F(1/2)>(F(0)+F(1))/2$. □

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

$$\theta\mapsto \sup_{\beta:\mathcal{W}_p(\beta,\hat{\alpha}_n)\leqslant\rho}\int \ell_\theta\,\mathrm{d}\beta$$

is convex as a supremum of convex functions of $\theta$. This explains why Wasserstein DRO can preserve convex learning formulations while still modeling adversarial distributional shifts.

**$\mathcal{W}_\infty$ robustness.** The limiting distance

$$\mathcal{W}_\infty(\alpha,\beta):=\inf_{\pi\in\mathcal{U}(\alpha,\beta)}\operatorname*{ess\,sup}_{(x,y)\sim\pi} d(x,y) \tag{3.13}$$

is the limit of $\mathcal{W}_p(\alpha,\beta)$ as $p\to\infty$ on bounded spaces, not the limit of the convex programs defining $\mathcal{W}_p^p$. It minimizes the worst displacement rather than an average displacement, and the resulting optimization is no longer a linear convex program because of the essential-supremum objective. This makes $\mathcal{W}_\infty$ less convenient algorithmically, but very natural in robust formulations where one wants to guarantee that every transported sample stays within a prescribed perturbation radius.

**Proposition 3.51** ($\mathcal{W}_\infty$ robust envelope around an empirical law)**.** *Let $(\mathcal{Z},d)$ be a Polish metric space. Let $\hat{\alpha}=\sum_{i=1}^n a_i\delta_{z_i}$ with $a_i>0$ and $\sum_i a_i=1$, and assume that the closed balls $\overline{B}(z_i,\rho)$ are compact. For any real-valued upper-semicontinuous loss $\ell$,*

$$\sup_{\beta:\mathcal{W}_\infty(\beta,\hat{\alpha})\leqslant\rho}\int \ell(z)\mathrm{d}\beta(z)=\sum_{i=1}^{n}a_i\sup_{z\in\overline{B}(z_i,\rho)}\ell(z).$$

*Proof.* If $\mathcal{W}_\infty(\beta, \hat{\alpha}) \leqslant \rho$, then, by symmetry, there are couplings $\pi_m \in \mathcal{U}(\hat{\alpha}, \beta)$ whose essential displacements are at most $\rho + 1/m$. Since the two marginals are fixed probability measures on a Polish space, $\mathcal{U}(\hat{\alpha}, \beta)$ is tight and closed, hence weakly compact by Prokhorov's theorem. After extraction, $\pi_m \rightharpoonup \pi \in \mathcal{U}(\hat{\alpha}, \beta)$. 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 $\pi$ with respect to the first marginal gives $\pi = \sum_i a_i \delta_{z_i} \otimes \nu_i$, where each $\nu_i$ 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 $\rho$, so this $\beta$ is feasible and attains the displayed value. □

## 3.9 Quantitative Central Limit Theorems

The weak topology only says whether laws converge; Wasserstein distances also quantify how fast they converge. The next result is a representative theoretical application: the central limit theorem becomes a rate estimate in $\mathcal{W}_1$. This is useful because $\mathcal{W}_1$ is exactly the supremum over 1-Lipschitz observables, so the bound controls the error of all stable numerical or statistical measurements of the normalized sum at once. Figure 3.8 illustrates the elementary Bernoulli case.

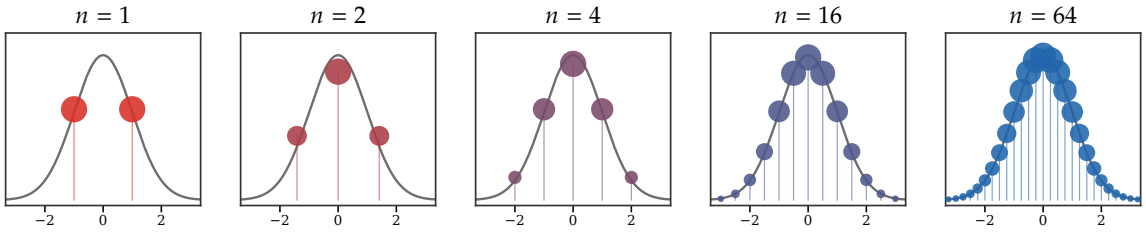



*Figure 3.8:* Central-limit theorem for normalized Bernoulli sums. Starting from $\alpha_0 = \frac{1}{2}(\delta_{-1} + \delta_1)$, the law of $Z_n = n^{-1/2} \sum_{i=1}^n X_i$ remains discrete, but its rescaled atom heights approach the standard Gaussian density shown in gray. The Wasserstein Berry–Esseen bound below quantifies this weak convergence by a $\mathcal{W}_1$ distance.

The following estimate is the Wasserstein form of the classical Berry–Esseen theorem [24, 88]. The proof sketch uses Stein's method, whose modern normal-approximation formulation is developed in [57].

**Proposition 3.52** (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 $\alpha_n$ is the law of $n^{-1/2} \sum_i X_i$ and $\gamma$ 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.* We give the standard Stein-method sketch. By Kantorovich–Rubinstein duality,

$$\mathcal{W}_1(\alpha_n, \gamma) = \sup_{\mathrm{Lip}(h) \leqslant 1} |\mathbb{E}h(S_n) - \mathbb{E}h(G)|, \qquad S_n = n^{-1/2} \sum_i X_i, \; G \sim \gamma.$$

For each such $h$, solve Stein's equation

$$f_h'(x) - x f_h(x) = h(x) - \mathbb{E}h(G).$$

The solution satisfies uniform derivative bounds depending only on the Lipschitz constant of $h$. Hence

$$\mathbb{E}h(S_n) - \mathbb{E}h(G) = \mathbb{E}[f_h'(S_n) - S_n f_h(S_n)].$$

Expanding $f_h'(S_n)$ and $f_h(S_n)$ by replacing the summands one at a time, the first- and second-order terms cancel because $\mathbb{E}X_i = 0$ and $\mathbb{E}X_i^2 = 1$. The Taylor remainder is bounded by $C\sum_i \mathbb{E}|X_i/\sqrt{n}|^3$, which gives the displayed $n^{-1/2}$ rate. Sharper constants and higher-order transport-distance refinements are studied in [34, 194]. □

CHAPTER 4

# 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 [28], while the continuous form is one of the central theorems of OT [225, 202].

## 4.1 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 4.1** (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:

$$R(a, b) := \left\{(f, g) \in \mathbb{R}^n \times \mathbb{R}^m \;;\; \forall\, (i, j) \in [\![n]\!] \times [\![m]\!], f_i + g_j \leqslant C_{i,j}\right\}. \tag{4.1}$$

Equivalently, $f \oplus g \leqslant C$ entrywise. The notation suppresses the dependence on $C$, which is fixed in the surrounding problem.

The two vectors play the role of source and target prices; admissibility means that no transported pair is priced above its travel cost.

**Proposition 4.2** (Discrete Kantorovich duality)**.** *One has*

$$L_C(a, b) = \max_{(f,g)\in R(a,b)} \langle f, a\rangle + \langle g, b\rangle \tag{4.2}$$

*Proof.* For the sake of completeness, let us derive this dual problem using Lagrangian duality. The Lagrangian associated to (3.2) reads

$$\min_{P\geqslant 0} \max_{(f,g)\in\mathbb{R}^n\times\mathbb{R}^m} \langle C, P\rangle + \langle a - P\mathbb{1}_m, f\rangle + \langle b - P^\top \mathbb{1}_n, g\rangle. \tag{4.3}$$

For a linear program, if the primal constraint set is non-empty, one can always exchange the min and the max and get the same value. We thus consider

$$\max_{(f,g)\in\mathbb{R}^n\times\mathbb{R}^m} \langle a, f\rangle + \langle b, g\rangle + \min_{P\geqslant 0} \langle C - f\mathbb{1}_m^\top - \mathbb{1}_n g^\top, P\rangle.$$

We conclude by remarking that

$$\min_{P\geqslant 0} \langle Q, P\rangle = \begin{cases} 0 & \text{if} \quad Q \geqslant 0 \\ -\infty & \text{otherwise} \end{cases}$$

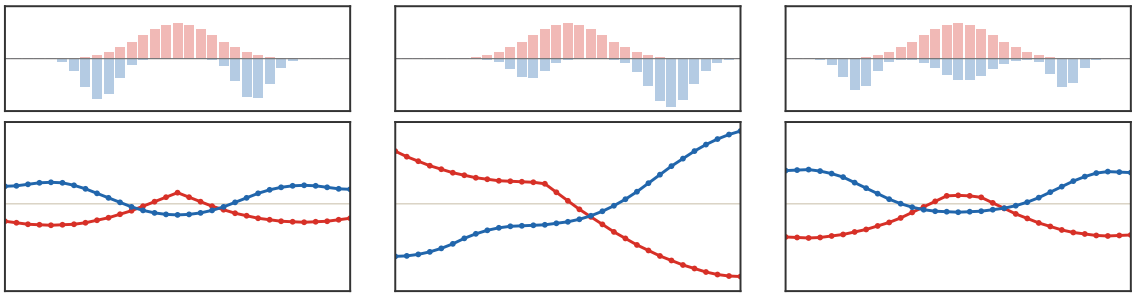
so that the constraint reads $C - f\mathbb{1}_m^\top - \mathbb{1}_n g^\top = C - f \oplus g \geqslant 0$. □

The primal-dual optimality relation for the Lagrangian (4.3) allows locating the support of the optimal transport plan

$$\mathrm{Supp}(P) \subset \left\{(i, j) \in [\![n]\!] \times [\![m]\!] \;;\; f_i + g_j = C_{i,j}\right\}. \tag{4.4}$$

Figure 4.1 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.

The formulation (4.2) shows that $(a, b) \mapsto L_C(a, b)$ is a convex function (as a supremum of linear functions). From the primal problem (3.2), one also sees that $C \mapsto L_C(a, b)$ is concave.



balanced target strongly shifted target three-mode target

*Figure 4.1:* Discrete Kantorovich dual potentials for the quadratic cost $C_{i,j} = |x_i - y_j|^2$. 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 $\langle f, a \rangle = 0$. Complementary slackness states that mass can be transported only through entries where $f_i + g_j = C_{i,j}$.

## 4.2 Auction Algorithm and Dual Prices

The assignment algorithms mentioned in Chapter 1 become more transparent once one has dual variables. The auction algorithm is a dual price method: it updates target prices and maintains an approximate complementary-slackness certificate. The small tolerance $\varepsilon$ removes ties, stabilizes the price updates and gives a quantitative optimality certificate [25, 26, 164].

Consider the square assignment problem with costs $C_{i,j}$ and rewrite it as the profit maximization problem with $a_{i,j} = -C_{i,j}$. The auction algorithm keeps prices $p_j$ on the target points and a partial assignment. For an unassigned source $i$, define the best and second-best reduced profits

$$v_i = \max_j (a_{i,j} - p_j), \qquad j_i \in \operatorname*{argmax}_j (a_{i,j} - p_j), \qquad w_i = \max_{j \neq j_i} (a_{i,j} - p_j).$$

Source $i$ bids for $j_i$ and increases its price by the gap to the second-best target, plus a margin:

$$p_{j_i} \leftarrow p_{j_i} + v_i - w_i + \varepsilon.$$

The target $j_i$ is then assigned to $i$, and its previous owner, if any, becomes unassigned. The iteration stops when all sources are assigned. Algorithm 4.1 records the bidding loop. Figure 4.2 displays actual auction iterates using the same assignment-state convention as Figure 1.8: flat rows denote unassigned bidders, and one-hot rows denote currently owned targets.

**Algorithm 4.1** Auction bidding with target prices

**Input:** Profit matrix $A = (a_{ij})$, bid increment $\varepsilon > 0$.
**Output:** Assignment map $\sigma$.
**Initialize:** Set prices $p_j = 0$, ownership map $o(j) = \emptyset$, and $U = \{1, \dots, n\}$.
**While** the unassigned set $U$ is nonempty **do**:
  **Set** $i = \min U$.
  **Set** $j_i = \min \operatorname{argmax}_j (a_{ij} - p_j)$.
  **Set** $v_i = a_{ij_i} - p_{j_i}$ and $w_i = \max_{j \neq j_i} (a_{ij} - p_j)$.
  **Update price:** $p_{j_i} \leftarrow p_{j_i} + v_i - w_i + \varepsilon$.
  **If** $o(j_i) = i' \neq \emptyset$ **then**:
    **Set** $U \leftarrow U \cup \{i'\}$.
  **Set** $o(j_i) = i$ and $U \leftarrow U \setminus \{i\}$.
**Return** $\sigma(i) = j$ iff $o(j) = i$.

For fixed prices $p$, eliminating the bidder utilities $u_i$ in the dual minimization gives the convex objective

$$D(p) = \sum_j p_j + \sum_i \max_j (a_{i,j} - p_j),$$

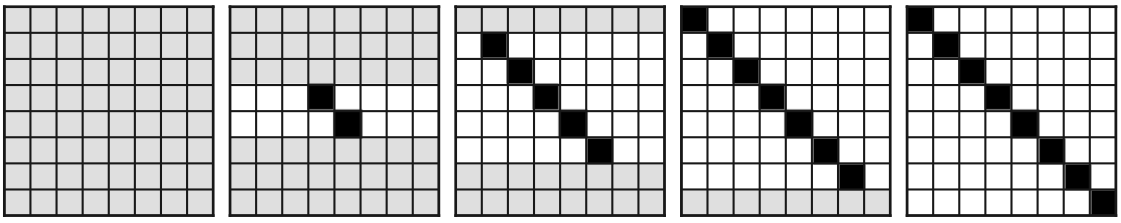



*Figure 4.2:* Matrix view of actual auction iterates on the same diagonally dominant one-dimensional squared-distance assignment as Figure 1.8. Each panel records the current ownership state: unassigned bidders are shown as flat rows, while assigned bidders are shown as one-hot rows at their currently held target. The snapshots show initialization, intermediate price updates, and the final identity assignment satisfying complementary slackness.

which comes from the dual constraints $u_i + p_j \geqslant a_{i,j}$. The auction update should thus be viewed as a price-adjustment method on this nonsmooth dual landscape, rather than as a generic gradient step: the proof of correctness is through approximate complementary slackness.

**Definition 4.3** ($\varepsilon$-complementary slackness)**.** An assignment $\sigma$ and prices $p$ satisfy $\varepsilon$-complementary slackness if, for every source $i$,

$$a_{i,\sigma(i)} - p_{\sigma(i)} \geqslant \max_j (a_{i,j} - p_j) - \varepsilon.$$

**Proposition 4.4** (Auction optimality certificate)**.** *If a complete assignment $\sigma$ satisfies $\varepsilon$-complementary slackness, then it is $n\varepsilon$-optimal for the profit maximization problem, or equivalently $n\varepsilon$-optimal for the original cost minimization problem. If all costs are integers and $\varepsilon < 1/n$, then $\sigma$ is optimal.*

*Proof.* Let $\tau$ be any assignment. By $\varepsilon$-complementary slackness,

$$a_{i,\tau(i)} - p_{\tau(i)} \leqslant \max_j (a_{i,j} - p_j) \leqslant a_{i,\sigma(i)} - p_{\sigma(i)} + \varepsilon.$$

Summing over $i$ cancels prices, because both $\sigma$ and $\tau$ are permutations:

$$\sum_i a_{i,\tau(i)} \leqslant \sum_i a_{i,\sigma(i)} + n\varepsilon.$$

Thus no assignment has profit more than $n\varepsilon$ above that of $\sigma$. Since $a = -C$, the same statement says that the cost of $\sigma$ is at most $n\varepsilon$ above the minimum cost. If the costs are integers, all assignment costs are integers; a gap strictly smaller than one therefore forces the gap to be zero. □

During the bidding process the last bid made by a source makes its chosen target better, up to the margin $\varepsilon$, than all alternatives. Subsequent price increases can only make targets less attractive, and one checks by induction that currently assigned pairs satisfy $\varepsilon$-complementary slackness. The standard finite-termination proof normalizes prices by subtracting their minimum, observes that each bid increases one target price by at least $\varepsilon$, and bounds the normalized price spread in terms of the range of the profits; see [26, 164] for the full bound and implementation details.

**Remark 4.5 ($\varepsilon$-scaling and relation with Sinkhorn).** In practice one starts with a coarse $\varepsilon$ and repeatedly decreases it, warm-starting the prices and assignment. This $\varepsilon$-scaling strategy is a homotopy method: large $\varepsilon$ regularizes the combinatorial problem by enforcing a visible margin between the best and second-best reduced profits, while small $\varepsilon$ recovers the exact dual certificate. If one wants a continuous-optimization analogy, the margin is closer to an exact-penalty or proximal continuation parameter than to a literal quadratic penalty.

Sinkhorn scaling plays a parallel role for entropic OT. There, the hard minimum in the dual $c$-transform is replaced by a soft minimum, or log-sum-exp, with temperature $\varepsilon$; in the auction algorithm, the hard maximum is kept but the complementary slackness condition is relaxed by $\varepsilon$. Both methods therefore use an $\varepsilon$-controlled dual continuation, and both recover the unregularized transport certificate as $\varepsilon \to 0$ under the usual assumptions. The outputs are different: Sinkhorn produces dense entropic couplings, whereas auction keeps a sparse assignment throughout.

## 4.3 General formulation

The continuous dual is the analytic counterpart of the discrete linear program. It uses continuous potentials because measures are probed through integration.

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 $\langle f,\, \alpha\rangle := \int f \mathrm{d}\alpha$.

**Proposition 4.6** (Kantorovich duality)**.** *Assume that $\mathcal{X}$ and $\mathcal{Y}$ are compact metric spaces and that $c \in \mathcal{C}(\mathcal{X}\times\mathcal{Y})$. Then*

$$\mathcal{L}_c(\alpha,\beta) = \max_{(f,g)\in\mathcal{R}(c)} \int_{\mathcal{X}} f(x)\mathrm{d}\alpha(x) + \int_{\mathcal{Y}} g(y)\mathrm{d}\beta(y), \tag{4.5}$$

*where the set of admissible dual potentials is*

$$\mathcal{R}(c) := \{(f,g) \in \mathcal{C}(\mathcal{X}) \times \mathcal{C}(\mathcal{Y}) \;;\; \forall (x,y),\, f(x) + g(y) \leqslant c(x,y)\}\,. \tag{4.6}$$

*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 "$\leqslant$".

For the reverse inequality, view the primal problem as a linear program over the locally convex space of Radon measures, paired with $\mathcal{C}(\mathcal{X}\times\mathcal{Y})$. The affine map $\pi \mapsto (\pi_1,\pi_2)$ is continuous for the weak topology, the feasible set is non-empty because it contains $\alpha\otimes\beta$, and the cost is continuous and bounded on compact sets. Since the set of probability measures on the compact product is weakly compact, the set of attainable cost-marginal triples is closed after adding the epigraph variable below. The separating-hyperplane theorem applied to the convex set of attainable triples

$$\left\{\left(\pi_1,\pi_2,\int c\mathrm{d}\pi + r\right) : \pi \geqslant 0,\; r \geqslant 0\right\}$$

gives a continuous affine separator, hence functions $(f,g)$ and a scalar multiplier which can be normalized so that $f\oplus g \leqslant c$. The separating inequality then states that the supremum over such potentials is at least the primal value. This proves equality. The same argument is the infinite-dimensional analogue of the finite linear-programming proof in Proposition 4.2. □

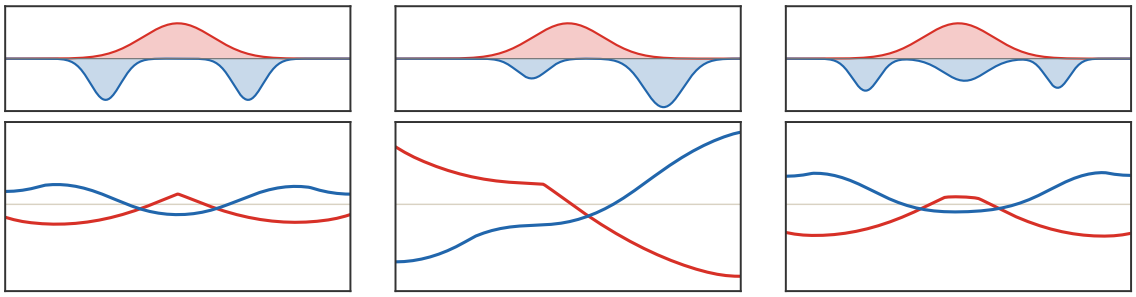
> **Remark 4.7 (Dual attainment from $c$-transforms).** Under the compactness and continuity assumptions of Proposition 4.6, the maximum in (4.5) is attained. If $c$ is Lipschitz, Proposition 4.10 shows that one may replace an admissible pair by its $c$-transforms without decreasing the dual objective; after fixing one additive gauge, the transformed potentials are uniformly bounded and equi-Lipschitz. Arzelà–Ascoli then gives a converging maximizing subsequence, and the closed constraint $f\oplus g\leqslant c$ passes to the limit.

The discrete case (4.2) corresponds to the dual vectors being samples of the continuous potentials, *i.e.* $(\mathrm{f}_i,\mathrm{g}_j) = (f(x_i), g(y_j))$. The primal-dual optimality conditions allow for tracking the support of the optimal plan, and (4.4) is generalized as

$$\mathrm{Supp}(\pi) \subset \{(x,y) \in \mathcal{X}\times\mathcal{Y} \;;\; f(x)+g(y) = c(x,y)\}\,. \tag{4.7}$$

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 $g = f^c$.

Note that in contrast to the primal problem (3.5), showing the existence of solutions to (4.5) is non-trivial, because the constraint set $\mathcal{R}(c)$ is not compact and the objective is not coercive. Using the machinery of $c$-transforms detailed in Section 4.4, one can show that optimal $(f,g)$ are necessarily Lipschitz regular, which enables the replacement of the constraint by a compact one.



balanced target strongly shifted target three-mode target

*Figure 4.3:* Continuous Kantorovich potentials for the same source and target families as Figure 4.1. The upper strips show the source density $\alpha$ in red and the target density $\beta$ 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.

## 4.4 $c$-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 (4.5), which leads one to consider

$$\sup_{g \in C(\mathcal{Y})} \left\{ \int g \mathrm{d}\beta \; ; \; \forall (x, y), g(y) \leqslant c(x, y) - f(x) \right\}.$$

The constraint can be replaced by

$$\forall y \in \mathcal{Y}, \quad g(y) \leqslant f^c(y)$$

.

**Definition 4.8** ($c$-transform)**.** For a function $f : \mathcal{X} \to \mathbb{R} \cup \{-\infty\}$, its $c$-transform is

$$\forall y \in \mathcal{Y}, \quad f^c(y) := \inf_{x \in \mathcal{X}} c(x, y) - f(x). \tag{4.8}$$

For a function $g : \mathcal{Y} \to \mathbb{R} \cup \{-\infty\}$, the $\bar{c}$-transform associated with $\bar{c}(y, x) = c(x, y)$ is

$$\forall x \in \mathcal{X}, \quad g^{\bar{c}}(x) := \inf_{y \in \mathcal{Y}} c(x, y) - g(y).$$

Since $\beta$ is positive, the maximization of $\int g \mathrm{d}\beta$ is thus achieved at those functions such that $g = f^c$ on the support of $\beta$, which means $\beta$-almost everywhere.

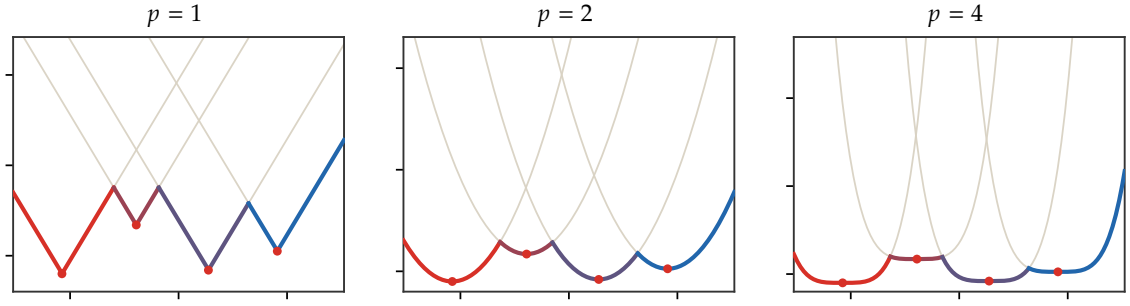



*Figure 4.4:* Discrete $c$-transform as a lower envelope for costs $c_p(x, y) = |x - y|^p$. The red circles are four source atoms $x_i$ with potential values $f_i$; the gray curves are the translated functions $y \mapsto c_p(x_i, y) - f_i$; the colored curve is their lower envelope $f^c(y) = \min_i c_p(x_i, y) - f_i$. This is the semi-discrete situation where $\mathcal{X}$ is finite, equivalently the source measure $\alpha$ is discrete; Chapter 5 studies the complementary case where the eliminated potential is supported on finitely many target atoms.

**Proposition 4.9** ($c$-transforms solve the semi-relaxed problems)**.** *For fixed $f$, the maximizers of the dual objective over all $g$ such that $f \oplus g \leqslant c$ are exactly the functions satisfying $g = f^c$ $\beta$-almost everywhere. Equivalently, $f^c$ gives the value of the one-marginal primal problem*

$$\inf_{\pi:\, \pi_2=\beta} \int c(x,y)\mathrm{d}\pi(x,y) - \int f(x)\mathrm{d}\pi_1(x) = \int f^c(y)\mathrm{d}\beta(y).$$

*Symmetrically, for fixed $g$, the maximizers over $f$ are the functions satisfying $f = g^{\bar{c}}$ $\alpha$-almost everywhere.*

*Proof.* The constraint $f(x) + g(y) \leqslant c(x,y)$ for all $x$ is equivalent, for each fixed $y$, to

$$g(y) \leqslant \inf_x c(x,y) - f(x) = f^c(y).$$

Since $\beta$ is nonnegative, the largest possible value of $\int g\mathrm{d}\beta$ is obtained by saturating this pointwise upper bound on the support of $\beta$. The proof for $f = g^{\bar{c}}$ is identical after exchanging the two marginals.

For the primal formula, disintegrate any feasible $\pi$ as $\pi(\mathrm{d}x, \mathrm{d}y) = \pi_y(\mathrm{d}x)\beta(\mathrm{d}y)$. Then

$$\int c\mathrm{d}\pi - \int f\mathrm{d}\pi_1 = \int \left( \int (c(x,y) - f(x))\mathrm{d}\pi_y(x) \right) \mathrm{d}\beta(y) \geqslant \int f^c(y)\mathrm{d}\beta(y).$$

If minimizers admit a measurable selection, equality is obtained by choosing $\pi_y$ supported on minimizers of $x \mapsto c(x,y) - f(x)$. Otherwise one uses approximate measurable selections and lets the approximation error vanish. □

The map $(f, g) \mapsto (g^{\bar{c}}, f^c)$ replaces dual potentials by better ones, in the sense that it preserves feasibility and improves the dual objective. Functions of the form $f^c$ and $g^{\bar{c}}$ are called $c$-concave and $\bar{c}$-concave functions. These partial minimizations define maximizers on the supports of $\alpha$ and $\beta$, while Definition 4.8 defines functions on the whole spaces $\mathcal{X}$ and $\mathcal{Y}$. This gives a canonical extension of dual solutions beyond the active supports.

**Proposition 4.10** (Lipschitz stability of $c$-transforms)**.** *If $c$ is $L$-Lipschitz with respect to its second variable, uniformly in the first one and for the metric $d_{\mathcal{Y}}$ on $\mathcal{Y}$, then $f^c$ is $L$-Lipschitz.*

*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 $F_x$ are $L$-Lipschitz,

$$|F(y) - F(y')| = \left| \inf_x F_x(y) - \inf_x F_x(y') \right| \leqslant \sup_x |F_x(y) - F_x(y')| \leqslant L d_{\mathcal{Y}}(y, y').$$

□

This stability is crucial for dual attainment. When $c$ is Lipschitz on compact spaces, one can replace arbitrary admissible potentials by $c$-transformed ones with a uniform Lipschitz bound; after fixing the harmless additive gauge, compactness follows from the Arzela–Ascoli theorem.

**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.

The special cost $c(x,y) = -\langle x, y\rangle$ on $\mathcal{X} = \mathcal{Y} = \mathbb{R}^d$ is central because it reduces the quadratic Wasserstein problem to convex duality. Indeed, for any $\pi \in \mathcal{U}(\alpha, \beta)$,

$$\int \|x - y\|^2 \mathrm{d}\pi(x,y) = \int \|x\|^2 \mathrm{d}\alpha(x) + \int \|y\|^2 \mathrm{d}\beta(y) - 2\int \langle x, y\rangle \mathrm{d}\pi(x,y).$$

For $c(x,y) = -\langle x, y\rangle$, one has

$$f^c(y) = \inf_x -\langle x, y\rangle - f(x) = -(-f)^*(y), \qquad h^*(y) \coloneqq \sup_x \langle x, y\rangle - h(x).$$

Thus $c$-concave functions are negatives of convex functions. In the one-dimensional bilinear model case, the hard double $c$-transform is therefore an operation of taking concave envelopes.

**Remark 4.11 (Proof of Brenier's theorem).** For $c(x,y) = \|x-y\|^2$, subtracting the harmless quadratic terms reduces the geometry to the bilinear cost $-\langle x, y\rangle$. The primal-dual relationship, together with the fact that one can replace $(f,g)$ by $(f^{c\bar{c}}, f^c)$, shows that an optimal plan satisfies

$$\mathrm{supp}(\pi) \subset \{(x,y)\,;\ \varphi(x) + \varphi^*(y) = \langle x, y\rangle\}\,,$$

where $\varphi = -f^{c\bar{c}}$ is convex and $-g = \varphi^*$. By the Fenchel inequality, equality holds exactly when $y \in \partial\varphi(x)$. If $\alpha$ has a density, convex functions are differentiable Lebesgue-almost everywhere, hence $\alpha$-almost everywhere, so $\partial\varphi(x)$ is a singleton for $\alpha$-almost every $x$. This yields the Brenier map $T = \nabla\varphi$ and explains why the optimal coupling is concentrated on a graph.

**The failure of alternate optimization.** A crucial property of the Legendre transform is that $f^{***} = f^*$, and that $f^{**}$ is the convex envelope of $f$ (the largest convex function below $f$). These properties carry over for the more general setting of $c$-transforms. The required convex-analytic background is standard in Rockafellar's theory [195].

**Proposition 4.12** (Algebra of $c$-transforms)**.** *The following identities, in which the inequality sign between vectors should be understood elementwise, hold, denoting $f^{c\bar{c}} := (f^c)^{\bar{c}}$:*

$$\text{(i) } f \leqslant f' \Rightarrow f^c \geqslant f'^c, \qquad \text{(ii) } f^{c\bar{c}} \geqslant f, \qquad \text{(iii) } g^{\bar{c}c} \geqslant g, \qquad \text{(iv) } f^{c\bar{c}c} = f^c.$$

*Proof.* The first inequality (i) follows from the definition of $c$-transforms (because of the $-$ sign). To prove (ii), expanding the definition of $f^{c\bar{c}}$ we have

$$\left(f^{c\bar{c}}\right)(x) = \min_y c(x,y) - f^c(y) = \min_y c(x,y) - \min_{x'}(c(x',y) - f(x')).$$

Now, since $-\min_{x'}\left(c(x',y) - f(x')\right) \geqslant -(c(x,y) - f(x))$, we recover

$$(f^{c\bar{c}})(x) \geqslant \min_y c(x,y) - c(x,y) + f(x) = f(x).$$

The relation $g^{\bar{c}c} \geqslant g$ is obtained in the same way. Now, to prove (iv), we first apply (ii) and then (i) with $f' = f^{c\bar{c}}$ to have $f^c \geqslant f^{c\bar{c}c}$. Then we apply (iii) to $g = f^c$ to obtain $f^c \leqslant f^{c\bar{c}c}$. □

This invariance property shows that one can "improve" only once the dual potential this way. Indeed, starting from any pair $(f,g)$, one obtains the following iterates by alternating maximization

$$(f,g) \mapsto (f, f^c) \mapsto (f^{c\bar{c}}, f^c) \mapsto (f^{c\bar{c}}, f^{c\bar{c}c}) = (f^{c\bar{c}}, f^c)\ldots \tag{4.9}$$

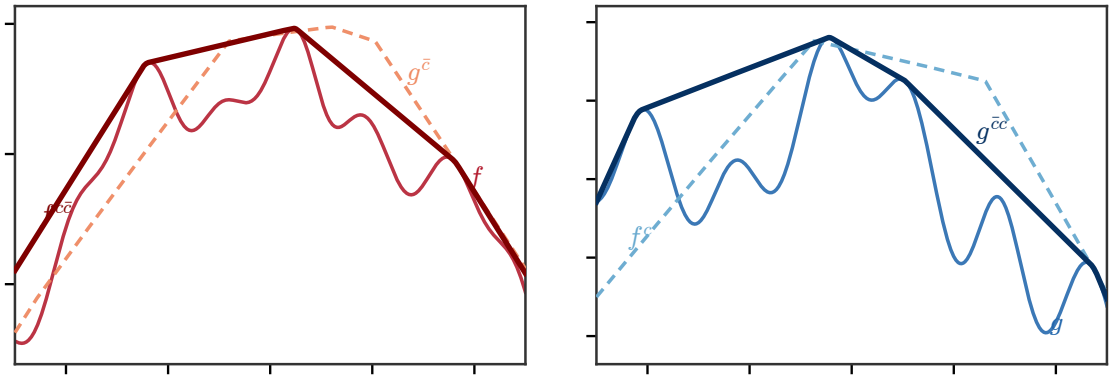
so that one reaches a stationary point.

**Algorithm 4.2** Hard alternating $c$-transform closure

**Input:** Source potential $f$ on $\mathcal{X}$, cost $c$.
**Output:** Closed $c$-concave pair $(\tilde{f}, \tilde{g})$.
**Set** target best response: $g = f^c$, $\quad g(y) = \inf_{x\in\mathcal{X}}\ c(x,y) - f(x)$.
**Set** source closure: $\tilde{f} = g^{\bar{c}} = f^{c\bar{c}}$.
**Set** closed target potential: $\tilde{g} = \tilde{f}^c = f^{c\bar{c}c} = f^c$. **Return** $(\tilde{f}, \tilde{g}) = (f^{c\bar{c}}, f^c)$.

This failure is the classical behavior of alternating maximization on a non-smooth problem, where the non-smooth part of the functional (here the constraint) mixes the two variables. The workaround is to introduce smoothing, which is the classical method of augmented Lagrangian, and that we will develop here using entropic regularization, which corresponds to Sinkhorn's algorithm.

For the bilinear cost $c(x,y) = -xy$ on a compact interval, the $c$-concave functions are ordinary concave functions and $f^{c\bar{c}}$ is the smallest concave majorant of $f$. In that model case, a hard transform removes non-concave oscillations in one closure step rather than producing a gradual ascent.




source-side closure target-side closure

*Figure 4.5:* 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 $g^{\bar{c}c}$, which are concave majorants, while dashed lighter curves are the one-sided best responses $g^{\bar{c}}$ and $f^c$ after a harmless vertical gauge shift. 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.

CHAPTER 5

# Semi-discrete and $\mathcal{W}_1$

This chapter focuses on two computationally useful degeneracies of the dual problem. Semi-discrete OT turns a continuous-to-discrete map into finite-dimensional geometry, while $\mathcal{W}_1$ replaces convex potentials by Lipschitz functions and flow fields. The material connects computational geometry [14, 161, 162] with the Kantorovich–Rubinstein and Beckmann formulations [129, 16].

## 5.1 Semi-dual

The semi-dual eliminates one potential by an exact $c$-transform. It keeps concavity while removing explicit inequality constraints.

Write the dual problem (4.5) as

$$\sup_{f,g\in C(\mathcal{X})\times C(\mathcal{Y})} \mathcal{E}(f,g)$$

where $\mathcal{E}(f,g)$ is the dual objective, with value $-\infty$ when the feasibility constraint fails. One can optimize out $g$ exactly and obtain the following semi-dual problem

$$\sup_{f\in C(\mathcal{X})} \tilde{\mathcal{E}}(f) \coloneqq \mathcal{E}(f,f^c) = \sup_g \mathcal{E}(f,g) = \int_{\mathcal{X}} f\,\mathrm{d}\alpha + \int_{\mathcal{Y}} f^c\,\mathrm{d}\beta. \tag{5.1}$$

Partial maximization of a concave problem preserves concavity, so $\tilde{\mathcal{E}}$ is still concave. The major advantage of the semi-dual is that it removes the explicit inequality constraint, which allows the use of simpler optimization algorithms.

## 5.2 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 $\beta = \sum_j \mathrm{b}_j \delta_{y_j}$ is discrete (of course the same construction applies if $\alpha$ is discrete by exchanging the role of $\alpha,\beta$). One can adapt the definition of the $\bar{c}$ transform (4.8) to this setting by restricting the minimization to the support $(y_j)_j$ of $\beta$,

$$\forall\, \mathrm{g}\in\mathbb{R}^m,\ \forall\, x\in\mathcal{X},\quad \mathrm{g}^{\bar{c}}(x) \coloneqq \min_{j\in[\![m]\!]} c(x,y_j) - \mathrm{g}_j. \tag{5.2}$$

This transform maps a vector g to a continuous function $\mathrm{g}^{\bar{c}} \in C(\mathcal{X})$ under the same regularity assumptions on $c$ as in the continuous setting. Note that this definition coincides with (4.8) when the target space $\mathcal{Y}$ is restricted to the support of $\beta$.

Crucially, using the discrete $\bar{c}$-transform, when $\beta$ is a discrete measure, yields a finite-dimensional optimization,

$$\mathcal{L}_c(\alpha,\beta) = \max_{\mathrm{g}\in\mathbb{R}^m} \mathcal{E}(\mathrm{g}) \coloneqq \int_{\mathcal{X}} \mathrm{g}^{\bar{c}}(x)\mathrm{d}\alpha(x) + \sum_j \mathrm{g}_j \mathrm{b}_j. \tag{5.3}$$

The geometric object encoded by the dual weights is a weighted nearest-neighbor diagram: each point of the source space is assigned to the target atom that realizes the discrete $\bar{c}$-transform.

**Definition 5.1** (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

$$\mathbb{L}_j(\mathrm{g}) \coloneqq \left\{ x\in\mathcal{X}\ ;\ \forall\, j'\neq j, c(x,y_j) - \mathrm{g}_j \leqslant c(x,y_{j'}) - \mathrm{g}_{j'} \right\}. \tag{5.4}$$

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.

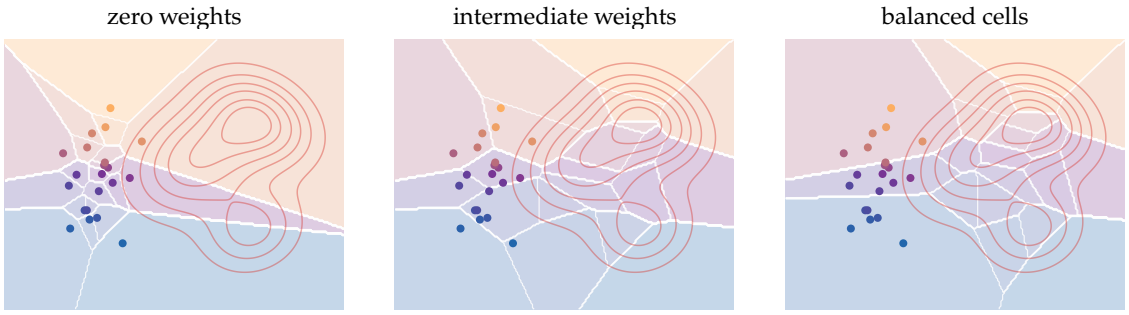



*Figure 5.1:* Laguerre cells for semi-discrete quadratic transport. The red contours show a continuous source density $\alpha$ given by a three-component Gaussian mixture on the right. The twenty-one colored circular sites are the atoms of the discrete target $\beta$ sampled from a compact Gaussian cloud 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 $\alpha$-mass captured by each cell approaches the prescribed target mass.

**Mass balance.** This allows one to conveniently rewrite the semi-dual energy as

$$\mathcal{E}(\mathrm{g}) = \sum_{j=1}^{m} \int_{\mathbb{L}_j(\mathrm{g})} \big(c(x, y_j) - \mathrm{g}_j\big)\, \mathrm{d}\alpha(x) + \langle \mathrm{g}, \mathrm{b} \rangle. \tag{5.5}$$

The following proposition provides a formula for the gradient of this concave function.

**Proposition 5.2** (Gradient of the semi-discrete dual)**.** *If $\alpha$ gives zero mass to the Laguerre cell boundaries, then $\mathcal{E}$ is differentiable at* g *and*

$$\forall\, j \in [\![m]\!], \quad \nabla\mathcal{E}(\mathrm{g})_j = \mathrm{b}_j - \int_{\mathbb{L}_j(\mathrm{g})} \mathrm{d}\alpha.$$

*Proof.* For $\alpha$-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

$$\frac{\mathrm{d}}{\mathrm{d}t}\bigg|_{t=0} \min_j \big(c(x, y_j) - (\mathrm{g}_j + t h_j)\big) = -h_{j(x)}.$$

Dominated convergence gives

$$\mathrm{d}\mathcal{E}(\mathrm{g})[h] = -\sum_j h_j \int_{\mathbb{L}_j(\mathrm{g})} \mathrm{d}\alpha + \sum_j h_j \mathrm{b}_j,$$

which is the announced gradient formula. □

The first-order optimality condition shows that solving the dual semi-discrete problem amounts to choosing the weights g so that $\int_{\mathbb{L}_j(\mathrm{g})} \mathrm{d}\alpha = \mathrm{b}_j$, i.e. each cell captures the prescribed amount of mass. In this case, the optimal transport $T$ with $T_\sharp \alpha = \beta$ is piecewise constant and maps $x \in \mathbb{L}_j(\mathrm{g})$ to $y_j$; for the quadratic cost, uniqueness follows from Brenier's theorem when $\alpha$ has a density.

The quadratic power diagrams of Definition 5.1 have polyhedral cells and can be computed efficiently using computational geometry algorithms [13, 14, 161]. One classical construction lifts the sites to points $(y_j, \|y_j\|^2 - \mathrm{g}_j) \in \mathbb{R}^{d+1}$ and obtains the power diagram by projecting the lower envelope of their convex hull. In dimensions two and three, Chan's output-sensitive convex-hull algorithm [56] has complexity $O(m \log Q)$ for $m$ sites and $Q$ hull vertices.

**Algorithm 5.1** Semi-discrete Laguerre descent

**Input:** Source measure $\alpha$, target atoms $(y_j, \mathrm{b}_j)$, cost $c$, steps $\tau_k$.
**Output:** Semi-discrete dual weights g and Laguerre cells.
**Initialize:** Set $\mathrm{g}^{(0)} = 0$.
**For** $k = 0, 1, \dots$ **do**:
  **Compute cells:** $\mathbb{L}_j(\mathrm{g}^{(k)}) = \left\{x \;;\; c(x, y_j) - \mathrm{g}_j^{(k)} \leqslant c(x, y_\ell) - \mathrm{g}_\ell^{(k)} \quad \forall \ell\right\}$.
  **Compute masses:** $m_j^{(k)} = \int_{\mathbb{L}_j(\mathrm{g}^{(k)})} \mathrm{d}\alpha$.
  **Update** $\mathrm{g}^{(k+1)} = \mathrm{g}^{(k)} + \tau_k\big(\mathrm{b} - m^{(k)}\big)$.
  **If** $\max_j \left|m_j^{(k)} - \mathrm{b}_j\right| \leqslant \mathrm{tol}$ **then**:
    **Return** $\mathrm{g}^{(k+1)}$ and the cells.

**Stochastic optimization.** The semi-discrete formulation (5.5) is useful because the objective is an expectation with respect to $\alpha$,

$$\mathcal{E}(\mathrm{g}) = \int_{\mathcal{X}} E(\mathrm{g}, x)\mathrm{d}\alpha(x) = \mathbb{E}_X(E(\mathrm{g}, X)), \qquad E(\mathrm{g}, x) \coloneqq \mathrm{g}^{\bar{c}}(x) + \langle \mathrm{g}, \mathrm{b}\rangle. \tag{5.6}$$

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 $\nabla\mathcal{E}(\mathrm{g})$ when cell boundaries have $\alpha$-measure zero. One can therefore maximize (5.5) without first discretizing $\alpha$: 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

$$\mathrm{g}^{(\ell+1)} \coloneqq \mathrm{g}^{(\ell)} + \tau_\ell \nabla_{\mathrm{g}} E(\mathrm{g}^{(\ell)}, x_\ell). \tag{5.7}$$

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 step size must decay so that the sampling noise averages out. A typical schedule is

$$\tau_\ell \coloneqq \frac{\tau_0}{1 + \ell/\ell_0}, \tag{5.8}$$

where $\ell_0$ is a warmup scale. Under standard stochastic-approximation assumptions, one obtains the usual sublinear rate

$$\mathcal{E}(\mathrm{g}^\star) - \mathbb{E}\big(\mathcal{E}(\mathrm{g}^{(\ell)})\big) = O(\ell^{-1/2}),$$

where $\mathrm{g}^\star$ is a maximizer and the expectation is over the i.i.d. samples. This stochastic viewpoint is one of the main algorithmic advantages of the semi-discrete formulation [161, 104].

## 5.3 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.

The optimal quantization problem for a measure $\alpha$ is

$$\mathcal{Q}_m(\alpha) \coloneqq \min_{Y=(y_j)_{j=1}^m,\ \mathrm{b}\in\Sigma_m} \mathcal{W}_p\left(\alpha, \sum_{j=1}^m \mathrm{b}_j \delta_{y_j}\right). \tag{5.9}$$

This problem is classical in approximation theory and information theory [111, 149]. The OT formulation emphasizes that one optimizes both the support locations $Y$ and, unless prescribed, the masses b.

**Algorithm 5.2** Stochastic semi-discrete ascent

**Input:** Source sampler $x \sim \alpha$, target atoms $(y_j, \mathrm{b}_j)$, steps $\tau_\ell$.
**Output:** Stochastic semi-discrete dual weights g.
**Initialize:** Set $\mathrm{g}^{(0)} = 0$.
**For** $\ell = 0, 1, \dots$ **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** $\mathrm{g}^{(\ell)}$ or its running average.

**Proposition 5.3** (Quantization rate and curse of dimensionality)**.** *Let $\Omega \subset \mathbb{R}^d$ be a bounded Lipschitz domain and assume $\alpha = \rho \, \mathrm{d}x$ on $\Omega$, with $0 < \rho_- \leqslant \rho \leqslant \rho_+ < +\infty$. 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.* For the upper bound, partition $\Omega$ into $m$ cells of diameter at most $Cm^{-1/d}$, up to boundary effects, and put one codepoint in each non-empty cell. Sending each point to the codepoint in its cell gives a transport distance bounded by $Cm^{-1/d}$.

For the lower bound, fix any set $Y$ of $m$ codepoints and write $d_Y(x) = \min_j \|x - y_j\|$. Since the density is bounded above, the mass of the $t$-neighborhood of $Y$ is at most $Cmt^d$. Choosing $t_0 \simeq m^{-1/d}$ small enough gives $\alpha(\{d_Y > t\}) \geqslant c$ for $0 < t < t_0$. Hence

$$\int d_Y(x)^p \mathrm{d}\alpha(x) = \int_0^{+\infty} p t^{p-1} \alpha(\{d_Y > t\}) \mathrm{d}t \geqslant c t_0^p \simeq c m^{-p/d}.$$

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. For fixed codepoints $Y$, the problem is convex with respect to the weights b. The dependence on $Y$ is non-convex and is generally computationally hard. The one-dimensional case is substantially simpler: monotonicity fixes the ordering of the cells, reducing the problem to interval endpoints and centroids; for the uniform law with the quadratic cost, the optimal centroids are equally spaced.

**Proposition 5.4** (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*

$$\min_{\mathrm{b} \in \Sigma_m} \mathcal{W}_p^p \left( \alpha, \sum_j \mathrm{b}_j \delta_{y_j} \right) = \int_{\mathcal{X}} \min_{1 \leqslant j \leqslant m} c(x, y_j) \mathrm{d}\alpha(x).$$

*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 $\alpha$ 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

$$\mathcal{Q}_m(\alpha)^p = \min_Y \mathcal{F}(Y), \qquad \mathcal{F}(Y) := \int_{\mathcal{X}} \min_{1 \leqslant j \leqslant m} c(x, y_j) \mathrm{d}\alpha(x).$$

At a differentiability point of this energy, each local minimizer 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 [149]. With standard tie-breaking, the objective decreases at each step. Since the problem is non-convex in $Y$, the iterates generally converge only to a local minimum. Good seeding matters; for the squared Euclidean objective, $k$-means++ gives a logarithmic approximation guarantee in expectation [12].

**Algorithm 5.3** Lloyd quantization

**Input:** Source measure $\alpha$, initial codepoints $Y^{(0)} = (y_j^{(0)})_{j=1}^m$, squared Euclidean cost, tolerance tol.
**Output:** Codepoints $Y = (y_j)_{j=1}^m$.
**Initialize:** Set $d_0 = +\infty$ and $k = 0$.
**While** $d_k >$ tol **do**:
  **Set** $k \leftarrow k+1$.
  **Compute Voronoi cells:** $\mathbb{V}_j(Y^{(k-1)}) = \left\{x \;;\; c(x, y_j^{(k-1)}) \leqslant c(x, y_\ell^{(k-1)}) \quad \forall \ell\right\}$.
  **For** each nonempty cell $\mathbb{V}_j$ **do**
    $y_j^{(k)} = \frac{\int_{\mathbb{V}_j(Y^{(k-1)})} x\,\mathrm{d}\alpha(x)}{\int_{\mathbb{V}_j(Y^{(k-1)})} \mathrm{d}\alpha(x)}$.
  **For** each empty cell $\mathbb{V}_j$ **do**:
    **Set** $y_j^{(k)} = y_j^{(k-1)}$.
  **Set** $d_k = \max_j \|y_j^{(k)} - y_j^{(k-1)}\|$.
**Return** $Y^{(k)}$.

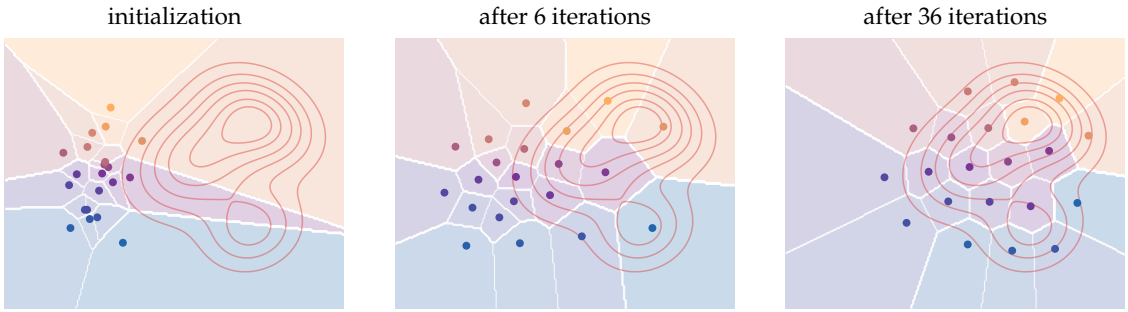



*Figure 5.2:* Lloyd quantization for the same continuous density and twenty-one initial sites as Figure 5.1. The red contours show the density $\alpha$, while the colored disks are the current codepoints and have the same colors as their Voronoi cells. The iterations move the initially left-located sites toward the high-density region and reshape the cells according to centroidal Voronoi geometry.

## 5.4 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 5.5** (Lipschitz constant)**.** For a function $f : \mathcal{X} \to \mathbb{R}$ on a metric space $(\mathcal{X}, d)$, its Lipschitz constant is

$$\mathrm{Lip}(f) \coloneqq \sup \left\{ \frac{|f(x) - f(y)|}{d(x, y)} \;;\; x \neq y \right\}. \tag{5.10}$$

The function is 1-Lipschitz when $\mathrm{Lip}(f) \leqslant 1$.

**Proposition 5.6** ($c$-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. □

Using the alternating $c$-transform scheme (4.9), one can replace the dual pair by $(f, -f)$ with $\mathrm{Lip}(f) \leqslant 1$. The Kantorovich dual therefore becomes the Kantorovich–Rubinstein formula

$$\mathcal{W}_1(\alpha, \beta) = \max_f \left\{ \int_{\mathcal{X}} f \mathrm{d}(\alpha - \beta) \;;\; \mathrm{Lip}(f) \leqslant 1 \right\}. \tag{5.11}$$

This expression depends only on the signed measure $\alpha - \beta$. It therefore extends to finite signed measures of total mass zero and defines the Kantorovich–Rubinstein norm on that space [129].

For a discrete signed measure $\alpha - \beta = \sum_k r_k \delta_{z_k}$ with $\sum_k r_k = 0$, (5.11) becomes the finite-dimensional linear program

$$\mathcal{W}_1(\alpha, \beta) = \max_{(f_k)_k} \left\{ \sum_k f_k r_k \;;\; \forall\, k, \ell,\ |f_k - f_\ell| \leqslant d(z_k, z_\ell) \right\}. \tag{5.12}$$

This linear program can be solved by generic interior-point or first-order methods; structured graph versions admit the flow formulations described below. 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 closed-form cumulative formula introduced earlier.

**$\mathcal{W}_1$ on Euclidean spaces.** In the special case of Euclidean spaces $\mathcal{X} = \mathcal{Y} = \mathbb{R}^d$, using $c(x, y) = \|x - y\|$, the global Lipschitz constraint in the Kantorovich–Rubinstein formula can be made local as a uniform bound on the gradient of $f$,

$$\mathcal{W}_1(\alpha, \beta) = \sup_f \left\{ \int_{\mathbb{R}^d} f(\mathrm{d}\alpha - \mathrm{d}\beta) \;;\; \|\nabla f\|_\infty \leqslant 1 \right\}. \tag{5.13}$$

Here the constraint $\|\nabla f\|_\infty \leqslant 1$ signifies that the norm of the gradient of $f$ at any point $x$ is upper bounded by 1, $\|\nabla f(x)\|_2 \leqslant 1$ for any $x$.

Considering the dual problem to (5.13), denoting $\xi \coloneqq \alpha - \beta$, and using the equivalent form

$$-\iota_{\|\cdot\|_{\mathbb{R}^d} \leqslant 1}(u) = \inf_v \langle u,\, v \rangle + \|v\|_{\mathbb{R}^d},$$

one has a maximization on flow vector fields $s : \mathbb{R}^d \to \mathbb{R}^d$

$$\begin{aligned}\mathcal{W}_1(\alpha,\beta) &= \sup_f \inf_{s(x)\in\mathbb{R}^d} \int_{\mathbb{R}^d} f\mathrm{d}\xi + \int \langle \nabla f(x),\, s(x)\rangle \mathrm{d}x + \int \|s(x)\|_{\mathbb{R}^d}\mathrm{d}x \\ &= \inf_{s(x)\in\mathbb{R}^d} \int \|s(x)\|\mathrm{d}x + \sup_f \int f(x)(\mathrm{d}\xi - \mathrm{div}(s)\mathrm{d}x)\end{aligned}$$

one obtains an optimization problem under a fixed divergence constraint

$$\mathcal{W}_1(\alpha,\beta) = \inf_s \left\{ \int_{\mathbb{R}^d} \|s(x)\|_{\mathbb{R}^d}\mathrm{d}x \;;\; \mathrm{div}(s) = \alpha - \beta \right\}, \tag{5.14}$$

which is often called the Beckmann formulation [16]. Here the vectorial function $s(x) \in \mathbb{R}^d$ can be interpreted as a flow field, describing locally the movement of mass. Outside the support of the two input measures, $\mathrm{div}(s) = 0$, which is the conservation of mass constraint. Once properly discretized using finite elements, Problems (5.13) and (5.14) become a nonsmooth convex optimization problem.

The previous formulations (5.13) and (5.14) of $\mathcal{W}_1$ can be generalized to the setting where $\mathcal{X}$ is a Riemannian manifold, i.e. $c(x,y) = d(x,y)$ where $d$ is the associated geodesic distance (and then for smooth manifolds, the gradient and divergence should be understood as differential operators on manifolds).

**Definition 5.7** (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 $\gamma$ joining $i$ to $j$.

This graph distance turns $\mathcal{W}_1$ into a finite-dimensional flow problem.

**Proposition 5.8** ($\mathcal{W}_1$ 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 $\mathrm{a}, \mathrm{b}$ on $V$, set $r = \mathrm{a} - \mathrm{b}$ and orient each edge $e = (i, j)$. If*

$$(\nabla_G f)_e = f_j - f_i, \qquad \mathrm{div}_G = -\nabla_G^*$$

*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,*

$$\mathcal{W}_{1,G}(\mathrm{a},\mathrm{b}) = \max_f \left\{ \sum_{i\in V} f_i r_i \;;\; |f_i - f_j| \leqslant \ell_e \quad \forall e = (i,j) \right\} = \min_m \left\{ \sum_{e\in E} \ell_e |m_e| \;;\; \mathrm{div}_G\, m = r \right\}.$$

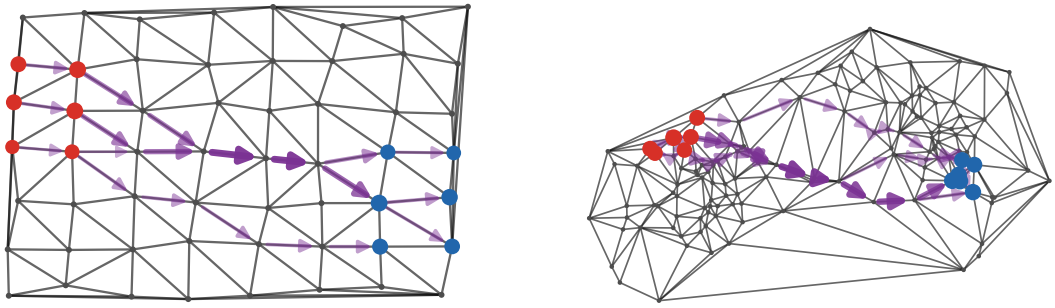
*The vector $m_e$ is an oriented edge flow, and the constraint $\mathrm{div}_G\, m = r$ is conservation of mass at each vertex.*

*Proof.* The edge constraint $|f_i - f_j| \leqslant \ell_e$ implies, by summing along paths, that $|f_i - f_j| \leqslant d_G(i,j)$ for all vertices. Conversely, any 1-Lipschitz function for $d_G$ satisfies the edge constraints because each edge is a path of length $\ell_e$. 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 $f$:

$$\inf_m \sum_e \ell_e |m_e| + \sup_f \sum_i f_i (r_i - (\mathrm{div}_G\, m)_i).$$

Using $\mathrm{div}_G = -\nabla_G^*$, 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 non-empty feasible set: connectedness and $\sum_i r_i = 0$ allow the signed surplus to be routed along paths. This proves the graph Beckmann formula. □



quasi-regular Delaunay graph denser nonuniform Delaunay graph

*Figure 5.3:* 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 $r = \alpha - \beta$, 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 5.9 (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 $\ell_e$, and impose the node balances $\sum_j u_{ij} - \sum_j u_{ji} = r_i$. 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 [27, 175]. 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.

**Algorithm 5.4** Graph Beckmann network-simplex pivot

**Input:** Graph $G = (V, E)$, edge lengths $\ell_e$, node balances $r_i$ with $\sum_i r_i = 0$.
**Output:** Minimum-cost graph flow $u$.
**Replace** each undirected edge by two directed arcs.
**Impose balances:** $\sum_j u_{ij} - \sum_j u_{ji} = r_i$.
**Initialize:** Add artificial root arcs and compute a feasible tree flow on a spanning tree $T$ with node potentials.
**While** $\min_{e \notin T} \bar{c}_e < 0$ **do**:
  **Set** entering arc $e$ to the first minimizer of $\bar{c}_a$ over $a \notin T$ in the prescribed arc order.
  **Add** it to the tree.
  **Set** $C$ = unique induced cycle, oriented in the direction of the entering arc $e$.
  **Set** $\theta = \min\{u_a : a \in C^-\}$, where $C^-$ are the arcs opposed to the cycle orientation.
  **Update** $u_a \leftarrow u_a + \theta$ for $a \in C^+$ and $u_a \leftarrow u_a - \theta$ for $a \in C^-$.
  **Remove** the first arc in $C^-$ attaining the minimum $\theta$, using the prescribed cycle order.
  **Update** the tree, potentials, and reduced costs.
**Return** $u$.

This graph formulation is the transshipment version of $\mathcal{W}_1$. It is the natural discrete analogue of (5.14): 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 $\mathcal{W}_1$ [16, 183].

CHAPTER 6

# Divergences and Dual Norms

This chapter compares OT with divergence-based and adversarial ways of measuring discrepancy. The main stake is topological: $\varphi$-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 [67, 3] with modern integral probability metrics and generative modeling [216, 109, 11].

## 6.1 Dual norms (Integral Probability Metrics)

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 (5.13) is a special case of a dual norm. This viewpoint designs "weak" discrepancies by testing signed differences of measures against a controlled class of functions.

**Definition 6.1** (Dual norm and integral probability metric)**.** For a symmetric convex set $B$ of measurable functions, define on signed measures $\xi$

$$\|\xi\|_B := \sup_f \left\{ \int_{\mathcal{X}} f(x) \mathrm{d}\xi(x) \;;\; f \in B \right\}. \tag{6.1}$$

When this quantity is applied to $\alpha - \beta$ for probability measures, it is often called an integral probability metric.

The choice of the test-function class $B$ determines both the topology and the statistical behavior of the discrepancy; see [217, 216, 218].

**Example 6.2 (Total variation).** As recalled in Definition 2.6 and Proposition 2.7, 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 $\varphi$-divergence and a dual norm; see [216].

**Example 6.3 ($\mathcal{W}_1$ norm).** $\mathcal{W}_1$, as defined in (5.13), is the dual norm (6.1) associated with

$$B = \left\{ f \;;\; \mathrm{Lip}(f) \leqslant 1 \right\}$$

the set of 1-Lipschitz functions.

**Example 6.4 (Flat norm and Dudley metric).** If the set $B$ is bounded and separates measures, then $\|\cdot\|_B$ is a norm on the whole space $\mathcal{M}(\mathcal{X})$ of finite measures. This is not the case of $\mathcal{W}_1$, which is only finite on signed measures $\xi$ such that $\int_{\mathcal{X}} \mathrm{d}\xi = 0$; otherwise $\|\xi\|_B = +\infty$ 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,

$$B = \left\{ f \;;\; \mathrm{Lip}(f) \leqslant 1 \quad \text{and} \quad \|f\|_\infty \leqslant 1 \right\}. \tag{6.2}$$

On compact metric spaces, it metrizes weak convergence on the whole space $\mathcal{M}(\mathcal{X})$ of finite measures. 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 [118] and has been used as a fidelity term for inverse problems in imaging [140]. The flat norm is similar to the Dudley metric, which uses

$$B = \{ f \;;\; \|\nabla f\|_\infty + \|f\|_\infty \leqslant 1 \}.$$

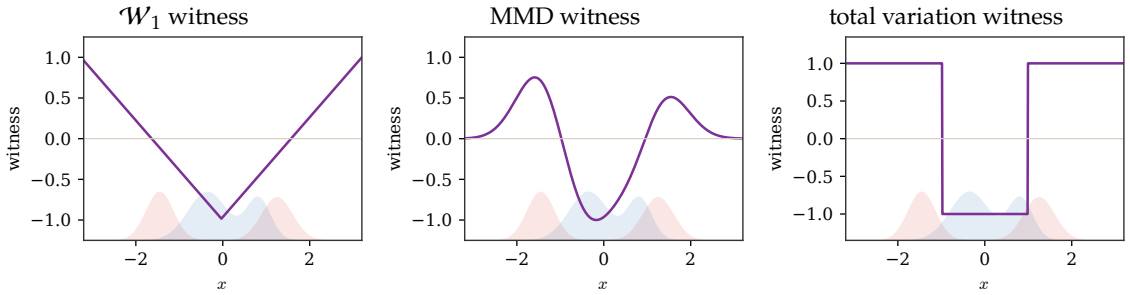



*Figure 6.1:* Dual witnesses for integral probability metrics. The red and blue curves are two one-dimensional probability densities and the violet curve is a normalized optimal dual witness $f^\star_{\alpha,\beta}$ for the IPM variational problem (6.1). $\mathcal{W}_1$ restricts the slope through Kantorovich–Rubinstein duality (5.13), MMD restricts the RKHS norm as in Proposition 6.10, 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 6.5** (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* 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

$$|\langle f,\, \alpha_n - \alpha\rangle| \leqslant \|\alpha_n - \alpha\|_B \quad \text{for} \quad f \in B.$$

By linearity, integrals converge for every $h \in \mathrm{Span}(B)$. Let $u \in C(\mathcal{X})$ and choose $h \in \mathrm{Span}(B)$ with $\|u - h\|_\infty \leqslant \eta$. Since $\alpha_n$ and $\alpha$ 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 6.6** (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 6.5.

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 6.5 gives $\mathcal{W}_1(\alpha_n, \alpha) \to 0$. Proposition 3.37 shows that all $\mathcal{W}_p$ distances induce the same topology on a compact space, so the result follows for every $p \geqslant 1$. □

## 6.2 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 norms 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 6.7** (Positive and conditionally positive 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 $r \in \mathbb{R}^n$,

$$\sum_{i,j=1}^{n} r_i r_j k(x_i, x_j) \geqslant 0. \tag{6.3}$$

It is conditionally positive definite if the same inequality is required only for zero-sum vectors, $\langle r, \mathbb{1}_n \rangle = 0$.

The conditional version is the right notion for probability distances, because one applies the quadratic form to signed measures $\xi = \alpha - \beta$ of total mass zero. Adding a term of the form $a(x) + a(y)$ to the kernel does not change $\iint k(x, y)\mathrm{d}\xi(x)\mathrm{d}\xi(y)$ on such measures, and many natural distance kernels are only conditionally positive definite.

**Example 6.8 (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 [22]. 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 the squared energy distance up to a dimensional constant [204, 222].

Shifted kernels replace $(-\Delta)^{-s}$ by $(-\Delta + \lambda I)^{-s}$ 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 [233]. 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 6.9** (Kernel norm 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 $\xi$ with finite kernel energy, the associated norm is

$$\|\xi\|_k^2 := \iint_{\mathcal{X}\times\mathcal{X}} k(x, y)\mathrm{d}\xi(x)\mathrm{d}\xi(y). \tag{6.4}$$

For two probability measures, the maximum mean discrepancy associated with $k$ is

$$\mathrm{MMD}_k(\alpha, \beta) := \|\alpha - \beta\|_k.$$

These norms are usually called maximum mean discrepancies in statistics and machine learning [112, 169], and kernel norms in shape analysis [123]. If $X, X'$ are independent with law $\alpha$, then $\|\alpha\|_k^2 = \mathbb{E}_{X,X'}(k(X, X'))$, whenever this expression is finite.

**Proposition 6.10** (Kernel norm 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 $\xi$. Then*

$$\|\xi\|_k = \sup_{\|h\|_{\mathcal{H}} \leqslant 1} \int h(x)\mathrm{d}\xi(x),$$

*so $\|\cdot\|_k$ is the dual norm in the sense of (6.1) associated with the RKHS unit ball.*

*Proof.* By the reproducing property,

$$\int h(x)\mathrm{d}\xi(x) = \left\langle h, \int k(x,\cdot)\mathrm{d}\xi(x)\right\rangle_{\mathcal{H}} = \langle h, m_\xi\rangle_{\mathcal{H}}.$$

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 (6.4). □

**Proposition 6.11** (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 integrals of all RKHS functions converge. For any $h\in C(\mathcal{X})$ and any $\eta>0$, choose $g\in\mathcal{H}$ with $\|h-g\|_\infty\leqslant\eta$. Since $\alpha_n$ and $\alpha$ 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 [23, 123, 205] for more details on RKHS functional spaces.

> **Remark 6.12 (Universal kernels).** The hypothesis in Proposition 6.11 is called universality of the kernel. Equivalently, finite sums of the form $\sum_{i=1}^n a_i k(x_i,\cdot)$ are dense in $C(\mathcal{X})$ for the uniform norm. For translation-invariant kernels on $\mathcal{X}=\mathbb{R}^d$, $k(x,y)=k_0(x-y)$, this is equivalent, in the usual sense on compact sets or with suitable decay assumptions, to the Fourier transform not vanishing on its support [218, 217].

In the special case where $\alpha$ is a discrete measure, one thus has the simple expression

$$\|\alpha\|_k^2 = \sum_{i=1}^n\sum_{i'=1}^n \mathrm{a}_i\mathrm{a}_{i'}\mathrm{k}_{i,i'} = \langle \mathrm{ka},\,\mathrm{a}\rangle \quad \text{where} \quad \mathrm{k}_{i,i'} := k(x_i,x_{i'}).$$

In particular, when $\alpha=\sum_{i=1}^n \mathrm{a}_i\delta_{x_i}$ and $\beta=\sum_{i=1}^n \mathrm{b}_i\delta_{x_i}$ are supported on the same set of points, $\|\alpha-\beta\|_k^2=\langle \mathrm{k}(\mathrm{a}-\mathrm{b}),\,\mathrm{a}-\mathrm{b}\rangle$, so that $\|\cdot\|_k$ is a Euclidean norm (proper if k is positive definite, degenerate otherwise if k is semidefinite) on the simplex $\Sigma_n$. To compute the discrepancy between two discrete measures, one can use

$$\|\alpha-\beta\|_k^2 = \sum_{i,i'}\mathrm{a}_i\mathrm{a}_{i'}k(x_i,x_{i'}) + \sum_{j,j'}\mathrm{b}_j\mathrm{b}_{j'}k(y_j,y_{j'}) - 2\sum_{i,j}\mathrm{a}_i\mathrm{b}_j k(x_i,y_j). \tag{6.5}$$

## 6.3 $\varphi$-divergences

This section develops divergences based on pointwise density ratios. They are computationally simple and statistically classical, but they do not see small spatial displacements between singular measures.

**Definition by density ratios.** We now consider a radically different class of methods to compare distributions, which are simpler to compute ($O(n)$ for discrete distributions) but never metrize weak-* convergence. Note that yet another way is possible, using Bregman divergence, which might metrize weak-* convergence when the associated entropy function is weak-* regular.

**Definition 6.13** (Entropy function)**.** A function $\varphi : \mathbb{R} \to \mathbb{R} \cup \{\infty\}$ is an entropy function if it is lower semicontinuous, convex, $\operatorname{dom}\varphi \subset [0, \infty[$, and satisfies the feasibility condition $\operatorname{dom}\varphi \cap (0, +\infty) \neq \emptyset$. The speed of growth of $\varphi$ at $\infty$ is described by

$$\varphi'_\infty = \lim_{x \to +\infty} \varphi(x)/x \in \mathbb{R} \cup \{\infty\}\,.$$

If $\varphi'_\infty = \infty$, then $\varphi$ grows faster than any linear function and $\varphi$ is said to be *superlinear*. Any entropy function $\varphi$ induces a $\varphi$-divergence (also known as Ciszár divergence [67, 3] or $f$-divergence) as follows.

**Definition 6.14** ($\varphi$-Divergences)**.** Let $\varphi$ be an entropy function. For $\alpha, \beta \in \mathcal{M}(\mathcal{X})$, let $\frac{\mathrm{d}\alpha}{\mathrm{d}\beta}\beta + \alpha^\perp$ be the Lebesgue decomposition of $\alpha$ with respect to $\beta$: this means that $\alpha$ is uniquely decomposed as $\alpha^{\mathrm{ac}} + \alpha^\perp$, with $\alpha^{\mathrm{ac}} \ll \beta$, $\alpha^\perp \perp \beta$, and $\alpha^{\mathrm{ac}} = (\mathrm{d}\alpha/\mathrm{d}\beta)\beta$. The divergence $\mathcal{D}_\varphi$ is defined by

$$\mathcal{D}_\varphi(\alpha|\beta) \coloneqq \int_{\mathcal{X}} \varphi\left(\frac{\mathrm{d}\alpha}{\mathrm{d}\beta}\right) \mathrm{d}\beta + \varphi'_\infty \alpha^\perp(\mathcal{X}) \tag{6.6}$$

if $\alpha, \beta$ are nonnegative and $\infty$ otherwise.

The additional term $\varphi'_\infty \alpha^\perp(\mathcal{X})$ in (6.6) 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 (6.10) defining the TV norm. If $\varphi$ has superlinear growth, e.g.the usual entropy (6.9), then $\varphi'_\infty = +\infty$ so that $\mathcal{D}_\varphi(\alpha|\beta) = +\infty$ if $\alpha$ does not have a density with respect to $\beta$.

In the discrete setting, assuming

$$\alpha = \sum_i \mathrm{a}_i \delta_{x_i} \quad \text{and} \quad \beta = \sum_i \mathrm{b}_i \delta_{x_i} \tag{6.7}$$

are supported on the same set of $n$ points $(x_i)_{i=1}^n \subset \mathcal{X}$, (6.6) defines a divergence on $\Sigma_n$

$$\mathrm{D}_\varphi(\mathrm{a}|\mathrm{b}) = \sum_{i \in \mathrm{Supp}(\mathrm{b})} \varphi\left(\frac{\mathrm{a}_i}{\mathrm{b}_i}\right) \mathrm{b}_i + \varphi'_\infty \sum_{i \notin \mathrm{Supp}(\mathrm{b})} \mathrm{a}_i, \tag{6.8}$$

where $\mathrm{Supp}(\mathrm{b}) \coloneqq \{i \in [\![n]\!]\,;\, \mathrm{b}_i \neq 0\}$.

**Proposition 6.15** (Basic properties of $\varphi$-divergences)**.** *If $\varphi$ is an entropy function, then $\mathcal{D}_\varphi$ is jointly 1-homogeneous, convex and weak-* lower semicontinuous in $(\alpha, \beta)$.*

*Proof.* One defines the associated perspective function

$$\forall\,(u, v) \in (\mathbb{R}_+)^2, \quad \psi(u, v) = \begin{cases} \varphi(u/v)v & \text{if} \quad v \neq 0, \\ u\varphi'_\infty & \text{if} \quad v = 0 \end{cases}$$

The joint 1-homogeneity follows from the definition of this perspective. We prove convexity in the discrete case, where

$$\mathrm{D}_\varphi(\mathrm{a}|\mathrm{b}) = \sum_i \psi(\mathrm{a}_i, \mathrm{b}_i),$$

and it is enough to show that $\psi$ is convex on $(\mathbb{R}_+)^2$. We first prove this on $\mathbb{R}_+ \times \mathbb{R}_+^*$; the extension to $v = 0$ follows by lower semicontinuity of the recession value $u\varphi'_\infty$. Indeed, for any $\lambda \in [0, 1]$, $\tau = 1 - \lambda$, set

$$\theta_1 = \frac{\tau v_1}{\tau v_1 + \lambda v_2}, \qquad \theta_2 = \frac{\lambda v_2}{\tau v_1 + \lambda v_2}.$$

Then $\theta_1 + \theta_2 = 1$ and

$$\frac{\tau u_1 + \lambda u_2}{\tau v_1 + \lambda v_2} = \theta_1 \frac{u_1}{v_1} + \theta_2 \frac{u_2}{v_2}.$$

Convexity of $\varphi$ therefore gives

$$\varphi\left(\frac{\tau u_1 + \lambda u_2}{\tau v_1 + \lambda v_2}\right)(\tau v_1 + \lambda v_2) \leqslant \tau v_1 \varphi\left(\frac{u_1}{v_1}\right) + \lambda v_2 \varphi\left(\frac{u_2}{v_2}\right).$$

In the general measure case, weak-* lower semicontinuity is the standard lower-semicontinuity theorem for convex integral functionals with recession extension; in the discrete case it is immediate from the lower semicontinuity of $\psi$. □

The following proposition records when $\mathcal{D}_\varphi$ is nonnegative.

**Proposition 6.16** (Non-negativity of $\varphi$-divergences)**.** *Assume that $\varphi$ 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 $\varphi$ is strictly convex, then one has $\mathcal{D}_\varphi(\alpha|\beta) = 0$ if and only if $\alpha = \beta$. This property extends to arbitrary distributions $(\alpha, \beta) \in \mathcal{M}_+(\mathcal{X})$ if one furthermore imposes that $\varphi \geqslant 0$.*

*Proof.* Let $m = \alpha + \beta$ and write $a = \frac{\mathrm{d}\alpha}{\mathrm{d}m}$ and $b = \frac{\mathrm{d}\beta}{\mathrm{d}m}$. Using the perspective function $\psi$ from the previous proof,

$$\mathcal{D}_\varphi(\alpha|\beta) = \int \psi(a, b)\mathrm{d}m.$$

Since $\alpha$ and $\beta$ are probabilities, $\int a\mathrm{d}m = \int b\mathrm{d}m = 1$. Jensen's inequality and $\psi(1, 1) = \varphi(1) = 0$ give

$$\mathcal{D}_\varphi(\alpha|\beta) \geqslant \psi\left(\int a\mathrm{d}m, \int b\mathrm{d}m\right) = 0.$$

If $\varphi$ is strictly convex, equality in Jensen forces $a = b$ $m$-almost everywhere, hence $\alpha = \beta$. In the general non-probability case, if $\varphi \geqslant 0$ then the divergence is positive by construction. □

**Classical examples and topology.** The following examples calibrate the strength of $\varphi$-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 6.17 (Kullback–Leibler divergence).** The Kullback–Leibler divergence $\mathrm{KL} := \mathcal{D}_{\varphi_{\mathrm{KL}}}$, also known as the relative entropy, was already introduced in (7.12) and (7.6). It is the divergence associated to the Shannon–Boltzmann entropy function $\varphi_{\mathrm{KL}}$, given by

$$\varphi_{\mathrm{KL}}(s) = \begin{cases} s\log(s) - s + 1 & \text{for } s > 0, \\ 1 & \text{for } s = 0, \\ +\infty & \text{otherwise.} \end{cases} \tag{6.9}$$

**Example 6.18 (Total variation).** The total variation distance $\mathrm{TV} := \mathcal{D}_{\varphi_{\mathrm{TV}}}$ is the divergence associated to

$$\varphi_{\mathrm{TV}}(s) = \begin{cases} |s - 1| & \text{for } s \geqslant 0, \\ +\infty & \text{otherwise.} \end{cases} \tag{6.10}$$

It actually defines a norm on the full space of measures $\mathcal{M}(\mathcal{X})$ where

$$\mathrm{TV}(\alpha|\beta) = \|\alpha - \beta\|_{\mathrm{TV}}, \quad \text{where} \quad \|\alpha\|_{\mathrm{TV}} = |\alpha|(\mathcal{X}) = \int_{\mathcal{X}} \mathrm{d}|\alpha|(x). \tag{6.11}$$

If $\alpha$ has a density $\rho_\alpha$ 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 $\alpha$ is discrete as in (6.7), then the TV norm is the $\ell^1$ norm of vectors in $\mathbb{R}^n$, $\|\alpha\|_{\mathrm{TV}} = \sum_i |\mathrm{a}_i| = \|\mathrm{a}\|_{\ell^1}$.

**Remark 6.19 (Strong vs. weak topology).** The total variation norm (6.11) defines the so-called "strong" topology on the space of measures. On a compact domain $\mathcal{X}$ of radius $R$, one has

$$\mathcal{W}_1(\alpha,\beta) \leqslant R\|\alpha-\beta\|_{\mathrm{TV}}$$

so that this strong notion of convergence implies the weak convergence metrized by Wasserstein distances. The converse is, however, not true, since $\delta_x$ does not converge strongly to $\delta_y$ if $x \to y$ (note that $\|\delta_x - \delta_y\|_{\mathrm{TV}} = 2$ if $x \neq y$). A chief advantage is that $\mathcal{M}_+^1(\mathcal{X})$ (once again on a compact ground space $\mathcal{X}$) is compact for the weak topology so that from any sequence of probability measures $(\alpha_k)_k$, one can always extract a converging subsequence, which makes it a suitable space for several optimization problems.

**Main families of $\varphi$-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 between the Pearson $\chi^2$ divergence at $\gamma = 2$, the Hellinger-type behavior at $\gamma = 1/2$, and, by taking limits, the KL divergence as $\gamma \to 1$ and the reverse KL or Burg entropy $\varphi_0(s) = -\log s + s - 1$ as $\gamma \to 0$. The Hellinger divergence is often written separately with $\varphi_H(s) = (\sqrt{s}-1)^2$, giving $\mathfrak{h}(\alpha,\beta) = \|\sqrt{\rho_\alpha} - \sqrt{\rho_\beta}\|_{L^2}$ when both measures have densities. The Jensen–Shannon divergence is the symmetrized and 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 is generated by a bounded entropy equivalent to $\varphi_{\mathrm{JS}}(s) = s\log s - (s+1)\log((s+1)/2)$ up to an irrelevant affine term. Total variation corresponds to the non-smooth entropy $\varphi_{\mathrm{TV}}(s) = |s-1|$ and is exceptional because it is both a $\varphi$-divergence and an integral probability metric.

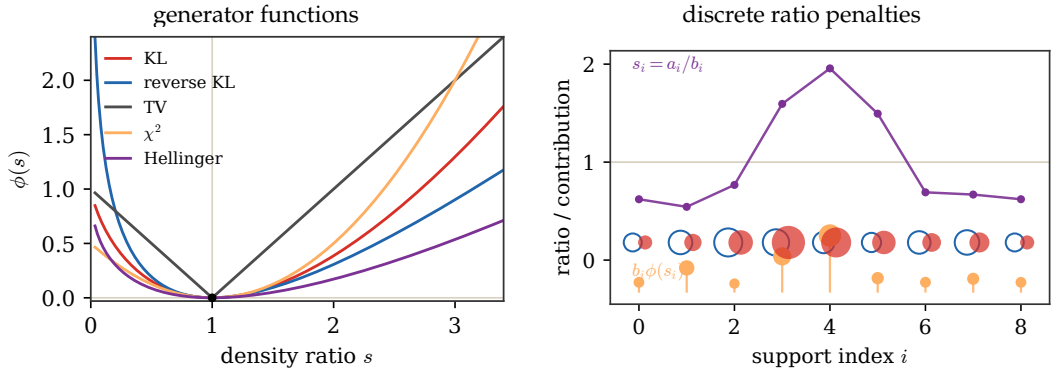



*Figure 6.2:* $\varphi$-divergences through density ratios. The left panel shows normalized generators for common divergences as functions of $s = \mathrm{d}\alpha/\mathrm{d}\beta$; all curves vanish at $s = 1$ up to affine normalization. The right panel shows the discrete formula $D_\varphi(a|b) = \sum_i b_i \varphi(a_i/b_i)$: 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 6.20 ($\varphi$-divergences versus Bregman divergences).** Except for KL-type entropies, $\varphi$-divergences should not be confused with Bregman divergences. A $\varphi$-divergence compares measures pointwise through the density ratio $\mathrm{d}\alpha/\mathrm{d}\beta$ and is invariant under measurable changes of variables. A Bregman divergence is generated by a convex functional on a linear space and compares two points through 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 $\varphi$-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 $\varphi$-divergences, of the Kantorovich dual formula for transport costs.

**Proposition 6.21** (Dual expression). *A $\varphi$-divergence can be expressed using the Legendre transform*

$$\varphi^{*,\geqslant 0}(s) \coloneqq \sup_{t\in\mathbb{R}^+} st - \varphi(t)$$

*(notice that we restrict the function to the positive real) of $\varphi$ as*

$$\mathcal{D}_\varphi(\alpha|\beta) = \sup_{f:\mathcal{X}\to\mathbb{R}} \int_{\mathcal{X}} f(x)\mathrm{d}\alpha(x) - \int_{\mathcal{X}} \varphi^{*,\geqslant 0}(f(x))\mathrm{d}\beta(x). \tag{6.12}$$

*which equivalently reads that the Legendre transform of $\mathcal{D}_\varphi(\cdot|\beta)$ reads*

$$\forall\, f \in C(\mathcal{X}), \quad \mathcal{D}^*_\varphi(f|\beta) = \int_{\mathcal{X}} \varphi^{*,\geqslant 0}(f(x))\mathrm{d}\beta(x). \tag{6.13}$$

*Proof.* We first consider the superlinear case $\varphi'_\infty = +\infty$, so that $\mathcal{D}_\varphi(\alpha|\beta) = +\infty$ if $\alpha$ does not have a density $\rho \geqslant 0$ with respect to $\beta$, $\mathrm{d}\alpha = \rho\mathrm{d}\beta$. Thus the Legendre-Fenchel transform of $\mathcal{D}_\varphi(\cdot|\beta)$ reads

$$\begin{aligned}\mathcal{D}^*_\varphi(f|\beta) &= \sup_{\rho\geqslant 0} \int_{\mathcal{X}} f(x)\rho(x)\mathrm{d}\beta(x) - \int_{\mathcal{X}} \varphi(\rho(x))\mathrm{d}\beta(x) \\ &= \int_{\mathcal{X}} \sup_{\rho(x)\geqslant 0} \left(f(x)\rho(x) - \varphi(\rho(x))\right)\mathrm{d}\beta(x) = \int_{\mathcal{X}} \varphi^{*,\geqslant 0}(f(x))\mathrm{d}\beta(x).\end{aligned}$$

Fenchel–Moreau then gives the displayed dual expression. For a general entropy, the same argument is applied to the perspective with its recession term; the singular part is exactly encoded by the effective domain of $\varphi^{*,\geqslant 0}$. □

## 6.4 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 $\beta = \frac{1}{m}\sum_j \delta_{y_j}$, where $\zeta \in \mathcal{M}^1_+(\mathcal{Z})$ is a fixed density over the latent space and $g_\theta : \mathcal{Z} \to \mathcal{X}$ is the generator, often a neural network.

**Divergence-based adversarial losses.** Any $\varphi$-divergence can be written in adversarial form through the dual formula (6.12):

$$\min_\theta \mathcal{D}_\varphi(\alpha_\theta|\beta) = \min_\theta \sup_f \int_{\mathcal{X}} f(x)\mathrm{d}\alpha_\theta(x) - \mathcal{D}^*_\varphi(f|\beta) = \min_\theta \sup_f \int_{\mathcal{X}} f(g_\theta(z))\mathrm{d}\zeta(z) - \frac{1}{m}\sum_j \varphi^*(f(y_j)).$$

Replacing the unrestricted potential $f$ by a neural network $f_\xi$ gives a saddle problem

$$\min_\theta \max_\xi \int_{\mathcal{Z}} f_\xi(g_\theta(z))\mathrm{d}\zeta(z) - \frac{1}{m}\sum_j \varphi^*(f_\xi(y_j)).$$

The original vanilla GAN [109] is this construction for the Jensen–Shannon generator discussed above,

$$\varphi_{\mathrm{JS}}(s) = s\log s - (s+1)\log\frac{s+1}{2}, \qquad \varphi^*_{\mathrm{JS}}(u) = -\log(2 - e^u), \quad u < \log 2,$$

up to affine normalizations and the usual reparametrization of the potential by a discriminator with values in $(0,1)$. In practice the min–max problem is solved by alternating stochastic gradient descent/ascent on $(\theta,\xi)$. Unlike the convex-concave variational formula, the neural parametrization is non-convex in $\theta$ and non-concave in $\xi$, which explains the instability and mode-collapse pathologies of divergence-based GAN training. These losses estimate density ratios, which is statistically meaningful when the measures overlap but can saturate when the model and data are mutually singular; for example, the Jensen–Shannon divergence is maximal for disjoint supports.

**Dual norms and integral probability metrics.** Instead of a density-ratio divergence, one can minimize a dual norm (6.1), also called an integral probability metric,

$$\min_{\theta} \|\alpha_\theta - \beta\|_B = \min_{\theta} \sup_{f \in B} \int_{\mathcal{X}} f(x) \mathrm{d}(\alpha_\theta - \beta)(x) = \min_{\theta} \sup_{f \in B} \int_{\mathcal{Z}} f(g_\theta(z)) \mathrm{d}\zeta - \frac{1}{m} \sum_j f(y_j).$$

MMD-GANs take $B$ to be a unit ball in an RKHS [85]; Wasserstein GANs take $B$ to be a Lipschitz ball, following Kantorovich–Rubinstein duality [11, 95]. The advantage of such choices is topological: for bounded continuous RKHS balls, or for bounded Lipschitz balls on compact spaces, the resulting objective is weak-* continuous. It can therefore 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, either by a kernel norm, a Lipschitz constraint or a related regularization.

> **Remark 6.22 (Weight clipping is only a proxy).** Wasserstein GANs originally used weight clipping, constraining $\|\xi\|_\infty \leqslant 1$ as a proxy for enforcing $f_\xi \in B = \{f \ ; \ \mathrm{Lip}(f) \leqslant 1\}$. This parameter set is both smaller than the true Lipschitz ball and non-convex, so clipping should be understood as a practical heuristic rather than a faithful implementation of the Kantorovich–Rubinstein dual constraint.

CHAPTER 7

# 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, explains the path-space Schrödinger problem behind the static coupling formulation, develops the dual soft-transform picture, and presents the main convex regularization variants and the debiased Sinkhorn divergence. The presentation connects the older matrix-scaling literature [209, 211, 210] with modern entropic OT [70, 185].

## 7.1 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.

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 7.1** (Discrete Shannon–Boltzmann entropy)**.** For a nonnegative matrix $P$, its Shannon–Boltzmann entropy is

$$H(P) := -\sum_{i,j} P_{i,j} \log(P_{i,j}),$$

with the convention $0 \log(0) = 0$.

Using this entropy as a regularizing function gives approximate solutions to the original transport problem (3.2)

$$L_C^\varepsilon(a, b) := \min_{P \in U(a,b)} \langle P, C \rangle - \varepsilon H(P). \tag{7.1}$$

**Proposition 7.2** (Existence and uniqueness of entropic OT)**.** *Assume that* $a, b$ *are probability histograms and that* $C$ *is finite. For every* $\varepsilon > 0$*, problem* (7.1) *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 $U(a, b)$ is non-empty and compact, and the objective is continuous on it with the convention $0 \log 0 = 0$, so a minimizer exists. On the relative interior of the polytope,

$$-\partial^2 H(P) = \operatorname{diag}(1/P_{i,j})$$

is positive definite on every non-zero feasible direction. Hence $-H$ is strictly convex on the polytope and $\langle P, C \rangle - \varepsilon H(P)$ is strictly convex, which implies uniqueness.

If $a_i, b_j > 0$ and a minimizer had $P_{i,j} = 0$, then for small $t > 0$ the perturbation $P_t = (1 - t)P + t\, a \otimes b$ 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 7.2, problem (7.1) has a unique optimal solution. This smoothing, beyond providing uniqueness, actually leads to $L_C^\varepsilon(a, b)$ being a smooth function of $a$, $b$ and $C$ whenever these variables stay in the relative interior of their domains. In finite dimension, this follows from strict convexity and the envelope theorem applied to the dual problem. The effect of the entropy is to act as a barrier function for the positivity constraint. As we will show later, this forces the solution $P$ to be strictly positive on the support of $a \otimes b$. We will also show that as $\varepsilon \to +\infty$, the solution satisfies $P \to a \otimes b$.

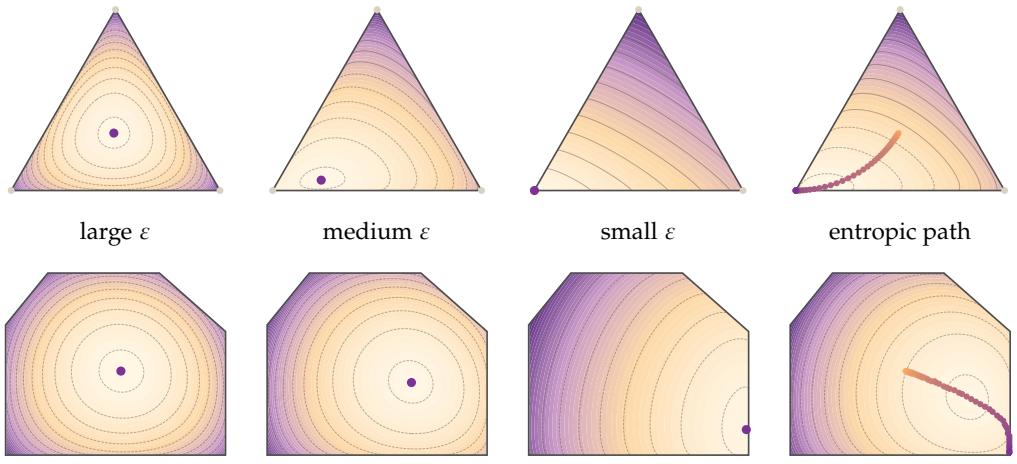



*Figure 7.1:* Entropic regularization and slack barriers. The first row shows the penalized objective $\langle c, p\rangle + \varepsilon \sum_i p_i(\log p_i - 1)$ on a triangular face of the transport polytope; color and level sets represent the regularized functional itself, not only the linear part. 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 $\varepsilon$ selects an interior reference point, while small $\varepsilon$ moves the minimizer toward a low-cost face.

**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 7.1, 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 [172]. 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.

## 7.2 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 [209, 211, 210]. Its modern role in OT was transformed by Cuturi's entropic formulation [70]: 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 (7.1) has a specific form, which can be parameterized using $n + m$ variables. That parameterization is therefore essentially dual, in the sense that a coupling P in $U(a, b)$ has $nm$ variables but $n + m$ constraints.

**Proposition 7.3** (Scaling form of entropic OT)**.** P *is the unique solution to* (7.1) *if and only if there exists* $(u, v) \in \mathbb{R}_+^n \times \mathbb{R}_+^m$ *such that*

$$\forall (i, j) \in [\![n]\!] \times [\![m]\!], \quad P_{i,j} = u_i K_{i,j} v_j \quad \textit{where} \quad K_{i,j} := e^{-\frac{C_{i,j}}{\varepsilon}}, \tag{7.2}$$

*and* $P \in \mathcal{U}(a, b)$.

*Proof.* Without loss of generality, we assume $a_i, b_j > 0$ (otherwise, rows or columns with zero mass are fixed to zero and can be removed). By Proposition 7.2, 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 (7.1) reads

$$\mathcal{E}(P, f, g) = \langle P, C\rangle + \varepsilon \sum_{i,j} P_{i,j} \log(P_{i,j}) + \langle f, a - P\mathbb{1}_m\rangle + \langle g, b - P^\top \mathbb{1}_n\rangle.$$

Considering first-order conditions (where we ignore the positivity constraint as explained above), we have

$$\frac{\partial \mathcal{E}(P, f, g)}{\partial P_{i,j}} = C_{i,j} + \varepsilon(\log\left(P_{i,j}\right) + 1) - f_i - g_j = 0.$$

which results, in an optimal P coupling of the regularized problem, in the expression $P_{i,j} = e^{\frac{f_i + g_j - C_{i,j}}{\varepsilon} - 1}$ which can be rewritten in the form provided in the proposition using non-negative vectors $u_i := e^{f_i/\varepsilon - 1}$ and $v_j := e^{g_j/\varepsilon}$. □

The factorization of the optimal solution exhibited in Equation (7.2) can be conveniently rewritten in matrix form as $P = \mathrm{diag}(u) K \,\mathrm{diag}(v)$. $u, v$ must therefore satisfy the following nonlinear equations which correspond to the mass conservation constraints inherent to $U(a, b)$,

$$\mathrm{diag}(u) K \,\mathrm{diag}(v)\mathbb{1}_m = a, \quad \text{and} \quad \mathrm{diag}(v) K^\top \,\mathrm{diag}(u)\mathbb{1}_n = b, \tag{7.3}$$

These two equations can be further simplified, since $\mathrm{diag}(v)\mathbb{1}_m$ is $v$, and the multiplication of $\mathrm{diag}(u)$ times $Kv$ is

$$u \odot (Kv) = a \quad \text{and} \quad v \odot (K^\top u) = b \tag{7.4}$$

where $\odot$ corresponds to the entry-wise multiplication of vectors. This problem is known in the numerical analysis community as the matrix scaling problem (see [171] and references therein).

The problem of normalizing a positive matrix K has a long history, from iterative proportional fitting in statistics and economics [135, 236, 97] to modern matrix balancing algorithms [125, 68]. The problem of normalizing a positive matrix K by diagonal scaling is well known, in particular when $n = m$ and a and b are uniform. This corresponds to diagonal scaling toward bistochasticity, which is a very old problem. The previous result shows that there is a unique such scaled matrix P, thanks to the strong convexity of the regularized problem. The remaining question is how to compute this scaled matrix in practice. If some entries of K vanish (equivalently, if the cost matrix C can have infinite values), additional support conditions are needed; here we focus on the strictly positive case.

An intuitive way to try to solve these equations is to solve them iteratively, by modifying first $u$ so that it satisfies the left-hand side of Equation (7.4) and then $v$ to satisfy its right-hand side. These two updates define Sinkhorn's algorithm

$$u^{(\ell+1)} := \frac{a}{Kv^{(\ell)}} \quad \text{and} \quad v^{(\ell+1)} := \frac{b}{K^\top u^{(\ell+1)}}, \tag{7.5}$$

initialized with an arbitrary positive vector, for instance $v^{(0)} = \mathbb{1}_m$. The division operator used above between two vectors is to be understood entry-wise. Note that a different initialization will likely lead to a different solution for $u, v$, since $u, v$ are only defined up to a multiplicative constant (if $u, v$ satisfy (7.3) then so do $\lambda u, v/\lambda$ for any $\lambda > 0$). 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.

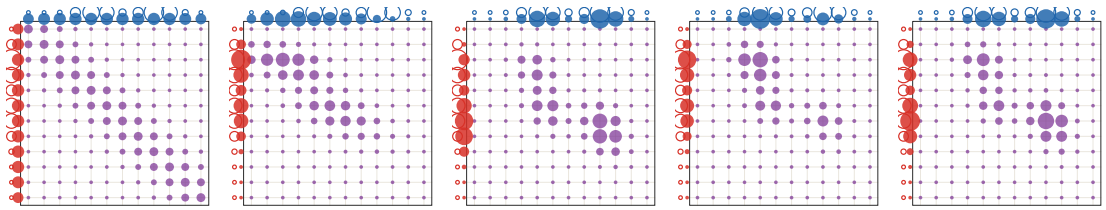



*Figure 7.2:* Marginal constraints during Sinkhorn scaling on a twelve-bin one-dimensional problem. Violet circles represent the current coupling matrix, framed by the thin black box; the red source marginal is displayed on the left and the blue target marginal below the matrix. Hollow side circles show 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.

**Algorithm 7.1** Sinkhorn scaling

**Input:** Weights $a, b$, cost matrix $C$, regularization $\varepsilon > 0$, tolerance tol.
**Output:** Entropic coupling $P$.
**Initialize:** Set $K_{ij} = e^{-C_{ij}/\varepsilon}$, $v^{(0)} = \mathbb{1}_m$, $r_0 = +\infty$, and $k = 0$.
**While** $r_k >$ tol **do**:
  **Set** $k \leftarrow k + 1$.
  $u^{(k)} = \frac{a}{Kv^{(k-1)}}$.
  $v^{(k)} = \frac{b}{K^\top u^{(k)}}$.
  $P^{(k)} = \operatorname{diag}(u^{(k)}) K \operatorname{diag}(v^{(k)})$.
  **Set** $r_k = \max\{\|P^{(k)}\mathbb{1}_m - a\|_1, \|(P^{(k)})^\top \mathbb{1}_n - b\|_1\}$.
**Return** $P^{(k)}$.

The same alternating projection mechanism is clearer on a dense one-dimensional discretization, where the marginal defects appear as continuous side curves.

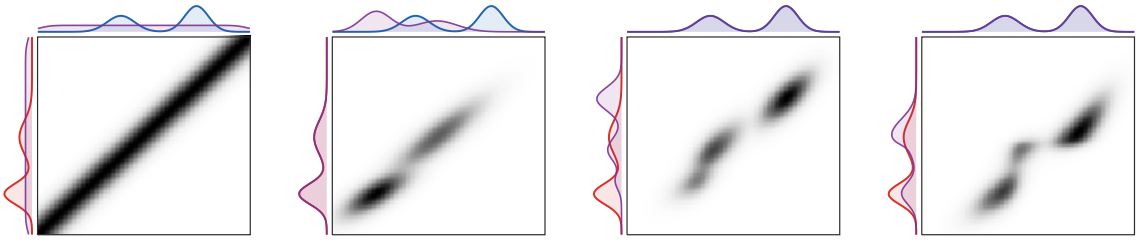



*Figure 7.3:* Dense Sinkhorn scaling for one-dimensional Gaussian-mixture marginals. The grayscale matrix is the current coupling, with a thin box delimiting only the matrix and not the side marginal plots. The red and blue side curves are the prescribed source and target marginals, while the violet 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 $\varepsilon$ produces a concentrated transport band, while larger $\varepsilon$ spreads the same marginals into a smoother coupling.

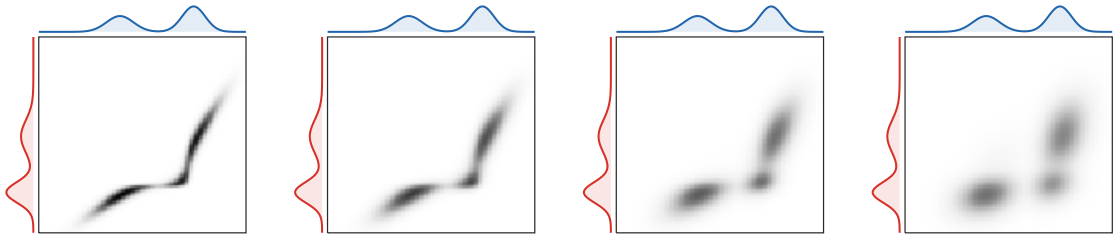



*Figure 7.4:* Final Sinkhorn couplings for the same one-dimensional Gaussian-mixture marginals and four regularization strengths. Each boxed matrix is the converged solution of the KL-regularized problem and uses a common grayscale normalization; the side curves display the fixed source and target marginals. Decreasing $\varepsilon$ sharpens the plan toward an optimal-transport graph, whereas increasing $\varepsilon$ keeps more of the diffuse product-measure structure.

Chapter 8 gives the formal convergence analysis. Before that, the following figure shows how the dual potentials stabilize along the same scaling iteration.

Complexity bounds for Sinkhorn and comparisons with accelerated first-order methods are discussed in [4, 84, 132]. A chief computational advantage of Sinkhorn's algorithm, besides its simplicity, is that the only expensive step is multiplication by the Gibbs kernel. Its complexity therefore scales like $Cnm$, where $C$ is the number of Sinkhorn iterations. Chapter 8 gives a more precise convergence statement: for a fixed regularization strength $\varepsilon$, the entropic dual gap has an $O(1/k)$ bound with constants proportional to $1/\varepsilon$,

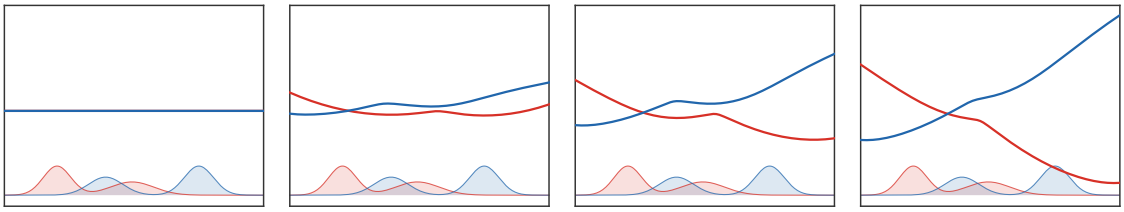



*Figure 7.5:* KL-normalized Sinkhorn dual potentials along the scaling iteration for the same one-dimensional Gaussian-mixture setting as Figure 7.7, with fixed regularization strength $\varepsilon = 0.045$. The bottom silhouettes show both the source histogram a in red and the target histogram b in blue, while the red and blue curves are the logarithmic scaling potentials. All panels share the same axes, making stabilization visible.

while Hilbert-metric arguments give eventual linear convergence when the Gibbs kernel is uniformly positive. To approximate the unregularized OT value to accuracy $\delta$, one must also balance the entropic bias, which is typically $O(\varepsilon)$ in finite dimension. Choosing $\varepsilon$ proportional to $\delta$ and solving the entropic problem to accuracy $O(\delta)$ leads to the familiar iteration scaling of order $1/\delta^2$ for the unregularized value, up to logarithmic and cost-range factors [4, 84].

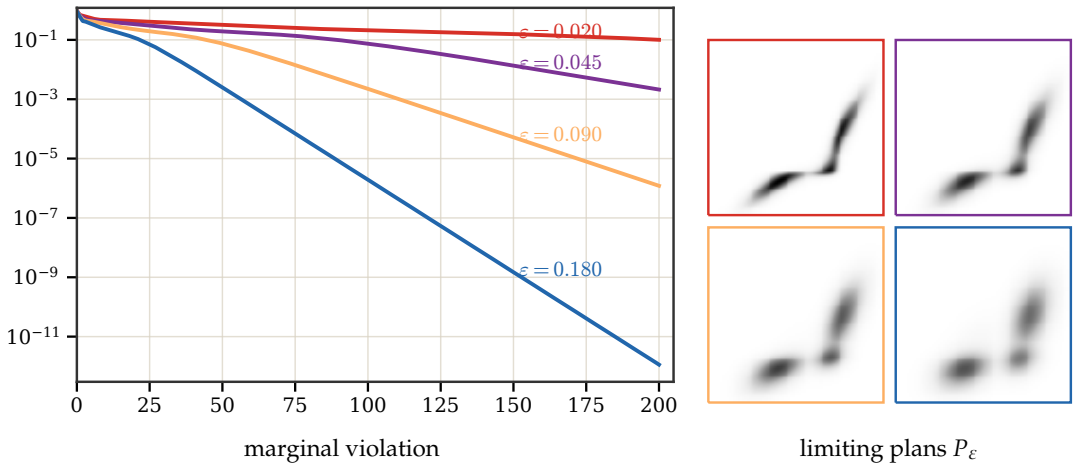



*Figure 7.6:* Marginal violation along Sinkhorn half-steps for four values of $\varepsilon$ on the same one-dimensional Gaussian-mixture problem, together with the corresponding limiting plans. The plotted error is $\frac{1}{2}(\|P_k\mathbb{1} - \mathrm{a}\|_1 + \|P_k^\top \mathbb{1} - \mathrm{b}\|_1)$ on a logarithmic scale. The colored boxes on the right use the same colors as the convergence curves. For fixed positive $\varepsilon$, the curves enter a linear regime; smaller $\varepsilon$ 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. In such settings, $C$ is often moderate. This should be contrasted with generic interior-point methods, which use the logarithmic barrier discussed above and require solving Newton systems. For a dense transportation linear program, these algorithms typically have worst-case complexity of order $O(n^6 \log(1/\delta))$ to reach accuracy $\delta$, up to problem-dependent conditioning factors.

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 7.4 (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

$$O(d\, q^{d+1}) = O(d\, N^{1+1/d})$$

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.

## 7.3 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.

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 7.5** (Discrete relative entropy)**.** For nonnegative matrices $\mathrm{P}, \mathrm{Q}$ of the same size, the generalized relative entropy, or Kullback–Leibler divergence, is

$$\mathrm{KL}(\mathrm{P}|\mathrm{Q}) \coloneqq \sum_{i,j} \mathrm{P}_{i,j} \log\left(\frac{\mathrm{P}_{i,j}}{\mathrm{Q}_{i,j}}\right) - \mathrm{P}_{i,j} + \mathrm{Q}_{i,j}. \tag{7.6}$$

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 $\varphi$-divergence (as defined in Section 6.3) 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 7.6** (Relative entropy is distance-like)**.** *Let* $\mathrm{P}, \mathrm{Q} \in \mathbb{R}_+^{n\times m}$ *have the same total mass and assume* $\mathrm{Q}_{i,j} > 0$ *on the support of* P*. Then* $\mathrm{KL}(\mathrm{P}|\mathrm{Q}) \geqslant 0$*, 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

$$\mathrm{KL}(\mathrm{P}|\mathrm{Q}) = \sum_{i,j} \mathrm{Q}_{i,j}\varphi(\mathrm{P}_{i,j}/\mathrm{Q}_{i,j}) \geqslant 0,$$

with equality only when $\mathrm{P}_{i,j}/\mathrm{Q}_{i,j} = 1$ for all entries with $\mathrm{Q}_{i,j} > 0$. The support convention rules out positive P where $\mathrm{Q} = 0$, so equality is equivalent to $\mathrm{P} = \mathrm{Q}$. □

Equivalently, when P and Q have the same total mass, it reads

$$\mathrm{KL}(\mathrm{P}|\mathrm{Q}) = \sum_{i,j} \varphi(\mathrm{P}_{i,j}/\mathrm{Q}_{i,j})\mathrm{Q}_{i,j}.$$

where $\varphi(s) = s\log(s)$. For any convex $\varphi$ such that $\varphi(1) = 0$, one has indeed by Jensen

$$\sum_{i,j} \varphi(\mathrm{P}_{i,j}/\mathrm{Q}_{i,j})\mathrm{Q}_{i,j} \geqslant \varphi(\sum_{i,j} \mathrm{P}_{i,j}/\mathrm{Q}_{i,j}\mathrm{Q}_{i,j}) = \varphi(\sum_{i,j} \mathrm{P}_{i,j}) = \varphi(1) = 0.$$

For instance, one can use as reference measure the tensor product $\mathrm{a}\otimes\mathrm{b} = (\mathrm{a}_i\mathrm{b}_j)_{i,j}$ and consider

$$\min_{\mathrm{P}\in\mathrm{U}(\mathrm{a},\mathrm{b})} \langle \mathrm{P}, \mathrm{C}\rangle + \varepsilon\mathrm{KL}(\mathrm{P}|\mathrm{a}\otimes\mathrm{b}). \tag{7.7}$$

This normalization will matter again for unbalanced OT, where changing the reference measure is no longer merely a harmless notational choice.

For the balanced problem with fixed positive marginals, however, the choice of tensor-product reference does not affect the selected coupling: it only adds a constant to the objective, as shown in the following proposition. In particular, (7.7) and (7.1) have the same unique solution.

**Proposition 7.7** (Reference measure shift for KL)**.** *After removing zero-mass rows and columns, assume that* $\mathrm{a}, \mathrm{a}' \in \Sigma_n$ *and* $\mathrm{b}, \mathrm{b}' \in \Sigma_m$ *have positive entries. For every* $\mathrm{P} \in \mathrm{U}(\mathrm{a}, \mathrm{b})$*, one has*

$$\mathrm{KL}(\mathrm{P}|\mathrm{a}\otimes\mathrm{b}) = \mathrm{KL}(\mathrm{P}|\mathrm{a}'\otimes\mathrm{b}') - \mathrm{KL}(\mathrm{a}|\mathrm{a}') - \mathrm{KL}(\mathrm{b}|\mathrm{b}').$$

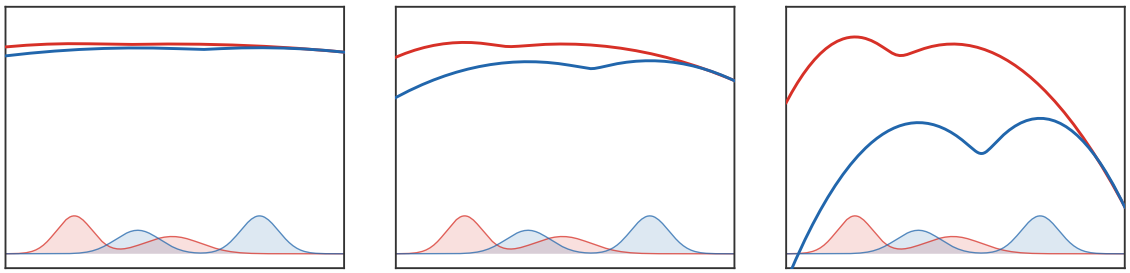
*Consequently, for fixed positive marginals, changing the positive tensor-product reference measure only adds a constant on the transport polytope. In particular,* (7.7) *and* (7.1) *have the same unique minimizer.*

*Proof.* Expanding the logarithm and using the marginal constraints gives

$$\begin{aligned}\mathrm{KL}(\mathrm{P}|\mathrm{a}\otimes\mathrm{b}) &= \mathrm{KL}(\mathrm{P}|\mathrm{a}'\otimes\mathrm{b}') + \sum_i \mathrm{a}_i \log\frac{\mathrm{a}'_i}{\mathrm{a}_i} + \sum_j \mathrm{b}_j \log\frac{\mathrm{b}'_j}{\mathrm{b}_j} \\ &= \mathrm{KL}(\mathrm{P}|\mathrm{a}'\otimes\mathrm{b}') - \mathrm{KL}(\mathrm{a}|\mathrm{a}') - \mathrm{KL}(\mathrm{b}|\mathrm{b}').\end{aligned}$$

□

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.



$\varepsilon = 0.010$ $\varepsilon = 0.045$ $\varepsilon = 0.20$

*Figure 7.7:* 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 $\mathrm{f}_i^\varepsilon = \varepsilon\log u_i^\varepsilon$ and $\mathrm{g}_j^\varepsilon = \varepsilon\log v_j^\varepsilon$, with gauge $\langle \mathrm{f}^\varepsilon, \mathrm{a}\rangle = 0$, computed from $\mathrm{P}_{i,j} = u_i\,\mathrm{a}_i\mathrm{b}_j e^{-\mathrm{C}_{i,j}/\varepsilon}\,v_j$. The squared Euclidean cost is normalized by its median positive entry, and all panels use the same axes; increasing $\varepsilon$ turns the 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 7.8** (Convergence with $\varepsilon$)**.** *Assume, after removing zero-mass rows and columns, that* a *and* b *are positive and that* C *is finite. The unique solution* $\mathrm{P}_\varepsilon$ *of* (7.1) *converges to the optimal solution with maximal entropy within the set of all optimal solutions of the Kantorovich problem, namely*

$$\mathrm{P}_\varepsilon \xrightarrow{\varepsilon\to 0} \operatorname*{argmin}_{\mathrm{P}} \ \{-\mathrm{H}(\mathrm{P}) \ ; \ \mathrm{P}\in \mathrm{U}(\mathrm{a},\mathrm{b}), \langle \mathrm{P}, \mathrm{C}\rangle = \mathrm{L}_\mathrm{C}(\mathrm{a},\mathrm{b})\} \tag{7.8}$$

*so that in particular*

$$\mathrm{L}^\varepsilon_\mathrm{C}(\mathrm{a},\mathrm{b}) \xrightarrow{\varepsilon\to 0} \mathrm{L}_\mathrm{C}(\mathrm{a},\mathrm{b}).$$

*Moreover,*

$$\mathrm{P}_\varepsilon \xrightarrow{\varepsilon\to\infty} \mathrm{a}\otimes\mathrm{b}. \tag{7.9}$$

*Proof.* **Case** $\varepsilon\to 0$**.** We consider a sequence $(\varepsilon_\ell)_\ell$ such that $\varepsilon_\ell\to 0$ and $\varepsilon_\ell>0$. We denote $\mathrm{P}_\ell$ the solution of (7.1) for $\varepsilon=\varepsilon_\ell$. Since $\mathrm{U}(\mathrm{a},\mathrm{b})$ is bounded, we can extract a sequence (that we do not relabel for the sake of simplicity) such that $\mathrm{P}_\ell\to\mathrm{P}^\star$. Since $\mathrm{U}(\mathrm{a},\mathrm{b})$ is closed, $\mathrm{P}^\star\in\mathrm{U}(\mathrm{a},\mathrm{b})$. We consider any P such that $\langle \mathrm{C},\mathrm{P}\rangle = \mathrm{L}_\mathrm{C}(\mathrm{a},\mathrm{b})$. Using the equivalent KL-normalized formulation (7.7), optimality of P and $\mathrm{P}_\ell$ for their respective optimization problems (for $\varepsilon=0$ and $\varepsilon=\varepsilon_\ell$) gives

$$0 \leqslant \langle \mathrm{C},\mathrm{P}_\ell\rangle - \langle \mathrm{C},\mathrm{P}\rangle \leqslant \varepsilon_\ell(\mathrm{KL}(\mathrm{P}|\mathrm{a}\otimes\mathrm{b}) - \mathrm{KL}(\mathrm{P}_\ell|\mathrm{a}\otimes\mathrm{b})). \tag{7.10}$$

Since KL is continuous, taking the limit $\ell\to+\infty$ in this expression shows that $\langle \mathrm{C},\mathrm{P}^\star\rangle = \langle \mathrm{C},\mathrm{P}\rangle$ so that $\mathrm{P}^\star$ is a feasible point of (7.8). Furthermore, dividing by $\varepsilon_\ell$ in (7.10) and taking the limit shows that $\mathrm{KL}(\mathrm{P}^\star|\mathrm{a}\otimes\mathrm{b}) \leqslant \mathrm{KL}(\mathrm{P}|\mathrm{a}\otimes\mathrm{b})$, which shows that $\mathrm{P}^\star$ is a solution of (7.8). Since the solution $\mathrm{P}^\star_0$ to this program is unique by strict convexity of $\mathrm{KL}(\cdot|\mathrm{a}\otimes\mathrm{b})$ on the optimal face, one has $\mathrm{P}^\star=\mathrm{P}^\star_0$, and the whole sequence is converging.

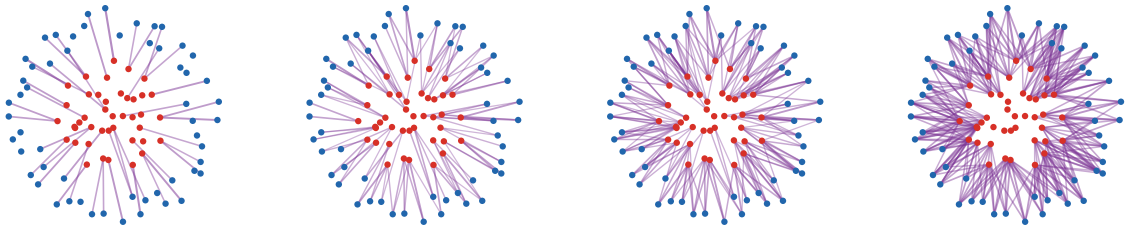
**Case** $\varepsilon\to+\infty$**.** Subtracting $\min_{i,j}\mathrm{C}_{i,j}$ from the cost changes every feasible objective by the same constant, so it does not change the minimizer. We can therefore assume $\mathrm{C}\geqslant 0$. Evaluating the energy at $\mathrm{a}\otimes\mathrm{b}$ (which belongs to the constraint set $\mathrm{U}(\mathrm{a},\mathrm{b})$), one has

$$\langle \mathrm{C},\mathrm{P}_\varepsilon\rangle + \varepsilon\mathrm{KL}(\mathrm{P}_\varepsilon|\mathrm{a}\otimes\mathrm{b}) \leqslant \langle \mathrm{C},\mathrm{a}\otimes\mathrm{b}\rangle + \varepsilon\times 0$$

and since $\langle \mathrm{C},\mathrm{P}_\varepsilon\rangle\geqslant 0$, this leads to

$$\mathrm{KL}(\mathrm{P}_\varepsilon|\mathrm{a}\otimes\mathrm{b}) \leqslant \varepsilon^{-1}\langle \mathrm{C},\mathrm{a}\otimes\mathrm{b}\rangle \leqslant \frac{\|\mathrm{C}\|_\infty}{\varepsilon}$$

so that $\mathrm{KL}(\mathrm{P}_\varepsilon|\mathrm{a}\otimes\mathrm{b})\to 0$ and thus $\mathrm{P}_\varepsilon\to\mathrm{a}\otimes\mathrm{b}$ since KL is a valid divergence. □



$\varepsilon = 0.018$ $\varepsilon = 0.045$ $\varepsilon = 0.120$ $\varepsilon = 0.320$

*Figure 7.8:* 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 sparse to diffuse as $\varepsilon$ increases.

## 7.4 General Formulation

The continuous formulation replaces matrices by measures and discrete KL by relative entropy. It is the static endpoint problem solved by Sinkhorn; the next section explains how it is obtained from an optimization problem on stochastic paths.

One can consider arbitrary measures by replacing the discrete entropy with the relative entropy with respect to the product measure $\mathrm{d}\alpha \otimes \mathrm{d}\beta(x,y) \coloneqq \mathrm{d}\alpha(x)\mathrm{d}\beta(y)$, and propose a regularized counterpart to (3.5) using

$$\mathcal{L}_c^{\varepsilon}(\alpha,\beta) \coloneqq \min_{\pi\in\mathcal{U}(\alpha,\beta)} \int_{\mathcal{X}\times\mathcal{Y}} c(x,y)\mathrm{d}\pi(x,y) + \varepsilon\,\mathrm{KL}(\pi|\alpha\otimes\beta) \tag{7.11}$$

The measure-theoretic definition is the exact analogue of Definition 7.5, with absolute continuity replacing the entrywise support condition.

**Definition 7.9** (Relative entropy of measures)**.** For nonnegative measures $\pi$ and $\xi$ on $\mathcal{X}\times\mathcal{Y}$, the relative entropy is

$$\mathrm{KL}(\pi|\xi) \coloneqq \int_{\mathcal{X}\times\mathcal{Y}} \log\left(\frac{\mathrm{d}\pi}{\mathrm{d}\xi}(x,y)\right)\mathrm{d}\pi(x,y) + \int_{\mathcal{X}\times\mathcal{Y}} (\mathrm{d}\xi(x,y) - \mathrm{d}\pi(x,y)). \tag{7.12}$$

By convention, $\mathrm{KL}(\pi|\xi) = +\infty$ if $\pi$ is not absolutely continuous with respect to $\xi$.

For fixed balanced marginals, the specific product reference in the entropy only matters up to additive constants, exactly as in Proposition 7.7, provided the alternative reference marginals are mutually absolutely continuous with $\alpha$ and $\beta$. 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. The path-space meaning of this static problem is developed next. The main point is that a noisy reference dynamics first defines a probability law on trajectories; after optimizing out the conditional law of the path given its endpoints, only the endpoint coupling remains.

## 7.5 Path-Space Schrödinger Problem

Schrödinger's reciprocal problem is naturally posed on paths rather than on endpoint pairs. The Sinkhorn problem appears after the path law is reduced to its two endpoint marginals.

**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 $\mathcal{Y}$. Let $\Omega = C([0,1];\mathcal{X})$ be a path space and denote by

$$e_t : \Omega \to \mathcal{X}, \qquad e_t(\omega) = \omega_t$$

the evaluation maps. A probability $M \in \mathcal{P}(\Omega)$ is a law of random trajectories. Imposing the endpoint constraints

$$(e_0)_\sharp M = \alpha, \qquad (e_1)_\sharp M = \beta$$

means that the random path starts with law $\alpha$ and ends with law $\beta$. Given a path action $\mathcal{A} : \Omega \to [0,+\infty]$, the unregularized path-space problem is

$$\inf_{M\in\mathcal{P}(\Omega)} \left\{ \int_\Omega \mathcal{A}(\omega)\,\mathrm{d}M(\omega)\ ;\ (e_0)_\sharp M = \alpha,\ (e_1)_\sharp M = \beta \right\}. \tag{7.13}$$

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 Remark 12.4.

The endpoint cost induced by the action is

$$c_{\mathcal{A}}(x,y) \coloneqq \inf_{\omega\in\Omega} \{\mathcal{A}(\omega)\ ;\ e_0(\omega)=x,\ e_1(\omega)=y\}\,. \tag{7.14}$$

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 7.10** (Endpoint reduction of path-space transport)**.** *Assume that minimizing paths in (7.14) can be selected measurably, or more generally that the infimum can be approximated by measurable selections. Then (7.13) 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 $\pi$ 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 $\alpha$ and $\beta$ and action equal, up to the selected approximation error, to $\int c_{\mathcal{A}}\,\mathrm{d}\pi$. Optimizing over $\pi$ 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 $\varepsilon$. Schrödinger's dynamic problem is the entropy projection

$$\mathrm{SB}_\varepsilon(\alpha,\beta) \coloneqq \inf_{M\in\mathcal{P}(\Omega)} \left\{\varepsilon\,\mathrm{KL}(M|\mathcal{R}^\varepsilon)\ ;\ (e_0)_\sharp M=\alpha,\ (e_1)_\sharp M=\beta\right\}. \tag{7.15}$$

This is the dynamic Schrödinger bridge problem. 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 [206] and is surveyed in modern OT language in [142, 143]; stochastic-control formulations are developed in [60].

**Viscous Benamou–Brenier formulations.** The Schrödinger interpolation also admits dynamic optimal-control descriptions, which are viscous analogues of the Benamou–Brenier formula. Write $\sigma$ for the diffusion normalization in this paragraph; it is independent of the entropic temperature $\varepsilon$ used elsewhere. Formally, for smooth positive densities and with the convention

$$\partial_t\rho_t + \mathrm{div}(\rho_t v_t) = \frac{\sigma}{2}\Delta\rho_t,$$

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 $(\rho_t, u_t)$ gives

$$\int_0^1 \int \frac{1}{2}\|v_t\|^2 \rho_t \,\mathrm{d}x\,\mathrm{d}t = \int_0^1 \int \left(\frac{1}{2}\|u_t\|^2 + \frac{\sigma^2}{8}\|\nabla \log \rho_t\|^2\right)\rho_t \,\mathrm{d}x\,\mathrm{d}t + \frac{\sigma}{2}\left[\int \rho_1 \log \rho_1 \,\mathrm{d}x - \int \rho_0 \log \rho_0 \,\mathrm{d}x\right].$$

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 $\sigma$ above by $2\sigma$, so the Fisher coefficient becomes $\sigma^2/2$.

The reduction to endpoint couplings follows from disintegration. We write

$$\mathcal{R}^\varepsilon(\mathrm{d}\omega) = \int \mathcal{R}^{\varepsilon,x,y}(\mathrm{d}\omega)\,\mathcal{R}^\varepsilon_{01}(\mathrm{d}x,\mathrm{d}y), \qquad \mathcal{R}^\varepsilon_{01} := (e_0,e_1)_\sharp \mathcal{R}^\varepsilon,$$

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 7.11** (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*

$$\mathrm{SB}_\varepsilon(\alpha,\beta) = \inf_{\pi\in\mathcal{U}(\alpha,\beta)} \varepsilon\,\mathrm{KL}(\pi|\mathcal{R}^\varepsilon_{01}). \tag{7.16}$$

*For a fixed endpoint coupling $\pi$ with finite $\mathrm{KL}(\pi|\mathcal{R}^\varepsilon_{01})$, the minimizing path law is the mixture of reference bridges*

$$M^\pi = \int \mathcal{R}^{\varepsilon,x,y}\,\mathrm{d}\pi(x,y). \tag{7.17}$$

*Consequently, if $\pi^\star$ solves (7.16), then $M^{\pi^\star}$ solves the path-space problem (7.15).*

*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 $\mathcal{R}^\varepsilon_{01}$. The chain rule for relative entropy gives

$$\mathrm{KL}(M|\mathcal{R}^\varepsilon) = \mathrm{KL}(\pi|\mathcal{R}^\varepsilon_{01}) + \int_{\mathcal{X}\times\mathcal{X}} \mathrm{KL}(M^{x,y}|\mathcal{R}^{\varepsilon,x,y})\,\mathrm{d}\pi(x,y). \tag{7.18}$$

The second term is nonnegative and vanishes exactly when $M^{x,y} = \mathcal{R}^{\varepsilon,x,y}$ for $\pi$-almost every $(x,y)$. Thus, once an endpoint coupling $\pi$ with finite $\mathrm{KL}(\pi|\mathcal{R}^\varepsilon_{01})$ is fixed, the best path law is (7.17), and the remaining minimization is precisely (7.16). 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 [142].

**Brownian bridges and Sinkhorn couplings.** For $\mathcal{X} = \mathbb{R}^d$, take $\mathcal{R}^\varepsilon$ to be a Brownian reference dynamics, up to the conventional scaling of $\varepsilon$. Its endpoint law has a heat-kernel density of the form

$$p_\varepsilon(x, y) \propto \exp\left(-\frac{\|x - y\|^2}{\varepsilon}\right)$$

after absorbing harmless constants into $\varepsilon$. More generally, suppose that the endpoint prior can be written, up to a normalization constant independent of $\pi$, as

$$\mathcal{R}^\varepsilon_{01}(\mathrm{d}x, \mathrm{d}y) \propto \exp\left(-\frac{c(x, y)}{\varepsilon}\right) \alpha(\mathrm{d}x)\beta(\mathrm{d}y).$$

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 $\pi \in \mathcal{U}(\alpha, \beta)$,

$$\varepsilon \,\mathrm{KL}(\pi|\mathcal{R}^\varepsilon_{01}) = \int c(x, y)\,\mathrm{d}\pi(x, y) + \varepsilon \,\mathrm{KL}(\pi|\alpha \otimes \beta) + \text{constant},$$

where the constant does not depend on $\pi$. Hence (7.16) is exactly the continuous Sinkhorn problem (7.11), 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

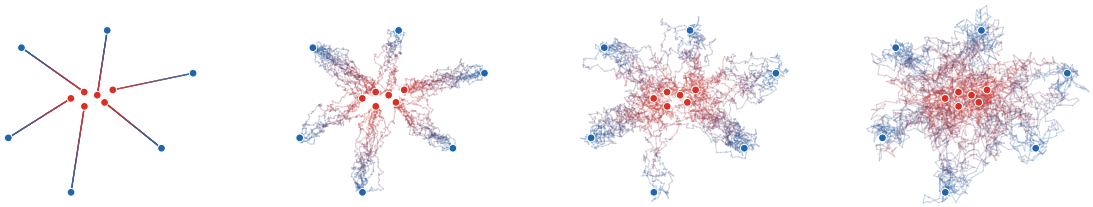
$$M^\star_\varepsilon = \int \mathcal{R}^{\varepsilon,x,y}\,\mathrm{d}\pi^\star_\varepsilon(x, y).$$

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 7.9 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 $\varepsilon$, which simultaneously softens the endpoint coupling and amplifies the Brownian fluctuations between paired endpoints.

**Algorithm 7.2** Endpoint-to-path Schrödinger lift

**Input:** Endpoint laws $\alpha, \beta$, cost $c$, regularization $\varepsilon > 0$, reference bridges $\mathcal{R}^{\varepsilon,x,y}$.
**Output:** Schrödinger path law $M^\star_\varepsilon$.
**Let** $\pi^\star_\varepsilon$ be a minimizer of the static entropic endpoint problem: $\pi^\star_\varepsilon \in \operatorname{argmin}_{\pi\in\mathcal{U}(\alpha,\beta)} \int c\,\mathrm{d}\pi + \varepsilon\,\mathrm{KL}(\pi|\alpha\otimes\beta)$.
**For** each endpoint pair $(x, y)$ sampled from $\pi^\star_\varepsilon$ **do**:
  **Draw** bridge path: $\omega \sim \mathcal{R}^{\varepsilon,x,y}$.
**Return** $M^\star_\varepsilon = \int \mathcal{R}^{\varepsilon,x,y}\,\mathrm{d}\pi^\star_\varepsilon(x, y)$.



$\varepsilon = 0$ $\varepsilon = 0.035$ $\varepsilon = 0.10$ $\varepsilon = 0.28$

*Figure 7.9:* 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 $\varepsilon$, 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 $\pi_{ij}$, 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.

**Definition 7.12** (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 (7.11) is equivalent to

$$\inf_{X\sim\alpha,\, Y\sim\beta} \mathbb{E}\big(c(X, Y)\big) + \varepsilon \mathcal{I}(X, Y).$$

Large $\varepsilon$ therefore favors nearly independent endpoints, while small $\varepsilon$ suppresses endpoint randomness and recovers an optimal Monge–Kantorovich coupling in the limit. When the unregularized quadratic problem has a Brenier map, this limiting coupling is deterministic.

## 7.6 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 (7.7). This formulation has the same minimizer as (7.1); its optimal value is shifted by the constant $\varepsilon \mathrm{H}(\mathrm{a}) + \varepsilon \mathrm{H}(\mathrm{b})$.

**Proposition 7.13** (Dual of entropic OT)**.** *The optimal value of* (7.7) *is*

$$\min_{\mathrm{P}\in\mathrm{U}(\mathrm{a},\mathrm{b})} \langle \mathrm{P}, \mathrm{C}\rangle + \varepsilon \mathrm{KL}(\mathrm{P}|\mathrm{a}\otimes\mathrm{b}) = \max_{\mathrm{f}\in\mathbb{R}^n, \mathrm{g}\in\mathbb{R}^m} \langle \mathrm{f}, \mathrm{a}\rangle + \langle \mathrm{g}, \mathrm{b}\rangle - \varepsilon \sum_{i,j} \exp\left(\frac{\mathrm{f}_i + \mathrm{g}_j - \mathrm{C}_{i,j}}{\varepsilon}\right) \mathrm{a}_i \mathrm{b}_j + \varepsilon. \tag{7.19}$$

*The optimal* $(\mathrm{f}, \mathrm{g})$ *are linked to scalings* $(\mathrm{u}, \mathrm{v})$ *appearing in* (7.2) *through*

$$\mathrm{u}_i = \mathrm{a}_i e^{\mathrm{f}_i/\varepsilon} \quad \textit{and} \quad \mathrm{v}_j = \mathrm{b}_j e^{\mathrm{g}_j/\varepsilon}. \tag{7.20}$$

*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 $(\mathrm{f}, \mathrm{g})$, 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 (6.13) 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 (7.19) 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 7.14** (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

$$\min_{\mathrm{b}}^{\varepsilon}(h) \coloneqq -\varepsilon \log \sum_j e^{-h_j/\varepsilon}\mathrm{b}_j.$$

It converges to $\min_j h_j$ as $\varepsilon \to 0$. Given a cost matrix C, the discrete soft $c$-transforms are

$$\mathrm{f}_i = \min_{\mathrm{b}}^{\varepsilon}(\mathrm{C}_{i,\cdot} - \mathrm{g}) \tag{7.21}$$

and

$$\mathrm{g}_j = \min_{\mathrm{a}}^{\varepsilon}(\mathrm{C}_{\cdot,j} - \mathrm{f}). \tag{7.22}$$

Exponentiating these iterations recovers exactly the Sinkhorn algorithm. These iterations, however, become unstable for small $\varepsilon$. 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 7.3** Log-domain Sinkhorn by soft transforms

**Input:** Weights a, b, cost matrix C, regularization $\varepsilon > 0$, tolerance tol.
**Output:** Entropic coupling P computed from stabilized potentials.
**Initialize:** Set $\mathrm{g}^{(0)} = 0$, $\eta_0 = +\infty$, and $k = 0$.
**While** $\eta_k >$ tol **do**:
  **Set** $k \leftarrow k + 1$.
  **Compute** stabilized soft transform: $\mathrm{f}_i^{(k)} = -\varepsilon \log \sum_j \exp\left(\frac{\mathrm{g}_j^{(k-1)} - \mathrm{C}_{ij}}{\varepsilon}\right)\mathrm{b}_j$.
  **Compute** stabilized reverse soft transform: $\mathrm{g}_j^{(k)} = -\varepsilon \log \sum_i \exp\left(\frac{\mathrm{f}_i^{(k)} - \mathrm{C}_{ij}}{\varepsilon}\right)\mathrm{a}_i$.
  **Set** $\eta_k = \max\{\|\mathrm{f}^{(k)} - \mathrm{f}^{(k-1)}\|_\infty, \|\mathrm{g}^{(k)} - \mathrm{g}^{(k-1)}\|_\infty\}$, with the first term ignored for $k = 1$.
**Return** $\mathrm{P}_{ij} = \mathrm{a}_i \mathrm{b}_j \exp\left(\frac{\mathrm{f}_i^{(k)} + \mathrm{g}_j^{(k)} - \mathrm{C}_{ij}}{\varepsilon}\right)$.

**Continuous dual and soft-transforms.** For general, not necessarily discrete, measures $(\alpha, \beta)$, the KL-regularized problem (7.11) has the concave dual

$$\mathcal{L}_c^{\varepsilon}(\alpha, \beta) = \sup_{f \in C(\mathcal{X}),\, g \in C(\mathcal{Y})} \mathcal{D}_{\varepsilon}(f, g), \tag{7.23}$$

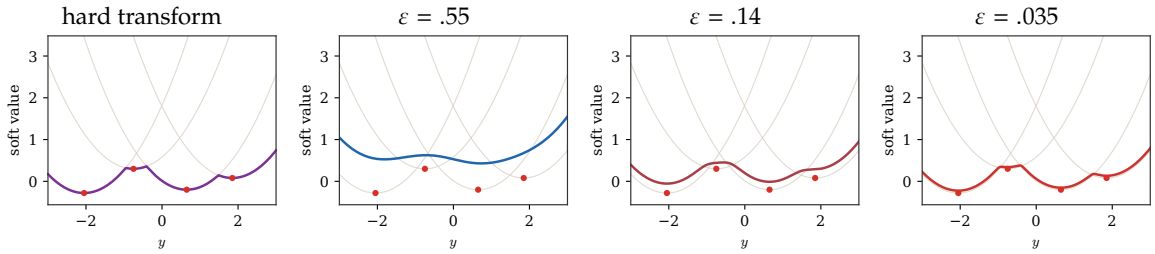



*Figure 7.10:* Soft $c$-transforms for decreasing temperatures. The hard transform is the lower envelope of shifted cost functions, while a positive $\varepsilon$ replaces the pointwise minimum by a log-sum-exp soft minimum. As $\varepsilon$ decreases, the soft transform sharpens and approaches the non-smooth hard envelope.

where

$$\mathcal{D}_\varepsilon(f,g) := \int_{\mathcal{X}} f(x)\mathrm{d}\alpha(x) + \int_{\mathcal{Y}} g(y)\mathrm{d}\beta(y) - \varepsilon \int_{\mathcal{X}\times\mathcal{Y}} \left(e^{\frac{f(x)+g(y)-c(x,y)}{\varepsilon}} - 1\right) \mathrm{d}\alpha(x)\mathrm{d}\beta(y). \tag{7.24}$$

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 7.14.

**Definition 7.15** (Continuous soft $c$-transforms)**.** For $f \in C(\mathcal{X})$ and $g \in C(\mathcal{Y})$, define

$$f^{c,\varepsilon}(y) := -\varepsilon \log\left(\int_{\mathcal{X}} e^{\frac{f(x)-c(x,y)}{\varepsilon}} \mathrm{d}\alpha(x)\right), \qquad y \in \mathcal{Y}, \tag{7.25}$$

$$g^{\bar{c},\varepsilon}(x) := -\varepsilon \log\left(\int_{\mathcal{Y}} e^{\frac{g(y)-c(x,y)}{\varepsilon}} \mathrm{d}\beta(y)\right), \qquad x \in \mathcal{X}. \tag{7.26}$$

In the case of discrete measures, these formulas reduce to (7.21) and (7.22). 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 7.16** (Existence and uniqueness of entropic dual potentials)**.** *Assume that $\mathcal{X}$ and $\mathcal{Y}$ are compact and that $c$ is continuous. The dual problem (7.23) has solutions, and the set of solutions is of the form*

$$(f^\star + \lambda, g^\star - \lambda), \qquad \lambda \in \mathbb{R}.$$

*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 (7.24) 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. Since the potentials are continuous on compact supports, this implies that $f - f'$ is constant and $g - g'$ is the opposite constant. □

**Remark 7.17 (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.

**Remark 7.18 (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 8.5 makes this finite-dimensional closure explicit.

## 7.7 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.

Let $\varphi$ be an entropy function in the sense of Definition 6.13, and recall the $\varphi$-divergence $\mathcal{D}_\varphi$ from Definition 6.14. For $\varepsilon > 0$, define the $\varphi$-regularized transport value

$$\mathcal{L}^{\varepsilon}_{c,\varphi}(\alpha,\beta) \coloneqq \min_{\pi\in\mathcal{U}(\alpha,\beta)} \int_{\mathcal{X}\times\mathcal{Y}} c(x,y)\mathrm{d}\pi(x,y) + \varepsilon\mathcal{D}_\varphi(\pi|\alpha\otimes\beta). \tag{7.27}$$

**Proposition 7.19** (Dual and density law for $\varphi$-regularized OT)**.** *Under the usual Fenchel–Rockafellar qualification assumptions, for instance compact spaces, continuous $c$, and finite value in (7.27), one has*

$$\mathcal{L}^{\varepsilon}_{c,\varphi}(\alpha,\beta) = \sup_{f\in C(\mathcal{X}),\, g\in C(\mathcal{Y})} \int f\mathrm{d}\alpha + \int g\mathrm{d}\beta - \varepsilon\int \varphi^{*,\geqslant 0}\left(\frac{f(x)+g(y)-c(x,y)}{\varepsilon}\right)\mathrm{d}\alpha(x)\mathrm{d}\beta(y). \tag{7.28}$$

*If an optimal plan has density $r^\star = \frac{\mathrm{d}\pi^\star}{\mathrm{d}(\alpha\otimes\beta)}$ and optimal potentials $(f^\star, g^\star)$, then*

$$\frac{f^\star(x)+g^\star(y)-c(x,y)}{\varepsilon} \in \partial\varphi(r^\star(x,y)) \qquad \alpha\otimes\beta\text{-a.e.}$$

*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 $\pi$ 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 (6.13) for the convex functional $\mathcal{D}_\varphi(\cdot|\alpha\otimes\beta)$, the infimum equals

$$-\varepsilon\int \varphi^{*,\geqslant 0}\left(\frac{f(x)+g(y)-c(x,y)}{\varepsilon}\right)\mathrm{d}\alpha(x)\mathrm{d}\beta(y),$$

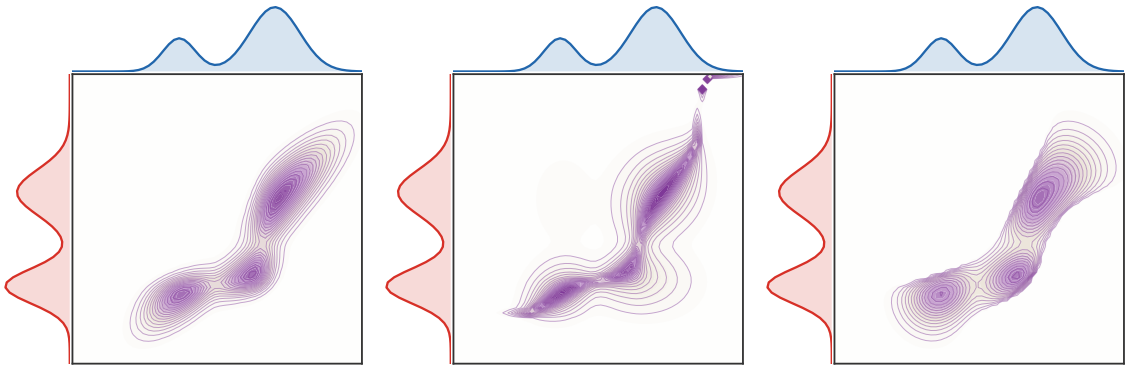
which gives (7.28). Equality in the Fenchel inequality is equivalent to the subgradient inclusion

$$\frac{f^\star\oplus g^\star - c}{\varepsilon} \in \partial\varphi(r^\star),$$

and inversion of $\varphi'$ gives the density law when $\varphi$ 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 $\varphi^{*,\geqslant 0}(s) = e^s - 1$. Taking this parameter to be the Sinkhorn temperature $\varepsilon$ in (7.28) recovers exactly the continuous Sinkhorn dual (7.24). Other choices replace the exponential law by another scalar transfer function:

$$\begin{aligned}
&\varphi(r) = r\log r - r + 1 \quad &&\Rightarrow \quad r^\star = e^s,\\
&\varphi(r) = r - \log r - 1 \quad &&\Rightarrow \quad r^\star = (1-s)^{-1} \quad (s<1),\\
&\varphi(r) = \tfrac{1}{2}(r-1)^2 \quad &&\Rightarrow \quad r^\star = (1+s)_+,
\end{aligned} \qquad s \coloneqq \frac{f^\star\oplus g^\star - c}{\varepsilon}.$$



$\varphi(r) = r\log r - r + 1$ $\varphi(r) = r - \log r - 1$ $\varphi(r) = \frac{1}{2}(r-1)^2$

*Figure 7.11:* 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 the quadratic density penalty can set entries exactly to zero through its positive-part law.

**Bregman vs. $\varphi$-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 7.20** (Measure Bregman divergence)**.** If $\Phi$ is a differentiable convex functional on a convex class of nonnegative measures and $\xi$ is a reference measure in its domain, the measure Bregman divergence generated by $\Phi$ is

$$B_\Phi(\pi|\xi) := \Phi(\pi) - \Phi(\xi) - \int \delta\Phi(\xi)\mathrm{d}(\pi - \xi), \tag{7.29}$$

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 8.1. The KL divergence is special because it is simultaneously a density-ratio divergence and a Bregman divergence.

**Proposition 7.21** (Dual comparison: Bregman vs. density-ratio penalties)**.** *Fix the marginals $\alpha, \beta$ and set $\xi := \alpha \otimes \beta$. Let $\Phi$ be a convex Gateaux-differentiable functional on nonnegative measures on $\mathcal{X} \times \mathcal{Y}$. Its convex conjugate is, for a continuous test function $u$,*

$$\Phi^*(u) := \sup_{\pi \geqslant 0} \left\{ \int u \,\mathrm{d}\pi - \Phi(\pi) \right\}.$$

*Define the Bregman-regularized value, using the same product reference as the density-ratio penalty,*

$$\mathcal{L}^{\varepsilon}_{c,\Phi}(\alpha,\beta) := \inf_{\pi \in \mathcal{U}(\alpha,\beta)} \int c \,\mathrm{d}\pi + \varepsilon B_\Phi(\pi|\xi).$$

*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*

$$\mathcal{L}^{\varepsilon}_{c,\Phi}(\alpha,\beta) = \sup_{f,g} \int f \mathrm{d}\alpha + \int g \mathrm{d}\beta - \varepsilon \left[ \Phi^*\left( \delta\Phi(\xi) + \frac{f \oplus g - c}{\varepsilon} \right) - \Phi^*(\delta\Phi(\xi)) \right]. \tag{7.30}$$

*If $(f^\star, g^\star)$ and $\pi^\star$ are optimal and the solution is interior, then*

$$\delta\Phi(\pi^\star) = \delta\Phi(\xi) + \frac{f^\star \oplus g^\star - c}{\varepsilon}.$$

*By contrast, the density-ratio formulation (7.27) has the scalar-integral dual (7.28) and the pointwise density law*

$$\frac{f^\star \oplus g^\star - c}{\varepsilon} \in \partial\varphi\left( \frac{\mathrm{d}\pi^\star}{\mathrm{d}(\alpha \otimes \beta)} \right).$$

*Proof.* Using (7.29), the Bregman-regularized objective can be written, up to a constant independent of $\pi$, 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 $\pi$ 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 $\xi$ gives

$$-\Phi(\xi) + \int \delta\Phi(\xi)\mathrm{d}\xi = \Phi^*(\delta\Phi(\xi)),$$

which yields (7.30). Equality in Fenchel's inequality gives the optimality condition for $\pi^\star$. The density-ratio dual and density law are exactly those of Proposition 7.19. Placing the two formulas side by side shows the structural difference: Bregman regularization translates the reference measure in the dual coordinate $\delta\Phi$, whereas $\varphi$-regularization applies a scalar nonlinearity to the density with respect to the moving product reference $\alpha\otimes\beta$. □

When $\Phi$ is separable with respect to a fixed dominating measure $\xi_0$, 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 $\varphi$ 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 7.22** (KL is the common Bregman and $\varphi$ case)**.** *Let $\omega$ be a finite reference measure with a nontrivial measurable subset. Work on probability measures $\alpha = p\omega$ and $\beta = q\omega$ whose densities are bounded above and below away from 0. Assume that $\Phi$ is twice Gateaux 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 $\alpha,\beta$, then there exist $c\geqslant 0$ and $a\in\mathbb{R}$ such that*

$$\varphi(t) = c\,t\log t + a(t-1).$$

*Hence the common divergence is $c\,\mathrm{KL}(\alpha|\beta)$, and $\Phi(p\omega)$ differs from $c\int p\log p\,\mathrm{d}\omega$ 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 $c = \varphi''(1)\geqslant 0$, the functional $\Psi(p\omega) = \Phi(p\omega) - c\int p\log p\mathrm{d}\omega$ has zero second variation along every zero-mass line segment in the positive simplex. It is therefore affine there, and $B_\Psi = 0$. Thus $B_\Phi = c\,\mathrm{KL}$.

It remains to identify the generator. Since $\mathcal{D}_\varphi = c\,\mathrm{KL}$, the function $g(t) = \varphi(t) - c\,t\log t$ generates the zero $\varphi$-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 $(y-1)g(x) + (1-x)g(y) = 0$. Hence $g(t)/(t-1)$ is constant on $(0,+\infty)\setminus\{1\}$, so $g(t) = a(t-1)$. 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 8.1: 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 (7.28) and, for interior solutions, the pointwise law $r^\star = (\varphi')^{-1}((f\oplus g - c)/\varepsilon)$. Proposition 7.21 makes the distinction explicit at the dual level. The Bregman dual contains the global conjugate $\Phi^*$ and the product reference $\alpha\otimes\beta$, whereas the $\varphi$-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.

## 7.8 Sinkhorn Divergences

Sinkhorn divergences remove the entropic self-bias while retaining smoothness. They interpolate between OT-like geometry and kernel-like norms, which explains their statistical behavior.

**Entropic bias.** A major issue with the value of the Sinkhorn problem (7.11) is that $\mathcal{L}_c^\varepsilon(\alpha,\beta) > 0$. In particular,

$$\alpha_\varepsilon = \operatorname*{argmin}_\beta \mathcal{L}_c^\varepsilon(\alpha,\beta)$$

does not satisfy $\alpha_\varepsilon = \alpha$ unless $\varepsilon = 0$. The following proposition shows that the bias induced by this entropic regularization dominates the large $\varepsilon$ limit.

**Proposition 7.23** (Large-temperature entropic bias)**.** *Assume that $c$ is bounded and 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 $\varepsilon$ 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. □

So in the large $\varepsilon$ limit, $\mathcal{L}_c^\varepsilon$ behaves like an inner product and not like a norm. The following special case makes the resulting attraction explicit.

**Example 7.24 (Large-temperature collapse for quadratic costs).** The limiting functional minimized by $\alpha_\varepsilon$ is linear in the second argument:

$$\beta \mapsto \int V_\alpha(y)\mathrm{d}\beta(y), \qquad V_\alpha(y) \coloneqq \int c(x,y)\mathrm{d}\alpha(x).$$

Thus any limiting minimizer is supported on $\operatorname{argmin} V_\alpha$. When this minimizer is unique,

$$\alpha_\varepsilon \rightharpoonup \delta_{y^\star(\alpha)}, \qquad y^\star(\alpha) = \operatorname*{argmin}_y V_\alpha(y).$$

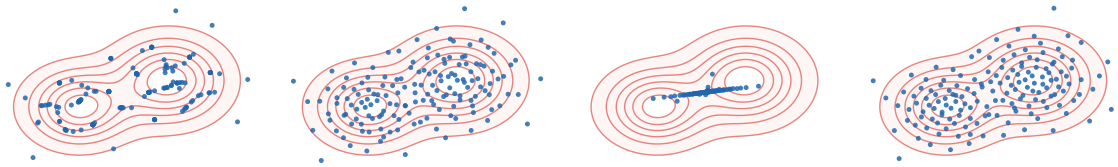
For the quadratic cost $c(x,y) = \|x-y\|^2$ on $\mathbb{R}^d$, assuming $\alpha$ has finite second moment, one has $V_\alpha(y) = \|y - \int x\mathrm{d}\alpha(x)\|^2 + \text{const}$, so the collapse is toward the Dirac mass at the mean of $\alpha$.

**Sinkhorn divergences.** The raw entropic OT value has a large-temperature attraction toward the product coupling. The standard debiasing subtracts the two self-interaction energies, in the same spirit as passing from a positive kernel value to the associated squared distance.

**Definition 7.25** (Sinkhorn divergence)**.** For $\varepsilon > 0$, the debiased Sinkhorn divergence associated with the entropic OT value $\mathcal{L}_c^\varepsilon$ is

$$\bar{\mathcal{L}}_c^\varepsilon(\alpha,\beta) := \mathcal{L}_c^\varepsilon(\alpha,\beta) - \frac{1}{2}\mathcal{L}_c^\varepsilon(\alpha,\alpha) - \frac{1}{2}\mathcal{L}_c^\varepsilon(\beta,\beta). \tag{7.31}$$

Although this formula is a debiasing by self-costs, its non-negativity is not automatic from the definition; Proposition 7.29 proves it below by a kernel Cauchy–Schwarz argument.



$\mathcal{L}_c^\varepsilon$, small $\varepsilon$ $\bar{\mathcal{L}}_c^\varepsilon$, small $\varepsilon$ $\mathcal{L}_c^\varepsilon$, large $\varepsilon$ $\bar{\mathcal{L}}_c^\varepsilon$, large $\varepsilon$

*Figure 7.12:* Visualization of the debiasing effect by point optimization. The red level sets show a fixed two-Gaussian target density $\beta$, while the blue atoms are an optimized empirical measure $\alpha_n$ initialized in the same way in all panels. For small $\varepsilon$, 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 $\varepsilon$, minimizing $\mathcal{L}_c^\varepsilon$ collapses the atoms toward the barycenter predicted by the large-temperature bias, whereas the self-cost subtraction in (7.31) keeps a bimodal cloud.

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.

**Lemma 7.26** (Entropic dual cost at optimum)**.** *Let $(f_{\alpha,\beta}, g_{\alpha,\beta})$ be optimal dual potentials, normalized arbitrarily. Then*

$$\mathcal{L}_c^\varepsilon(\alpha,\beta) = \langle f_{\alpha,\beta},\, \alpha\rangle + \langle g_{\alpha,\beta},\, \beta\rangle. \tag{7.32}$$

*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 (7.23) gives the result. □

The next proposition records the two limiting regimes of this debiased quantity.

**Proposition 7.27** (Asymptotics of Sinkhorn divergences)**.** *Assume that the two measures are supported on the same space and that $c$ is bounded, continuous, nonnegative and satisfies $c(x,x) = 0$. 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 already gives the correct intuition; we now use the standard continuous argument. **Case $\varepsilon \to 0$.** The first limit follows from the standard Γ-convergence argument for entropic optimal transport: the entropy term is lower semicontinuous along weakly converging couplings, while any finite-cost coupling can be approximated by couplings with finite entropy. Since $c(x,x) = 0$, the two self-costs in the debiased expression converge to zero, and 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 $f_\varepsilon$ (equivalently the Sinkhorn fixed point on $f_\varepsilon$) reads

$$f_\varepsilon = -\varepsilon \log \int \exp\left(\frac{g_\varepsilon(y) - c(\cdot, y)}{\varepsilon}\right) \mathrm{d}\beta(y) = -\varepsilon \log \int \left(1 + \frac{g_\varepsilon(y) - c(\cdot, y)}{\varepsilon} + o(1/\varepsilon)\right) \mathrm{d}\beta(y)$$
$$= -\varepsilon \log\left(1 + \frac{1}{\varepsilon}\int (g_\varepsilon(y) - c(\cdot, y))\mathrm{d}\beta(y) + o(1/\varepsilon)\right) = -\int g_\varepsilon \mathrm{d}\beta + \int c(\cdot, y)\mathrm{d}\beta(y) + o(1).$$

Plugging this relation in the dual expression (7.32)

$$\mathcal{L}_c^\varepsilon(\alpha, \beta) = \int f_\varepsilon \mathrm{d}\alpha + \int g_\varepsilon \mathrm{d}\beta = \iint c(x, y)\mathrm{d}\alpha(x)\mathrm{d}\beta(y) + o(1).$$

Applying this expansion to $(\alpha, \beta)$, $(\alpha, \alpha)$ and $(\beta, \beta)$ 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 7.28 (Large-temperature Hilbertian limit).** If $-c$ defines a conditionally positive definite kernel, the large-temperature limit in Proposition 7.27 is the square of a Hilbertian kernel norm. A typical example is $c(x, y) = \|x - y\|^p$ for $0 < p < 2$, which corresponds to the energy-distance kernel. This kernel norm is the dual of a homogeneous Sobolev norm.

We now show that this debiased Sinkhorn divergence is positive.

**Proposition 7.29** (Non-negativity of Sinkhorn divergences)**.** *If $k(x, y) = e^{-c(x,y)/\varepsilon}$ is positive definite, then $\bar{\mathcal{L}}_c^\varepsilon(\alpha, \beta) \geqslant 0$.*

*Proof.* In the following, we denote by $(f_{\alpha,\beta}, g_{\alpha,\beta})$ optimal dual potentials for the dual Schrödinger problem between $\alpha$ and $\beta$. We denote by $f_{\alpha,\alpha} = g_{\alpha,\alpha}$ (one can assume they are equal by symmetry) the solution for the problem between $\alpha$ and itself. Using the suboptimal function $(f_{\alpha,\alpha}, g_{\beta,\beta})$ in the dual maximization problem, and using relation (7.32) 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 $\beta$, so the previous inequality equivalently reads

$$\frac{1}{\varepsilon}\bar{\mathcal{L}}_c^\varepsilon(\alpha, \beta) \geqslant 1 - \langle e^{\frac{f_{\alpha,\alpha} \oplus g_{\beta,\beta} - c}{\varepsilon}}, \, \alpha \otimes \beta\rangle = 1 - \langle \tilde{\alpha}, \, \tilde{\beta}\rangle_k$$

where $\tilde{\alpha} = e^{f_{\alpha,\alpha}/\varepsilon}\alpha$, $\tilde{\beta} = e^{f_{\beta,\beta}/\varepsilon}\beta$ and we introduced the inner product, valid because $k$ is positive definite, $\langle \tilde{\alpha}, \, \tilde{\beta}\rangle_k := \int k(x, y)\mathrm{d}\tilde{\alpha}(x)\mathrm{d}\tilde{\beta}(y)$. The self Sinkhorn fixed point equation, once exponentiated, reads pointwise

$$e^{f_{\alpha,\alpha}(x)/\varepsilon}\int k(x, y)\mathrm{d}\tilde{\alpha}(y) = 1 \qquad \text{for } \alpha\text{-a.e. } x,$$

and hence

$$\|\tilde{\alpha}\|_k^2 = \langle k(\tilde{\alpha}), \, \tilde{\alpha}\rangle = \int e^{f_{\alpha,\alpha}(x)/\varepsilon} k(\tilde{\alpha})(x)\mathrm{d}\alpha(x) = 1$$

and similarly $\|\tilde{\beta}\|_k^2 = 1$. Therefore, by Cauchy–Schwarz, one has $1 - \langle \tilde{\alpha}, \, \tilde{\beta}\rangle_k \geqslant 0$. □

> **Remark 7.30 (Strict positivity).** Under additional assumptions on the kernel, one can furthermore show that $\bar{\mathcal{L}}_c^\varepsilon(\alpha, \beta) = 0$ implies $\alpha = \beta$, and that this debiased divergence metrizes convergence in law.

CHAPTER 8

# Entropic Regularization: Convergence

Convergence for entropic optimal transport has two complementary meanings. At fixed marginals and fixed temperature, one studies how Sinkhorn iterates approach the regularized optimizer; when the marginals are empirical, one also studies how the regularized value and potentials behave as the number of samples grows toward a mean-field limit. This chapter keeps these two scales together: first the algorithmic convergence of matrix scaling and soft transforms, then the Gaussian closed forms and sample-complexity consequences that explain the statistical role of the regularization.

The algorithmic part of the chapter revisits Sinkhorn convergence through three complementary lenses. Bregman projections explain the alternating-projection geometry, Fortet's order argument gives qualitative fixed-point convergence, robust Bregman estimates give a non-asymptotic $O(1/k)$ dual-gap bound, and Hilbert's metric gives a clean linear contraction when the kernel is uniformly positive. The first and last viewpoints explain convergence mechanisms, while the robust estimate is often the most useful for explicit complexity guarantees.

## 8.1 Sinkhorn Convergence: Bregman 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.

**Definition 8.1** (Bregman divergence)**.** Let $\Phi$ be a differentiable strictly convex function on a convex domain $\Omega$. The Bregman divergence generated by $\Phi$ is

$$B_\Phi(P|Q) := \Phi(P) - \Phi(Q) - \langle \nabla\Phi(Q),\, P - Q\rangle.$$

For the negative entropy $\Phi(P) = \sum_{i,j} P_{i,j} \log P_{i,j}$ on the positive orthant, one obtains $B_\Phi(P|Q) = \mathrm{KL}(P|Q)$ up to the harmless convention at the boundary.

Bregman divergences are useful because their geometry can encode constraints. A Legendre-type generator $\Phi$ 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 8.2** (Linear tilts of Bregman penalties)**.** *Let $\Phi$ be differentiable and strictly convex, and let $B_\Phi$ be its Bregman divergence. Fix a reference point $\mathrm{Q}$ in the interior of the domain. Assume that there exists $\mathrm{Q}^{\mathrm{C}}$ such that*

$$\nabla\Phi(\mathrm{Q}^{\mathrm{C}}) = \nabla\Phi(\mathrm{Q}) - \mathrm{C}/\varepsilon.$$

*Then, for all $\mathrm{P}$ in the domain,*

$$\langle \mathrm{P},\, \mathrm{C}\rangle + \varepsilon B_\Phi(\mathrm{P}|\mathrm{Q}) = \varepsilon B_\Phi(\mathrm{P}|\mathrm{Q}^{\mathrm{C}}) + \mathrm{cst},$$

*where the constant does not depend on $\mathrm{P}$.*

*Proof.* Subtract the two Bregman divergences:

$$B_\Phi(\mathrm{P}|\mathrm{Q}^{\mathrm{C}}) - B_\Phi(\mathrm{P}|\mathrm{Q}) = \langle \nabla\Phi(\mathrm{Q}) - \nabla\Phi(\mathrm{Q}^{\mathrm{C}}),\, \mathrm{P}\rangle + \mathrm{cst}.$$

Using $\nabla\Phi(\mathrm{Q}) - \nabla\Phi(\mathrm{Q}^{\mathrm{C}}) = \mathrm{C}/\varepsilon$ and multiplying by $\varepsilon$ gives the claim. □

For the negative entropy $\Phi(\mathrm{P}) = \sum_{i,j} \mathrm{P}_{i,j} \log \mathrm{P}_{i,j}$, one has $B_\Phi = \mathrm{KL}$. Taking $\mathrm{Q} = \mathrm{a} \otimes \mathrm{b}$ gives the tilted reference

$$\mathrm{K}^{\varepsilon}_{\mathrm{a,b}} := (\mathrm{a} \otimes \mathrm{b}) \odot e^{-\mathrm{C}/\varepsilon}.$$

Thus

$$\langle \mathrm{P},\, \mathrm{C}\rangle + \varepsilon \mathrm{KL}(\mathrm{P}|\mathrm{a} \otimes \mathrm{b}) = \varepsilon \mathrm{KL}(\mathrm{P}|\mathrm{K}^{\varepsilon}_{\mathrm{a,b}}) + \mathrm{cst}.$$

On the transport polytope, scaling $\mathrm{K}^{\varepsilon}_{\mathrm{a,b}}$ is equivalent to scaling the Gibbs kernel $\mathrm{K} = e^{-\mathrm{C}/\varepsilon}$ because the factors $\mathrm{a}_i$ and $\mathrm{b}_j$ can be absorbed into the Sinkhorn scalings.

Thus the unique solution $\mathrm{P}_\varepsilon$ of (7.1) is the KL projection of the tilted Gibbs reference onto $\mathrm{U}(\mathrm{a}, \mathrm{b})$:

$$\mathrm{P}_\varepsilon = \mathrm{Proj}^{\mathrm{KL}}_{\mathrm{U(a,b)}}(\mathrm{K}^{\varepsilon}_{\mathrm{a,b}}) := \operatorname*{argmin}_{\mathrm{P} \in \mathrm{U(a,b)}} \mathrm{KL}(\mathrm{P}|\mathrm{K}^{\varepsilon}_{\mathrm{a,b}}). \tag{8.1}$$

**Cyclic projection convergence.** The convergence mechanism is the classical one of Bregman [42].

**Proposition 8.3** (Cyclic Bregman projections on affine constraints)**.** *Let $\Phi$ be a Legendre strictly convex generator on a finite-dimensional convex domain, and let $C_1, C_2$ be affine constraint sets whose intersection meets the domain. Define*

$$P_{k+1} = \mathrm{Proj}^{B_\Phi}_{C_2} \mathrm{Proj}^{B_\Phi}_{C_1}(P_k),$$

*starting from an interior point $P_0$. Assume the projections are well-defined and that the iterates remain in a compact subset of the domain. Then $P_k$ converges to the Bregman projection of $P_0$ onto $C_1 \cap C_2$. In particular, the KL case converges for positive affine marginal constraints on a bounded transportation polytope.*

*Proof.* We first prove the Pythagorean identity used by the projection argument. For three interior points one has the Bregman three-point formula

$$B_\Phi(Q|P) = B_\Phi(Q|P^+) + B_\Phi(P^+|P) + \langle \nabla\Phi(P^+) - \nabla\Phi(P),\, Q - P^+\rangle.$$

If $P^+ = \mathrm{Proj}^{B_\Phi}_C(P)$ and $C$ is affine, the first-order optimality condition for minimizing $R \mapsto B_\Phi(R|P)$ over $R \in C$ is

$$\langle \nabla\Phi(P^+) - \nabla\Phi(P),\, R - P^+\rangle = 0 \qquad \forall R \in C,$$

because $R - P^+$ ranges over the tangent linear space of $C$. Taking $R = Q \in C$ cancels the last term and gives

$$B_\Phi(Q|P) = B_\Phi(Q|P^+) + B_\Phi(P^+|P) \qquad \forall Q \in C.$$

Let $(Z_\ell)_\ell$ be the half-step sequence obtained by alternating projections onto $C_1$ and $C_2$, so that $Z_{2k} = P_k$ and $Z_{2k+2} = P_{k+1}$. Fix $Q \in C_1 \cap C_2$. Applying the identity at each half-step gives

$$B_\Phi(Q|Z_\ell) - B_\Phi(Q|Z_{\ell+1}) = B_\Phi(Z_{\ell+1}|Z_\ell) \geqslant 0.$$

Thus $B_\Phi(Q|Z_\ell)$ decreases and the series $\sum_\ell B_\Phi(Z_{\ell+1}|Z_\ell)$ is finite. The compactness assumption gives cluster points. Since the projection drops tend to zero and $\Phi$ is strictly convex on compact subsets of the domain, $\|Z_{\ell+1} - Z_\ell\| \to 0$. Every cluster point of the even subsequence is therefore also a cluster point of the adjacent odd subsequence. Because these two subsequences lie alternately in the closed affine sets $C_1$ and $C_2$, every cluster point belongs to $C_1 \cap C_2$.

Let $\bar{P}$ be such a cluster point. For each half-step, the dual displacement

$$\nabla\Phi(Z_{\ell+1}) - \nabla\Phi(Z_\ell)$$

belongs to the normal space of the affine set onto which one projects. Telescoping and using the convergence of $Z_\ell$ gives

$$\nabla\Phi(\bar{P}) - \nabla\Phi(P_0) \in N_{C_1} + N_{C_2} = N_{C_1 \cap C_2},$$

where the last equality uses that the sets are affine. This is precisely the first-order optimality condition for minimizing $R \mapsto B_\Phi(R|P_0)$ over $R \in C_1 \cap C_2$. Thus $\bar{P}$ is the Bregman projection of $P_0$ onto the intersection. Strict convexity gives uniqueness of this minimizer, so all cluster points coincide and the whole sequence converges. The KL statement follows by choosing the negative entropy generator. □

**Algorithm 8.1** Cyclic Bregman projections

**Input:** Constraint sets $C_1, C_2$, Bregman divergence $B_\Phi$, interior point $P_0$, constraint defects $\mathrm{def}_{C_1}, \mathrm{def}_{C_2}$, tolerance tol.
**Output:** Point in $C_1 \cap C_2$ when the intersection is feasible.
**Specialize** to entropic OT, if needed: $C_1 = C_{\mathrm{a}}^1$, $C_2 = C_{\mathrm{b}}^2$, $B_\Phi = \mathrm{KL}$.
**Initialize:** Set $r_0 = +\infty$ and $k = 0$.
**While** $r_k >$ tol **do**:
  **Set** $k \leftarrow k+1$.
  $P_{k-1/2} = \mathrm{Proj}_{C_1}^{B_\Phi}(P_{k-1})$.
  $P_k = \mathrm{Proj}_{C_2}^{B_\Phi}(P_{k-1/2})$.
  **Set** $r_k = \max\{\mathrm{def}_{C_1}(P_k), \mathrm{def}_{C_2}(P_k)\}$.
**Return** $P_k$.

Denoting

$$C_{\mathrm{a}}^1 \coloneqq \{\mathrm{P}\,;\, \mathrm{P}\mathbb{1}_m = \mathrm{a}\} \quad \text{and} \quad C_{\mathrm{b}}^2 \coloneqq \left\{\mathrm{P}\,;\, \mathrm{P}^\top \mathbb{1}_n = \mathrm{b}\right\}$$

the rows and columns constraints, one has $\mathrm{U}(\mathrm{a},\mathrm{b}) = C_{\mathrm{a}}^1 \cap C_{\mathrm{b}}^2$. One can use KL, or more generally Bregman, iterative projections [42, 198, 199]

$$\mathrm{P}^{(\ell+1)} \coloneqq \mathrm{Proj}_{C_{\mathrm{a}}^1}^{\mathrm{KL}}(\mathrm{P}^{(\ell)}) \quad \text{and} \quad \mathrm{P}^{(\ell+2)} \coloneqq \mathrm{Proj}_{C_{\mathrm{b}}^2}^{\mathrm{KL}}(\mathrm{P}^{(\ell+1)}). \tag{8.2}$$

Since the sets $C_{\mathrm{a}}^1$ and $C_{\mathrm{b}}^2$ are affine, Proposition 8.3 applies with $P_0 = \mathrm{K}_{\mathrm{a},\mathrm{b}}^\varepsilon$ and shows convergence to the solution of (8.1).

**Row and column scalings.** The two projectors are simple to compute since they correspond to scaling respectively the rows and the columns, as explained in this proposition.

**Proposition 8.4** (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 $p = sq/\sum_i q_i$. □

These iterations are equivalent to Sinkhorn iterations (7.5) since defining

$$\mathrm{P}^{(2\ell)} \coloneqq \mathrm{diag}(\mathrm{u}^{(\ell)})\mathrm{K}\,\mathrm{diag}(\mathrm{v}^{(\ell)}),$$

one has

$$\begin{aligned} \mathrm{P}^{(2\ell+1)} &\coloneqq \mathrm{diag}(\mathrm{u}^{(\ell+1)})\mathrm{K}\,\mathrm{diag}(\mathrm{v}^{(\ell)}) \\ \text{and} \quad \mathrm{P}^{(2\ell+2)} &\coloneqq \mathrm{diag}(\mathrm{u}^{(\ell+1)})\mathrm{K}\,\mathrm{diag}(\mathrm{v}^{(\ell+1)}) \end{aligned}$$

In practice, however, one should prefer using (7.5), 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 $\varepsilon$. 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 $O(1/k)$ estimate explains what can be guaranteed before any asymptotic linear regime becomes visible. It is also possible to anneal $\varepsilon$ during the iterations and to rely on multiscale strategies in low dimensions.

## 8.2 Sinkhorn Convergence: Monotone Point of View

There is another, older way to understand convergence, going back to Fortet's proof of the Schrödinger system [91, 89, 144]. It does not primarily estimate a contraction factor. Instead, it uses the order structure of the soft transforms.

**Proposition 8.5** (Monotone fixed-point route to Sinkhorn convergence)**.** *Let $c$ be bounded and continuous on compact spaces, and define the double Sinkhorn map, normalized by subtracting its $\alpha$-mean,*

$$\mathcal{A}(f) := (f^{c,\varepsilon})^{\bar{c},\varepsilon} - \int (f^{c,\varepsilon})^{\bar{c},\varepsilon} \mathrm{d}\alpha.$$

*The map $\mathcal{A}$ is order preserving on the quotient by additive constants. If a representative of the initial class is chosen so that $f_0 \leqslant \mathcal{A}(f_0)$, then representatives of the iterates $f_{k+1} = \mathcal{A}(f_k)$ can be chosen to increase pointwise, remain uniformly bounded in oscillation, and converge to a fixed point. Since constants are free, any bounded initialization can be shifted downward to satisfy the subsolution inequality. The fixed point is the entropic potential, hence the associated Sinkhorn scalings converge.*

*Proof.* The soft $c$-transform is order reversing: if $g \leqslant g'$, then

$$-\varepsilon \log \int e^{(g-c)/\varepsilon} \mathrm{d}\beta \geqslant -\varepsilon \log \int e^{(g'-c)/\varepsilon} \mathrm{d}\beta.$$

The composition of two order-reversing transforms is therefore order preserving. The transform also commutes with additive constants in the projective sense, which is why the argument is naturally stated for equivalence classes modulo constants. Starting from a subsolution representative gives $f_0 \leqslant f_1$, and order preservation gives representatives satisfying $f_k \leqslant f_{k+1}$ for all $k$. Soft-transform oscillation bounds, controlled by $\sup c - \inf c$, prevent escape to infinity after normalization. Monotone pointwise convergence, compactness of equicontinuous soft transforms, and continuity of $\mathcal{A}$ then give a fixed point. Uniqueness of entropic potentials up to constants, Proposition 7.2 and the dual uniqueness statement above identify this fixed point with the Sinkhorn solution. □

This proof is qualitative rather than quantitative, but it is conceptually useful: Sinkhorn is not only alternating projection or projective contraction; it is also a monotone fixed-point iteration on potential classes once constants are quotiented out.

**Definition 8.6** (Variation seminorm)**.** For a bounded real-valued function $h$, the variation seminorm is

$$\|h\|_V := \sup h - \inf h.$$

It vanishes exactly on constant functions, hence becomes a norm after quotienting by additive constants.

This is the natural size for Sinkhorn potentials because adding constants changes their gauge but not the coupling.

**Proposition 8.7** (Topical maps are variation-nonexpansive)**.** *Let $E$ be a vector space of real-valued bounded functions, ordered pointwise, and write $\|\cdot\|_V$ for the variation seminorm of Definition 8.6. Let $\mathcal{T} : E \to E$ be monotone and additively homogeneous,*

$$f \leqslant g \Rightarrow \mathcal{T}(f) \leqslant \mathcal{T}(g), \qquad \mathcal{T}(f + \lambda) = \mathcal{T}(f) + \lambda \quad \forall \lambda \in \mathbb{R}.$$

*Then*

$$\|\mathcal{T}(f) - \mathcal{T}(g)\|_V \leqslant \|f - g\|_V.$$

*The same conclusion holds for order-reversing maps satisfying $\mathcal{T}(f + \lambda) = \mathcal{T}(f) - \lambda$.*

*Proof.* Set $a = \inf(f - g)$ and $b = \sup(f - g)$. Then

$$g + a \leqslant f \leqslant g + b.$$

If $\mathcal{T}$ is order preserving and additively homogeneous, applying $\mathcal{T}$ gives

$$\mathcal{T}(g) + a \leqslant \mathcal{T}(f) \leqslant \mathcal{T}(g) + b.$$

Hence every value of $\mathcal{T}(f) - \mathcal{T}(g)$ lies in $[a, b]$, so its oscillation is at most $b - a = \|f - g\|_V$. For an order-reversing, additively anti-homogeneous map, the same inequalities give

$$\mathcal{T}(g) - b \leqslant \mathcal{T}(f) \leqslant \mathcal{T}(g) - a,$$

and the oscillation bound is identical. □

**Corollary 8.8** (Soft transforms are nonexpansive)**.** *For every $\varepsilon > 0$, the soft c-transforms (7.25)–(7.26) are 1-Lipschitz for the variation seminorm. Consequently, the double Sinkhorn map used in Proposition 8.5, after quotienting constants, is also 1-Lipschitz for $\|\cdot\|_V$.*

*Proof.* The soft transform is order reversing and satisfies $(g + \lambda)^{\bar{c},\varepsilon} = g^{\bar{c},\varepsilon} - \lambda$, and similarly for the other block. Proposition 8.7 applies to each block, and the composition of two 1-Lipschitz maps is 1-Lipschitz. □

> **Remark 8.9 (Topical maps and projective geometry).** Order-preserving additively homogeneous maps are called topical maps in nonlinear Perron–Frobenius theory [141]. Proposition 8.7 is the basic nonexpansiveness mechanism behind Fortet's monotone argument. The Hilbert-metric analysis in Section 8.4 is stronger: under strict positivity assumptions on the kernel it upgrades nonexpansiveness to a genuine projective contraction.

## 8.3 Sinkhorn Convergence: Sublinear Robust Rate

Sinkhorn is cyclic coordinate ascent on the smooth dual objective $\mathcal{D}_\varepsilon$ defined in (7.24); equivalently, it alternates KL projections on the two marginal constraint sets. The following statement records the rate most useful for complexity estimates: before any possible linear regime becomes visible, the dual objective gap decreases at order $1/k$. We state it for balanced entropic OT, which is the specialization needed here. Related robust Bregman-projection rates are developed in [183, 4, 84], and statistical consequences of entropic smoothing are analyzed in [103, 30].

**Proposition 8.10** (Pinsker inequality)**.** *If $p, q \in \Sigma_n$, then*

$$\|p - q\|_1^2 \leqslant 2\,\mathrm{KL}(p|q).$$

*Proof.* Let $A = \{i : p_i \geqslant q_i\}$ and set $a = \sum_{i\in A} p_i$, $b = \sum_{i\in A} q_i$. Then $a - b = \frac{1}{2}\|p - q\|_1$. Applying data processing for relative entropy to the partition $(A, A^c)$ gives

$$\mathrm{KL}(p|q) \geqslant a \log \frac{a}{b} + (1 - a) \log \frac{1 - a}{1 - b}.$$

For fixed $b \in (0, 1)$, the function

$$h(a) = a \log \frac{a}{b} + (1 - a) \log \frac{1 - a}{1 - b} - 2(a - b)^2$$

satisfies $h(b) = h'(b) = 0$ and

$$h''(a) = \frac{1}{a} + \frac{1}{1 - a} - 4 \geqslant 0,$$

because $a(1 - a) \leqslant 1/4$. Hence the binary relative entropy is at least $2(a - b)^2 = \frac{1}{2}\|p - q\|_1^2$. The boundary cases follow by approximation. □

**Proposition 8.11** (A compact $O(1/k)$ dual rate)**.** *Assume that $\mathcal{X}$ and $\mathcal{Y}$ are compact, that $c$ is bounded, and write $R = \sup c - \inf c$. Let $(f_k, g_k)$ be Sinkhorn dual iterates normalized by $\int f_k \mathrm{d}\alpha = 0$, and let*

$$\Delta_k := \mathcal{D}_\varepsilon(f^\star, g^\star) - \mathcal{D}_\varepsilon(f_k, g_k)$$

*be the dual suboptimality gap for the entropic dual objective (7.24). Then there exists a numerical constant $C$ such that*

$$\Delta_k \leqslant \frac{CR^2}{\varepsilon(k + 1)}.$$

*Proof.* The proof uses three elementary ingredients. First, the soft $c$-transform bounds the oscillation of every normalized iterate and every normalized optimum:

$$\|f_k\|_V + \|g_k\|_V + \|f^\star\|_V + \|g^\star\|_V \leqslant CR.$$

Here $\|h\|_V := \sup h - \inf h$ is the variation seminorm, the same projective norm used in Hilbert's metric in Section 8.4. It is natural because dual potentials can be shifted by constants without changing the coupling.

Second, each Sinkhorn half-step is an exact KL projection. The Pythagorean identity for KL projections gives the ascent identity

$$\mathcal{D}_\varepsilon(f_{k+1}, g_{k+1}) - \mathcal{D}_\varepsilon(f_k, g_k) = \varepsilon\big[\,\mathrm{KL}(\pi^\star|\pi_k) - \mathrm{KL}(\pi^\star|\pi_{k+1})\big],$$

where $\pi_k = e^{(f_k \oplus g_k - c)/\varepsilon}\alpha \otimes \beta$ and $\pi^\star$ is the optimal entropic coupling. The KL drop controls the marginal residuals through Pinsker's inequality, Proposition 8.10. Third, convexity of the exponential dual objective gives a one-step estimate of the form

$$\Delta_k^2 \leqslant \frac{CR^2}{\varepsilon}\big(\mathcal{D}_\varepsilon(f_{k+1}, g_{k+1}) - \mathcal{D}_\varepsilon(f_k, g_k)\big).$$

This is the usual Bregman-projection estimate: the dual gap is controlled by the product of a bounded dual radius and the marginal residual corrected by the next projection, while the residual squared is controlled by the KL drop. Summing the reciprocal inequality obtained from the last display yields

$$\frac{1}{\Delta_{k+1}} - \frac{1}{\Delta_k} \geqslant \frac{\varepsilon}{CR^2},$$

and therefore $\Delta_k \leqslant CR^2/(\varepsilon(k+1))$. □

**Corollary 8.12** (Approximating unregularized OT by regularized dual costs)**.** *Consider discrete histograms* $\mathrm{a} \in \Sigma_n$, $\mathrm{b} \in \Sigma_m$ *and a finite cost matrix* C*. Let* $\mathcal{D}_{\varepsilon,k}$ *be the KL-normalized entropic dual value after* $k$ *Sinkhorn cycles, and let* $\Delta_k$ *be its dual gap. Define the entropy-corrected lower bound*

$$L_{\varepsilon,k} := \mathcal{D}_{\varepsilon,k} - \varepsilon\mathrm{H}(\mathrm{a}) - \varepsilon\mathrm{H}(\mathrm{b}), \qquad \mathrm{H}(\mathrm{a}) = -\sum_i a_i \log a_i.$$

*Then*

$$0 \leqslant \mathrm{L_C}(\mathrm{a},\mathrm{b}) - L_{\varepsilon,k} \leqslant \varepsilon\log(nm) + \Delta_k.$$

*Consequently, choosing* $\varepsilon \leqslant \delta/(2\log(nm))$ *and running Sinkhorn until* $\Delta_k \leqslant \delta/2$ *gives a* $\delta$*-accurate lower bound on the unregularized OT value. Under Proposition 8.11, the intermediate condition is* $k+1 \geqslant 2CR^2/(\varepsilon\delta)$*. With the above choice of* $\varepsilon$*, it is sufficient to take*

$$k+1 \geqslant \frac{4CR^2\log(nm)}{\delta^2},$$

*hence* $k = O(R^2\log(nm)/\delta^2)$*, up to constants and logarithmic stabilization factors.*

*Proof.* The KL-normalized objective differs from the entropy convention (7.1) 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 $E_\varepsilon$ 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 $L_{\varepsilon,k} = E_\varepsilon - \Delta_k$, which gives the displayed value bound. The final iteration estimate follows by combining $\Delta_k \leqslant CR^2/(\varepsilon(k+1))$ with the target $\Delta_k \leqslant \delta/2$. □

The same identity also yields computable stopping diagnostics. The KL drops are exactly marginal defects: after a row update the row marginal is correct and the remaining drop is measured by the column marginal, and conversely after a column update. Thus marginal violations monitor both feasibility and the remaining dual gap, up to the bounded-radius constant above.

**Algorithm 8.2** Certified entropic approximation of discrete OT

**Input:** Cost matrix $C \in \mathbb{R}^{n\times m}$, weights a, b, target accuracy $\delta > 0$.
**Output:** Certified lower bound $L_{\varepsilon,k}$ for exact OT.
**Set** $\varepsilon = \frac{\delta}{2\log(nm)}$.

**Initialize** stabilized Sinkhorn potentials and set $k = 0$, $\widehat{\Delta}_0 = +\infty$.
**While** $\widehat{\Delta}_k > \delta/2$ **do**:
  **Set** $k \leftarrow k + 1$.
  **Compute** one row soft-transform update and one column soft-transform update in stabilized log-domain variables.
  **Compute** the entropic dual value $\mathcal{D}_{\varepsilon,k}$ and the certified gap upper bound $\widehat{\Delta}_k$.

**Return** $L_{\varepsilon,k} = \mathcal{D}_{\varepsilon,k} - \varepsilon \mathrm{H}(\mathrm{a}) - \varepsilon \mathrm{H}(\mathrm{b}), \qquad 0 \leqslant \mathrm{L_C}(\mathrm{a},\mathrm{b}) - L_{\varepsilon,k} \leqslant \delta.$

## 8.4 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.

As initially explained by [93], the global convergence analysis of Sinkhorn is greatly simplified using Hilbert's projective metric on positive vectors.

**Definition 8.13** (Hilbert metric)**.** On $\mathbb{R}^n_{+,*}$, Hilbert's projective metric is

$$\forall\, (\mathrm{u},\mathrm{u}') \in (\mathbb{R}^n_{+,*})^2, \quad d_{\mathcal{H}}(\mathrm{u},\mathrm{u}') := \|\log(\mathrm{u}) - \log(\mathrm{u}')\|_V. \tag{8.3}$$

where, for vectors, $\|z\|_V = \max_i z_i - \min_i z_i$.

Multiplying both vectors by arbitrary positive constants does not change this quantity, so it is a distance only after passing to projective classes.

**Proposition 8.14** (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 [32] and [200] 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 8.15** (Birkhoff contraction theorem)**.** *Let* $\mathrm{K} \in \mathbb{R}^{n\times m}_{+,*}$, *then for* $(\mathrm{v},\mathrm{v}') \in (\mathbb{R}^m_{+,*})^2$

$$d_{\mathcal{H}}(\mathrm{Kv},\mathrm{Kv}') \leqslant \lambda(\mathrm{K}) d_{\mathcal{H}}(\mathrm{v},\mathrm{v}') \text{ where } \begin{cases} \lambda(\mathrm{K}) := \frac{\sqrt{\eta(\mathrm{K})}-1}{\sqrt{\eta(\mathrm{K})}+1} < 1 \\ \eta(\mathrm{K}) := \max\limits_{i,j,k,\ell} \frac{\mathrm{K}_{i,k}\mathrm{K}_{j,\ell}}{\mathrm{K}_{j,k}\mathrm{K}_{i,\ell}}. \end{cases}$$

*Proof.* We recall the finite-dimensional Birkhoff–Hopf estimate. For a positive linear map $A$ on a cone, define its projective diameter

$$\Delta(A) := \sup_{u,v>0} d_{\mathcal{H}}(Au, Av).$$

Then

$$d_{\mathcal{H}}(Au, Av) \leqslant \tanh(\Delta(A)/4) d_{\mathcal{H}}(u, v).$$

Indeed, after quotienting by positive scalings, write $r_k = u_k/v_k$ and normalize so that $e^{-h/2} \leqslant r_k \leqslant e^{h/2}$, where $h = d_{\mathcal{H}}(u, v)$. The ratio between two coordinates of $Au/Av$ is a quotient of two weighted averages

of the numbers $r_k$. A two-point extremal argument shows that the largest possible contraction is obtained when the mass of the two weights is placed on the two endpoints $e^{-h/2}$ and $e^{h/2}$; the cross-ratio bound defining $\Delta(A)$ then gives

$$d_{\mathcal{H}}(Au, Av) \leqslant 2 \log \frac{e^{\Delta(A)/4} e^{h/2} + e^{-\Delta(A)/4} e^{-h/2}}{e^{\Delta(A)/4} e^{-h/2} + e^{-\Delta(A)/4} e^{h/2}} \leqslant \tanh(\Delta(A)/4) h.$$

For the matrix K, its projective diameter is

$$\Delta(\mathrm{K}) = \log \eta(\mathrm{K}), \qquad \eta(\mathrm{K}) = \max_{i,j,k,\ell} \frac{\mathrm{K}_{i,k}\mathrm{K}_{j,\ell}}{\mathrm{K}_{j,k}\mathrm{K}_{i,\ell}}.$$

Therefore $\tanh(\Delta(\mathrm{K})/4) = (\sqrt{\eta(\mathrm{K})} - 1)/(\sqrt{\eta(\mathrm{K})} + 1)$, which is the claimed contraction factor. □

This result extends to arbitrary convex cones and affine mappings from the cone to its interior.

The following theorem of [93] uses Theorem 8.15 to show the linear convergence of Sinkhorn's iterations.

**Theorem 8.16** (Linear convergence of Sinkhorn). *One has* $(\mathrm{u}^{(\ell)}, \mathrm{v}^{(\ell)}) \to (\mathrm{u}^\star, \mathrm{v}^\star)$ *and*

$$d_{\mathcal{H}}(\mathrm{u}^{(\ell)}, \mathrm{u}^\star) = O(\lambda(\mathrm{K})^{2\ell}), \quad d_{\mathcal{H}}(\mathrm{v}^{(\ell)}, \mathrm{v}^\star) = O(\lambda(\mathrm{K})^{2\ell}). \tag{8.4}$$

*One also has*

$$d_{\mathcal{H}}(\mathrm{u}^{(\ell)}, \mathrm{u}^\star) \leqslant \frac{d_{\mathcal{H}}(\mathrm{P}^{(\ell)}\mathbb{1}_m, \mathrm{a})}{1 - \lambda(\mathrm{K})} \quad \text{and} \quad d_{\mathcal{H}}(\mathrm{v}^{(\ell)}, \mathrm{v}^\star) \leqslant \frac{d_{\mathcal{H}}(\mathrm{P}^{(\ell),\top}\mathbb{1}_n, \mathrm{b})}{1 - \lambda(\mathrm{K})}, \tag{8.5}$$

*where we denoted* $\mathrm{P}^{(\ell)} \coloneqq \operatorname{diag}(\mathrm{u}^{(\ell)})\mathrm{K}\operatorname{diag}(\mathrm{v}^{(\ell)})$. *Lastly, one has*

$$\|\log(\mathrm{P}^{(\ell)}) - \log(\mathrm{P}^\star)\|_\infty \leqslant d_{\mathcal{H}}(\mathrm{u}^{(\ell)}, \mathrm{u}^\star) + d_{\mathcal{H}}(\mathrm{v}^{(\ell)}, \mathrm{v}^\star) \tag{8.6}$$

*where* $\mathrm{P}^\star$ *is the unique solution of* (7.1).

*Proof.* Notice that for any $(\mathrm{v}, \mathrm{v}') \in (\mathbb{R}^m_{+,*})^2$, one has

$$d_{\mathcal{H}}(\mathrm{v}, \mathrm{v}') = d_{\mathcal{H}}(\mathrm{v}/\mathrm{v}', \mathbb{1}_m) = d_{\mathcal{H}}(\mathbb{1}_m/\mathrm{v}, \mathbb{1}_m/\mathrm{v}'),$$

since indeed $d_{\mathcal{H}}(\mathrm{a}/\mathrm{v}, \mathrm{a}/\mathrm{v}') = d_{\mathcal{H}}(\mathrm{v}, \mathrm{v}')$. This shows that

$$d_{\mathcal{H}}(\mathrm{u}^{(\ell+1)}, \mathrm{u}^\star) = d_{\mathcal{H}}\left(\frac{\mathrm{a}}{\mathrm{K}\mathrm{v}^{(\ell)}}, \frac{\mathrm{a}}{\mathrm{K}\mathrm{v}^\star}\right) = d_{\mathcal{H}}(\mathrm{K}\mathrm{v}^{(\ell)}, \mathrm{K}\mathrm{v}^\star) \leqslant \lambda(\mathrm{K}) d_{\mathcal{H}}(\mathrm{v}^{(\ell)}, \mathrm{v}^\star).$$

where we used Theorem 8.15. This shows (8.4). One also has, using the triangular inequality,

$$\begin{aligned} d_{\mathcal{H}}(\mathrm{u}^{(\ell)}, \mathrm{u}^\star) &\leqslant d_{\mathcal{H}}(\mathrm{u}^{(\ell+1)}, \mathrm{u}^{(\ell)}) + d_{\mathcal{H}}(\mathrm{u}^{(\ell+1)}, \mathrm{u}^\star) \leqslant d_{\mathcal{H}}\left(\frac{\mathrm{a}}{\mathrm{K}\mathrm{v}^{(\ell)}}, \mathrm{u}^{(\ell)}\right) + \lambda(\mathrm{K}) d_{\mathcal{H}}(\mathrm{u}^{(\ell)}, \mathrm{u}^\star) \\ &= d_{\mathcal{H}}\left(\mathrm{a}, \mathrm{u}^{(\ell)} \odot (\mathrm{K}\mathrm{v}^{(\ell)})\right) + \lambda(\mathrm{K}) d_{\mathcal{H}}(\mathrm{u}^{(\ell)}, \mathrm{u}^\star), \end{aligned}$$

which gives the first part of (8.5) since $\mathrm{u}^{(\ell)} \odot (\mathrm{K}\mathrm{v}^{(\ell)}) = \mathrm{P}^{(\ell)}\mathbb{1}_m$ (the second one being similar). The proof of (8.6) follows from [93, Lemma 3]. □

**Dual-potential form of the contraction.** The Hilbert-metric contraction above can also be read directly on the dual potentials $(f, g)$ through their variation norms. For bounded cost $c$ (e.g. on compact spaces),

$$\|f_k - f^\star\|_V = O(\lambda^k) \quad \text{and} \quad \|g_k - g^\star\|_V = O(\lambda^k)$$
$$\|\log \frac{\mathrm{d}\pi_k}{\mathrm{d}\pi^\star}\|_\infty = \|(f_k - f^\star) \oplus (g_k - g^\star)\|_\infty \leqslant \|f_k - f^\star\|_V + \|g_k - g^\star\|_V$$

where the contraction ratio is the Birkhoff factor of the Gibbs kernel $\mathrm{K}_\varepsilon = e^{-c/\varepsilon}$. Namely, with $\eta = \eta(\mathrm{K}_\varepsilon)$ as in Theorem 8.15 and $R \coloneqq \sup c - \inf c$,

$$\lambda = \frac{\sqrt{\eta} - 1}{\sqrt{\eta} + 1} \leqslant \tanh(R/(2\varepsilon)) < 1.$$

One also has the following bounds

$$\|f_k - f^\star\|_V \leqslant \frac{\|\log \frac{\mathrm{d}\pi_{k,1}}{\mathrm{d}\alpha}\|_\infty}{1-\lambda}$$

which can be used to provide a posterior estimate of the rate of convergence and serve as a stopping criterion.

The bound (8.5) shows that some error measures on the marginal constraints violation, for instance, $\|\mathrm{P}^{(\ell)}\mathbb{1}_m - \mathrm{a}\|_1$ and $\|{\mathrm{P}^{(\ell)}}^\top \mathbb{1}_n - \mathrm{b}\|_1$, are useful stopping criteria to monitor the convergence. This theorem shows that the Sinkhorn algorithm converges linearly, but the worst-case rate becomes exponentially bad as $\varepsilon \to 0$, since the global contraction factor is controlled by the cost range divided by $\varepsilon$. In practice, one often observes a much better local linear regime after enough iterations. 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.

## 8.5 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 [126]. This is the entropic counterpart of the Gaussian $\mathcal{W}_2$ formula in Proposition 2.39.

**Proposition 8.17** (Quadratic closure of Sinkhorn iterates). *Let $\beta = \mathcal{N}(\mathbf{m}_\beta, \mathbf{\Sigma}_\beta)$ 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*

$$f_1(x) = \frac{\varepsilon}{2}\log\det\left(\mathrm{Id} + \frac{2\mathbf{\Sigma}_\beta}{\varepsilon}\right) + \varepsilon\langle x - \mathbf{m}_\beta,\, (\varepsilon\mathrm{Id} + 2\mathbf{\Sigma}_\beta)^{-1}(x-\mathbf{m}_\beta)\rangle.$$

*Proof.* The exponent is the sum of a quadratic polynomial in $y$ and the logarithm of the Gaussian density of $\beta$. 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

$$\mathbb{E}\exp\left(-\frac{\|x-Y\|^2}{\varepsilon}\right) = \det\left(\mathrm{Id} + \frac{2\mathbf{\Sigma}_\beta}{\varepsilon}\right)^{-1/2} \exp\left(-\langle x-\mathbf{m}_\beta,\, (\varepsilon\mathrm{Id} + 2\mathbf{\Sigma}_\beta)^{-1}(x-\mathbf{m}_\beta)\rangle\right)$$

gives the displayed expression. □

**Proposition 8.18** (Balanced entropic OT between Gaussians). *Let $\alpha = \mathcal{N}(\mathbf{m}_\alpha, \mathbf{\Sigma}_\alpha)$ and $\beta = \mathcal{N}(\mathbf{m}_\beta, \mathbf{\Sigma}_\beta)$ with positive-definite covariances, and let*

$$\mathbf{\Sigma}_\alpha^{1/2}\mathbf{\Sigma}_\beta^{1/2} = U\,\mathrm{diag}(\sigma_i)V^\top$$

*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*

$$K_\varepsilon = \mathbf{\Sigma}_\alpha^{1/2} U\,\mathrm{diag}(s_i) V^\top \mathbf{\Sigma}_\beta^{1/2}, \qquad s_i = \frac{\sqrt{\varepsilon^2 + 16\sigma_i^2} - \varepsilon}{4\sigma_i}.$$

*The optimal value is*

$$\|\mathbf{m}_\alpha - \mathbf{m}_\beta\|^2 + \operatorname{tr}(\mathbf{\Sigma}_\alpha) + \operatorname{tr}(\mathbf{\Sigma}_\beta) + \sum_i \left(-2\sigma_i s_i - \frac{\varepsilon}{2}\log(1 - s_i^2)\right).$$

*As* $\varepsilon \downarrow 0$, $s_i \to 1$ *and the full covariance contribution, including the two trace terms, converges to* $\mathcal{B}(\mathbf{\Sigma}_\alpha, \mathbf{\Sigma}_\beta)^2$.

*Proof.* Let $(X, Y)$ be any coupling with finite second moments and cross-covariance

$$K = \mathbb{E}\big[(X - \mathbf{m}_\alpha)(Y - \mathbf{m}_\beta)^\top\big].$$

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

$$\begin{pmatrix} \mathbf{\Sigma}_\alpha & K \\ K^\top & \mathbf{\Sigma}_\beta \end{pmatrix}.$$

Write $K = \mathbf{\Sigma}_\alpha^{1/2} S \mathbf{\Sigma}_\beta^{1/2}$. 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

$$\|\mathbf{m}_\alpha - \mathbf{m}_\beta\|^2 + \operatorname{tr}(\mathbf{\Sigma}_\alpha) + \operatorname{tr}(\mathbf{\Sigma}_\beta) - 2\operatorname{tr}(K),$$

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 $\mathbf{\Sigma}_\alpha^{1/2}\mathbf{\Sigma}_\beta^{1/2}$, so $S = U\operatorname{diag}(s_i)V^\top$. 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

$$\sum_i \sigma_i = \operatorname{tr}\Big((\mathbf{\Sigma}_\alpha^{1/2}\mathbf{\Sigma}_\beta\mathbf{\Sigma}_\alpha^{1/2})^{1/2}\Big)$$

then gives the Bures–Wasserstein covariance contribution. □

**Algorithm 8.3** Closed-form Gaussian Sinkhorn coupling

**Input:** Gaussian marginals $\alpha = \mathcal{N}(\mathbf{m}_\alpha, \mathbf{\Sigma}_\alpha)$, $\beta = \mathcal{N}(\mathbf{m}_\beta, \mathbf{\Sigma}_\beta)$, scale $\varepsilon > 0$.
**Output:** Gaussian entropic coupling covariance.
**Compute singular value decomposition** $\mathbf{\Sigma}_\alpha^{1/2}\mathbf{\Sigma}_\beta^{1/2} = U\operatorname{diag}(\sigma_i)V^\top$.
**For** each $\sigma_i > 0$ **do**
  $s_i = \frac{\sqrt{\varepsilon^2 + 16\sigma_i^2} - \varepsilon}{4\sigma_i}$.

**Set** cross-covariance: $K_\varepsilon = \mathbf{\Sigma}_\alpha^{1/2} U \operatorname{diag}(s_i) V^\top \mathbf{\Sigma}_\beta^{1/2}$. **Return** Gaussian coupling with means $(\mathbf{m}_\alpha, \mathbf{m}_\beta)$, marginal covariances $(\mathbf{\Sigma}_\alpha, \mathbf{\Sigma}_\beta)$, and cross-covariance $K_\varepsilon$.

**Corollary 8.19** (Gaussian Sinkhorn divergence and smoothed Bures term)**.** *Let $\alpha = \mathcal{N}(\mathbf{m}_\alpha, \mathbf{\Sigma}_\alpha)$ and $\beta = \mathcal{N}(\mathbf{m}_\beta, \mathbf{\Sigma}_\beta)$ have positive-definite covariances. For $r > 0$, define*

$$\tau_\varepsilon(r) := \frac{\sqrt{\varepsilon^2 + 16r^2} - \varepsilon}{4r}, \qquad \psi_\varepsilon(r) := -2r\,\tau_\varepsilon(r) - \frac{\varepsilon}{2}\log\big(1 - \tau_\varepsilon(r)^2\big).$$

*If $\sigma_i(\Sigma, \Lambda)$ denotes the singular values of $\Sigma^{1/2}\Lambda^{1/2}$ and $\lambda_i(\Sigma)$ the eigenvalues of $\Sigma$, then the debiased Sinkhorn divergence (7.31) is*

$$\bar{\mathcal{L}}^{\varepsilon}_{\|\cdot-\cdot\|^2}(\alpha, \beta) = \|\mathbf{m}_\alpha - \mathbf{m}_\beta\|^2 + \mathcal{B}_\varepsilon(\mathbf{\Sigma}_\alpha, \mathbf{\Sigma}_\beta)^2,$$

*where the Gaussian covariance contribution is the debiased smoothed Bures term*

$$\mathcal{B}_\varepsilon(\Sigma, \Lambda)^2 := \sum_i \psi_\varepsilon\big(\sigma_i(\Sigma, \Lambda)\big) - \frac{1}{2}\sum_i \psi_\varepsilon\big(\lambda_i(\Sigma)\big) - \frac{1}{2}\sum_i \psi_\varepsilon\big(\lambda_i(\Lambda)\big).$$

*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 2.39.*

*Proof.* Proposition 8.18 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 $\sum_i \psi_\varepsilon(\sigma_i(\mathbf{\Sigma}_\alpha, \mathbf{\Sigma}_\beta))$. Applying the same formula to the self-costs $(\alpha, \alpha)$ and $(\beta, \beta)$ replaces these singular values by the eigenvalues of $\mathbf{\Sigma}_\alpha$ and $\mathbf{\Sigma}_\beta$. In the polarization formula (7.31), the trace terms cancel:

$$\operatorname{tr}\mathbf{\Sigma}_\alpha + \operatorname{tr}\mathbf{\Sigma}_\beta - \frac{1}{2}(2\operatorname{tr}\mathbf{\Sigma}_\alpha) - \frac{1}{2}(2\operatorname{tr}\mathbf{\Sigma}_\beta) = 0,$$

while the polarization of the squared mean terms leaves exactly $\|\mathbf{m}_\alpha - \mathbf{m}_\beta\|^2$. 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 (2.16). □

**Proposition 8.20** (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 $g_q(y) = qy^2 + \text{cst}$, then one soft transform has quadratic coefficient*

$$T_\varepsilon(q) = 1 - \frac{1}{1 - q + \varepsilon/2}, \qquad q < 1 + \varepsilon/2,$$

*and one full Sinkhorn cycle acts as $q \mapsto T_\varepsilon(T_\varepsilon(q))$. The fixed point $q_\star = T_\varepsilon(q_\star)$ is determined by*

$$A_\star^2 - \frac{\varepsilon}{2}A_\star - 1 = 0, \qquad A_\star := 1 - q_\star + \frac{\varepsilon}{2} = \frac{\varepsilon + \sqrt{\varepsilon^2 + 16}}{4}.$$

*Consequently the local asymptotic contraction factor of one full Sinkhorn cycle on the quadratic coefficient is*

$$\rho_\varepsilon = A_\star^{-4} = \left(\frac{4}{\varepsilon + \sqrt{\varepsilon^2 + 16}}\right)^4.$$

*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 $T_\varepsilon(q)$. 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

$$T_\varepsilon'(q) = -\frac{1}{(1 - q + \varepsilon/2)^2},$$

the derivative of the full-cycle map at the fixed point is $T_\varepsilon'(q_\star)^2 = A_\star^{-4}$. This gives the local asymptotic rate. □

This scalar calculation illustrates the general Gaussian convergence picture of Chizat, Delalande and Vaškevičius [64]: the rate improves when $\varepsilon$ is large or the covariance scales overlap well, and deteriorates in the small-temperature limit where the entropic coupling approaches a deterministic Brenier map.

## 8.6 Sample Complexity

This section separates two statistical regimes. Exact OT resolves geometry at all spatial scales and pays dimension-dependent empirical rates; fixed-temperature Sinkhorn divergences smooth the dual potentials and recover parametric fluctuations, at the price of regularization bias.

The sample complexity of unregularized OT suffers from the curse of dimensionality. Entropic regularization changes this picture: for a fixed $\varepsilon > 0$, Sinkhorn divergences have parametric $n^{-1/2}$ statistical rates, although the constant deteriorates when $\varepsilon \to 0$ [103, 30]. Related two-sample-testing viewpoints are developed in [192], and the large-$\varepsilon$ kernel limit connects to classical MMD tests [112]. This improvement should be balanced against the regularization bias, which vanishes only when $\varepsilon$ is sent to zero.

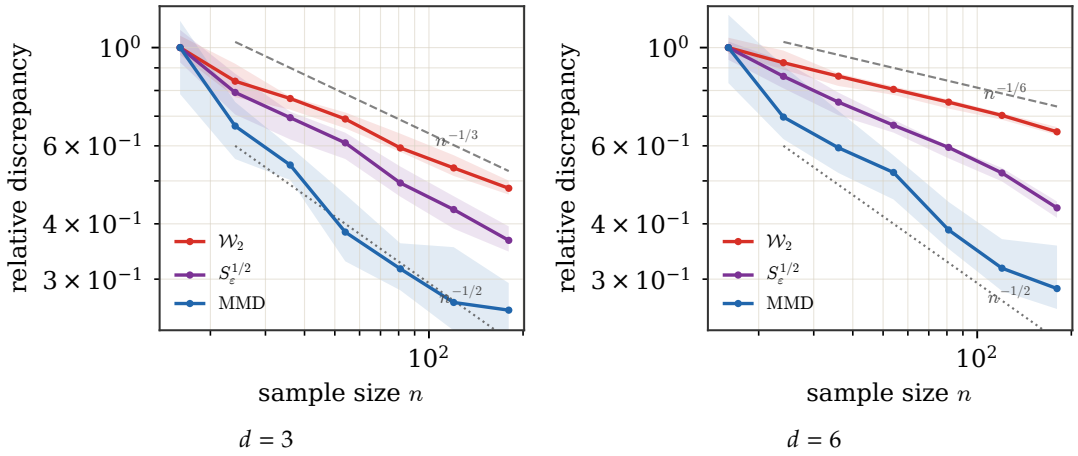



*Figure 8.1:* Empirical fluctuations in dimensions three and six. For each sample size $n$, two independent empirical measures $\alpha_n$ and $\alpha'_n$ are drawn from the same standard Gaussian law in $\mathbb{R}^d$. The curves show the median of $D(\alpha_n, \alpha'_n)$, normalized by its value at the smallest displayed $n$, with interquartile bands. Exact OT follows a slower dimension-dependent scale, while MMD and the fixed-$\varepsilon$ Sinkhorn divergence behave closer to the parametric $n^{-1/2}$ guide. This is a statistical illustration, not a solver benchmark.

**Proposition 8.21** (Empirical OT has dimension-dependent value rates)**.** *Let $\alpha$ and $\beta$ be probability distributions with densities bounded above and below on $[0,1]^d$, and let $\hat{\alpha}_n$ and $\hat{\beta}_m$ be independent empirical measures. For $d > 2p$, the expected empirical error for estimating the two-sample distance obeys*

$$\mathbb{E}\left|\mathcal{W}_p(\hat{\alpha}_n, \hat{\beta}_m) - \mathcal{W}_p(\alpha, \beta)\right| \lesssim n^{-1/d} + m^{-1/d}.$$

*The exponent changes in low dimension, but the important message is that exact OT deteriorates with the ambient dimension.*

*Proof.* By the triangle inequality,

$$\left|\mathcal{W}_p(\hat{\alpha}_n, \hat{\beta}_m) - \mathcal{W}_p(\alpha, \beta)\right| \leqslant \mathcal{W}_p(\hat{\alpha}_n, \alpha) + \mathcal{W}_p(\hat{\beta}_m, \beta).$$

We then recall the standard multiscale argument for each one-sample term, suppressing constants [82, 92, 232]. For the upper bound, partition $[0,1]^d$ into dyadic cubes. At scale $2^{-j}$, the empirical mass fluctuation over the cells is of order $n^{-1/2}2^{jd/2}$, while moving this excess mass inside cells costs $2^{-j}$. Summing the multiscale contributions up to the scale where the expected number of samples per cell is of order one gives $2^{-J}$ with $2^{Jd} \simeq n$, hence $n^{-1/d}$. The same estimate with $m$ samples gives the second term. Matching lower bounds for empirical OT follow from packing arguments; they show that this dimension dependence is intrinsic for exact OT. □

**Proposition 8.22** (MMD has a parametric value rate)**.** *Let $k$ be a bounded positive definite kernel with RKHS $\mathcal{H}_k$, and define*

$$\mathrm{MMD}_k(\alpha, \beta) = \| \int k(x, \cdot)\mathrm{d}(\alpha - \beta)(x) \|_{\mathcal{H}_k}.$$

*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. □

**Proposition 8.23** (Sinkhorn divergences interpolate the rates)**.** *Assume that $\alpha$ and $\beta$ are supported in a compact subset of $\mathbb{R}^d$ and that the cost is smooth. For fixed $\varepsilon > 0$, debiased Sinkhorn divergences satisfy representative empirical bounds of the form*

$$\mathbb{E}\left|\bar{\mathcal{L}}^\varepsilon_c(\hat{\alpha}_n, \hat{\beta}_m) - \bar{\mathcal{L}}^\varepsilon_c(\alpha, \beta)\right| \leqslant C_{c,d}\,\varepsilon^{-d/2}\left(\frac{1}{\sqrt{n}} + \frac{1}{\sqrt{m}}\right),$$

*up to constants and exponents depending on the precise smoothness class and support diameter. Thus regularization removes the $n^{-1/d}$ curse for fixed $\varepsilon$, while the prefactor deteriorates as $\varepsilon \to 0$.*

*Proof.* The proof follows the empirical-process argument of [103]. By the envelope theorem, the fluctuation of $\mathcal{L}^\varepsilon_c$ with respect to its first marginal is controlled by the class of entropic dual potentials. The soft $c$-transform smooths these potentials at spatial scale $\sqrt{\varepsilon}$ for a quadratic-type cost. Covering a bounded $d$-dimensional domain at this scale gives an effective complexity of order $\varepsilon^{-d/2}$. Standard Rademacher or Dudley entropy bounds then give an empirical-process fluctuation of order $\varepsilon^{-d/2}/\sqrt{n}$ for each marginal. Applying the same estimate to the three terms defining the debiased divergence gives the stated bound. □

> **Remark 8.24 (No free lunch when approximating exact OT).** The parametric rate in Proposition 8.23 holds for fixed $\varepsilon$. 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,
>
> $$\left|\bar{\mathcal{L}}^\varepsilon_c(\alpha, \beta) - \mathcal{L}_c(\alpha, \beta)\right| \leqslant C\varepsilon, \qquad \mathbb{E}\left|\bar{\mathcal{L}}^\varepsilon_c(\hat{\alpha}_n, \hat{\beta}_n) - \bar{\mathcal{L}}^\varepsilon_c(\alpha, \beta)\right| \leqslant C_{c,d}\varepsilon^{-d/2}n^{-1/2}.$$
>
> Balancing the two terms gives $\varepsilon \simeq n^{-1/(d+2)}$ and total error of order $n^{-1/(d+2)}$. Equivalently, target accuracy $\eta$ requires choosing $\varepsilon \simeq \eta$ and $n \simeq \eta^{-(d+2)}$ samples under this bound. Thus entropic smoothing improves the statistical behavior at fixed scale, but approximating exact OT still forces a bias-variance tradeoff whose exponent deteriorates with dimension.

CHAPTER 9

# Generalized Wasserstein Distances

The first family of extensions keeps the idea of a distance between measures, but changes the geometry used to compare them. The variants below relax mass conservation, reduce high-dimensional transport to one-dimensional projections, or replace the trace quadratic cost by spectral gauges and robust projected viewpoints. They are useful when standard $\mathcal{W}_p$ is too rigid or too expensive, while still preserving a metric interpretation.

## 9.1 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 [146, 65, 66].

**Relaxed formulation.** For nonnegative measures $(\alpha, \beta) \in \mathcal{M}_+(\mathcal{X}) \times \mathcal{M}_+(\mathcal{Y})$, a generic relaxed formulation is

$$\mathrm{UW}_c(\alpha,\beta) = \inf_{\pi \in \mathcal{M}_+(\mathcal{X}\times\mathcal{Y})} \int_{\mathcal{X}\times\mathcal{Y}} c(x,y)\mathrm{d}\pi(x,y) + \mathcal{D}_{\psi_1}(\pi_1|\alpha) + \mathcal{D}_{\psi_2}(\pi_2|\beta), \tag{9.1}$$

where $\psi_1, \psi_2$ are convex entropy functions. 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:

$$\mathrm{UW}_{c,\tau}(\alpha,\beta) = \inf_{\pi \geqslant 0} \int c\mathrm{d}\pi + \tau\mathcal{D}_{\bar{\psi}_1}(\pi_1|\alpha) + \tau\mathcal{D}_{\bar{\psi}_2}(\pi_2|\beta).$$

Large $\tau$ makes marginal mismatch expensive and approaches balanced OT when the total masses are compatible. Small $\tau$ makes creation and destruction cheap; after rescaling by $\tau$, the zero-transport part reveals the pure divergence geometry.

**Proposition 9.1** (Small-transport-scale limit for marginal penalties)**.** *Assume that $\alpha, \beta$ 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*

$$\lim_{\tau\downarrow 0} \frac{1}{\tau}\,\mathrm{UW}_{c,\tau}(\alpha,\beta) = \inf_{\rho\in\mathcal{M}_+(\mathcal{X})} \mathcal{D}_{\bar{\psi}_1}(\rho|\alpha) + \mathcal{D}_{\bar{\psi}_2}(\rho|\beta).$$

*The right-hand side is the infimal gluing divergence obtained by matching the two measures through a common zero-transport marginal $\rho$. In the dominated case, if $\alpha = a\lambda$, $\beta = b\lambda$, and the minimizing common marginal is absolutely continuous, $\rho = r\lambda$, 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,*

$$\inf_{\rho\in\mathcal{M}_+(\mathcal{X})} \mathrm{KL}(\rho|\alpha) + \mathrm{KL}(\rho|\beta) = \int (\sqrt{a} - \sqrt{b})^2\mathrm{d}\lambda.$$

*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 $\pi = (\mathrm{Id}, \mathrm{Id})_\sharp\rho$, whose transport cost is zero and whose two marginals are both $\rho$. This gives the desired upper bound after optimizing over $\rho$.

For the lower bound, let $\tau_n \downarrow 0$ and let $\pi_n$ 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 $\pi_0$ is concentrated on the diagonal. Its two marginals are therefore equal to a common measure $\rho$. Lower semicontinuity of the marginal divergences gives

$$\liminf_n \frac{1}{\tau_n}\mathrm{UW}_{c,\tau_n}(\alpha,\beta) \geqslant \mathcal{D}_{\bar\psi_1}(\rho|\alpha) + \mathcal{D}_{\bar\psi_2}(\rho|\beta),$$

and optimizing over $\rho$ gives the lower bound.

In the dominated case, writing $\rho = r\lambda$ gives

$$\mathcal{D}_{\bar\psi_1}(\rho|\alpha) + \mathcal{D}_{\bar\psi_2}(\rho|\beta) = \int a\,\bar\psi_1(r/a) + b\,\bar\psi_2(r/b)\mathrm{d}\lambda,$$

so the minimization over $\rho$ decouples into the scalar envelope $\mathfrak{m}_{\bar\psi_1,\bar\psi_2}$. For KL, no singular part is admissible when $\alpha$ and $\beta$ are dominated by $\lambda$. 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 9.2** (Dual of unbalanced optimal transport)**.** *Under the usual Fenchel–Rockafellar qualification assumptions, one has equality between* (9.1) *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 (6.13) 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

$$L_c(r,s) \coloneqq c + r\psi_1(1/r) + s\psi_2(1/s), \tag{9.2}$$

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 perspective transform of $L_c$,

$$H_c(r,s) \coloneqq \inf_{\theta>0}\theta L_c(r/\theta, s/\theta), \tag{9.3}$$

which is positively 1-homogeneous. It defines

$$\mathrm{HW}_c(\alpha,\beta) = \inf_{\pi\geqslant 0}\int H_{c(x,y)}(F(x),G(y))\mathrm{d}\pi(x,y) + \psi_1(0)\alpha^\perp(\mathcal{X}) + \psi_2(0)\beta^\perp(\mathcal{Y}). \tag{9.4}$$

**Proposition 9.3** (Homogenization does not change the unbalanced cost)**.** *One has* $\mathrm{HW}_c(\alpha,\beta) = \mathrm{UW}_c(\alpha,\beta)$.

*Proof.* The inequality HW $\leqslant$ UW follows from $H_c \leqslant L_c$ by taking $\theta = 1$. Conversely, take a feasible measure $\pi$ in the homogeneous formulation. By definition of the perspective transform, for every $(x,y)$ and every $\eta > 0$ there exists a scale $\theta(x,y) > 0$ such that

$$H_{c(x,y)}(F(x),G(y)) + \eta \geqslant \theta(x,y) L_{c(x,y)}\big(F(x)/\theta(x,y), G(y)/\theta(x,y)\big).$$

Replacing $\pi$ by the rescaled measure $\tilde\pi = \theta\pi$ and the densities by $F/\theta$ and $G/\theta$ gives an admissible competitor for the reverse formulation with cost no larger than the homogeneous cost plus $\eta\pi(\mathcal{X}\times\mathcal{Y})$. Letting $\eta \to 0$ yields UW $\leqslant$ HW. The singular terms are unchanged because the same rescaling is performed before taking the Lebesgue decomposition of the marginals. □

**Conic lifting.** Assume now that $\mathcal{X} = \mathcal{Y}$ and $\psi_1 = \psi_2 = \psi$. The last formulation lifts the problem to the cone space $\mathfrak{C}[\mathcal{X}] := (\mathcal{X}\times\mathbb{R}_+)/\!\sim$, where all points $(x,0)$ are identified at the apex. For an exponent $p \geqslant 1$, define

$$\Delta((x,r),(y,s)) := H_{c(x,y)}(r^p, s^p)^{1/p}.$$

Several classical unbalanced geometries are obtained by choosing $\psi$, $c$ and $p$ so that $\Delta$ is a distance on the cone:

- $\mathcal{D}_\psi$ = 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
$$\Delta((x,r),(y,s))^2 = r^2 + s^2 - 2rs\cos(d(x,y)\wedge\pi/2).$$
- $\mathcal{D}_\psi$ = KL, $p = 2$, and $c(x,y) = d(x,y)^2$ give the Gaussian Hellinger cone metric
$$\Delta((x,r),(y,s))^2 = r^2 + s^2 - 2rse^{-d(x,y)^2/2}.$$
- $\mathcal{D}_\psi$ = TV, $p = 1$, and $c(x,y) = d(x,y)$ give the partial-transport cone cost
$$\Delta((x,r),(y,s)) = r + s - (r\wedge s)(2 - d(x,y))_+.$$

The corresponding cone value is

$$\mathrm{CW}(\alpha,\beta) = \inf_{\gamma\in\mathcal{M}_+(\mathfrak{C}[\mathcal{X}]^2)} \left\{ \int \Delta((x,r),(y,s))^p \mathrm{d}\gamma \;;\; \int r^p \mathrm{d}\gamma_1(\cdot,r) = \alpha, \int s^p \mathrm{d}\gamma_2(\cdot,s) = \beta \right\}.$$

**Theorem 9.4** (Cone formulation of unbalanced OT)**.** *One has* UW = HW = CW. *If $\Delta$ is a distance, then* $\mathrm{CW}^{1/p}$ *is a distance between nonnegative measures.*

*Proof.* The equality UW = HW is Proposition 9.3. To prove HW = CW, disintegrate an admissible cone coupling $\gamma$ with respect to its spatial variables $(x,y)$ and radii $(r,s)$. The cone marginal constraints say precisely that the spatial marginals are recovered after weighting by $r^p$ and $s^p$. Since $\Delta((x,r),(y,s))^p = H_{c(x,y)}(r^p,s^p)$, integrating the cone cost gives the homogeneous objective. Conversely, any homogeneous competitor can be lifted to the cone by placing, over each $(x,y)$, radii whose $p$th powers are the two density factors appearing in $H_c$.

If $\Delta$ is a distance on the cone, then $\mathrm{CW}^{1/p}$ is the usual $p$-Wasserstein distance between lifted measures under the linear cone-marginal constraints. Symmetry and the triangle inequality follow from the corresponding Wasserstein properties and the gluing lemma on the cone. If the distance is zero, an optimal cone coupling is concentrated on the diagonal of the cone, so the weighted projections agree and therefore $\alpha = \beta$. □

**Entropic KL relaxation.** A generic entropic regularization of unbalanced OT reads

$$\inf_{\pi\in\mathcal{M}_+(\mathcal{X}\times\mathcal{Y})} \int c\mathrm{d}\pi + \mathcal{D}_{\psi_1}(\pi_1|\alpha) + \mathcal{D}_{\psi_2}(\pi_2|\beta) + \varepsilon\mathcal{D}_\varphi(\pi|\alpha\otimes\beta).$$

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 transforms

$$f(x) = -\frac{\tau\varepsilon}{\tau+\varepsilon}\log\int_{\mathcal{Y}} \exp\left(\frac{g(y) - c(x,y)}{\varepsilon}\right)\mathrm{d}\beta(y),$$
$$g(y) = -\frac{\tau\varepsilon}{\tau+\varepsilon}\log\int_{\mathcal{X}} \exp\left(\frac{f(x) - c(x,y)}{\varepsilon}\right)\mathrm{d}\alpha(x).$$

In the discrete case, with $K_{i,j} = e^{-\mathrm{C}_{i,j}/\varepsilon}\mathrm{a}_i\mathrm{b}_j$ and $\rho = \tau/(\tau+\varepsilon)$, this gives the generalized Sinkhorn scaling

$$u_i \leftarrow \left(\frac{\mathrm{a}_i}{(Kv)_i}\right)^\rho, \qquad v_j \leftarrow \left(\frac{\mathrm{b}_j}{(K^\top u)_j}\right)^\rho, \qquad \mathrm{P} = \mathrm{diag}(u)K\,\mathrm{diag}(v).$$

The exponent $\rho < 1$ is the visible difference with balanced Sinkhorn: marginal corrections are damped because violating the marginals is allowed.

**Algorithm 9.1** Unbalanced Sinkhorn scaling

**Input:** Weights a, b, cost matrix C, entropic scale $\varepsilon > 0$, KL strength $\tau > 0$, tolerance tol.
**Output:** Unbalanced entropic coupling P.
**Initialize:** Set $K_{ij} = e^{-\mathrm{C}_{ij}/\varepsilon}\mathrm{a}_i\mathrm{b}_j$, $\rho = \frac{\tau}{\tau+\varepsilon}$, $u^{(0)} = \mathbb{1}_n$, $v^{(0)} = \mathbb{1}_m$, $\eta_0 = +\infty$, $k = 0$.
**While** $\eta_k >$ tol **do**:
  **Set** $k \leftarrow k+1$.
  $u^{(k)} = \left(\frac{\mathrm{a}}{Kv^{(k-1)}}\right)^\rho, \qquad v^{(k)} = \left(\frac{\mathrm{b}}{K^\top u^{(k)}}\right)^\rho.$
  **Set** $\eta_k = \max\{\|u^{(k)} - u^{(k-1)}\|_\infty, \|v^{(k)} - v^{(k-1)}\|_\infty\}$.
**Return** $\mathrm{P}^{(k)} = \mathrm{diag}(u^{(k)})K\,\mathrm{diag}(v^{(k)})$.

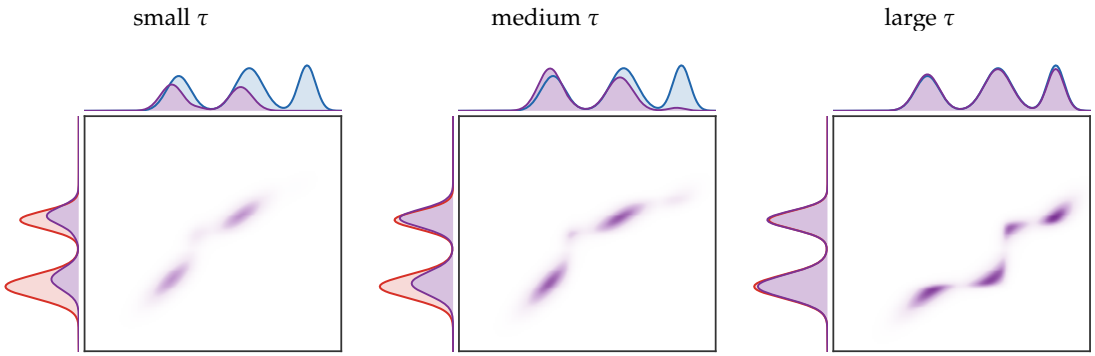



*Figure 9.1:* KL unbalanced OT on one-dimensional Gaussian-mixture densities. The central matrix is the transported coupling. On the left, the red curve is the prescribed source marginal and the violet curve is the transported source marginal; the red gap is destroyed mass. On the top, the blue curve is the prescribed target marginal and the violet curve is the transported target marginal; the blue gap is created mass. Increasing $\tau$ 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.

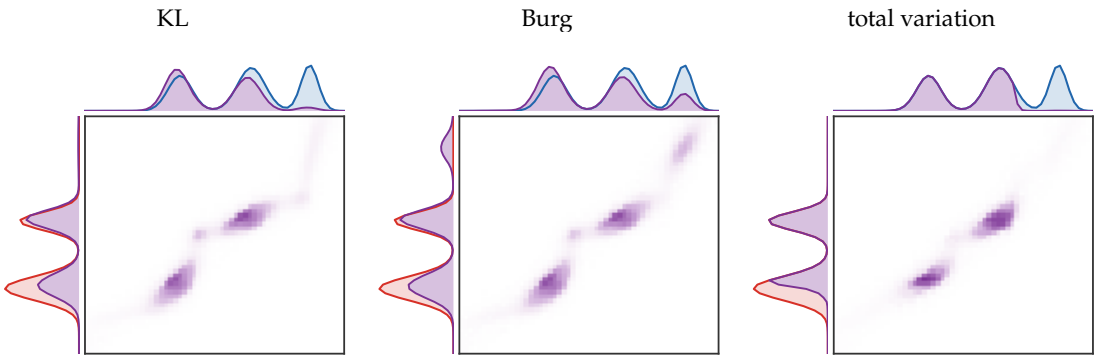



*Figure 9.2:* Effect of the marginal divergence in unbalanced entropic OT. The geometric cost, entropic plan regularization $\varepsilon$, and relaxation strength $\tau$ are fixed; only the marginal penalty changes. KL allows smooth mass variation, Burg keeps transported marginals from vanishing on prescribed modes, and total variation gives a sharper active-mass selection. The side plots use the same convention as Figure 9.1.

## 9.2 Sliced Wasserstein Distances

Sliced Wasserstein trades exact high-dimensional geometry for many one-dimensional projections. 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 $\theta$.

**Definition 9.5** (Sliced Wasserstein distance)**.** Let $\sigma$ be the uniform probability measure on the sphere $\mathbb{S}^{d-1}$. The sliced $p$-Wasserstein distance is

$$\mathrm{SW}_p(\alpha,\beta)^p := \int_{\mathbb{S}^{d-1}} \mathcal{W}_p\left((P_\theta)_\sharp \alpha, (P_\theta)_\sharp \beta\right)^p \mathrm{d}\sigma(\theta). \tag{9.5}$$

This construction is closely related to the Radon transform and is much cheaper to approximate numerically than high-dimensional OT, since each projected problem can be solved by sorting or quantiles [189, 40, 134]. 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 [170].

**Algorithm 9.2** Monte Carlo sliced Wasserstein

**Input:** Equal-weight point clouds $(x_i)_{i=1}^n$, $(y_i)_{i=1}^n$, exponent $p$, number of directions $L$.

**Output:** Monte Carlo estimate $\widehat{\mathrm{SW}}_p^p(\alpha,\beta)$.

**For** $\ell = 1, \dots, L$ **do**:

- **Sample** $\theta_\ell \sim \sigma$ on $\mathbb{S}^{d-1}$.
- **Set** $s_i^\ell = \langle \theta_\ell, x_i \rangle$ and $t_i^\ell = \langle \theta_\ell, y_i \rangle$.
- **Let** $\sigma_\ell, \tau_\ell$ be stable sorting permutations: $s^\ell_{\sigma_\ell(1)} \leqslant \cdots \leqslant s^\ell_{\sigma_\ell(n)}$, $\quad t^\ell_{\tau_\ell(1)} \leqslant \cdots \leqslant t^\ell_{\tau_\ell(n)}$.
- **Compute** $E_\ell = \frac{1}{n} \sum_{i=1}^n \left| s^\ell_{\sigma_\ell(i)} - t^\ell_{\tau_\ell(i)} \right|^p$.

**Return** $\widehat{\mathrm{SW}}_p^p(\alpha,\beta) = \frac{1}{L} \sum_{\ell=1}^L E_\ell$.

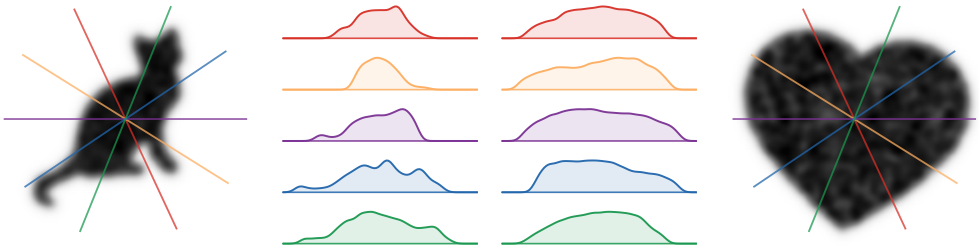



*Figure 9.3:* Sliced Wasserstein projections between two planar densities. The source and target are rendered as smoothed black-and-white density images obtained from dense farthest-point samples of two silhouettes. Five fixed directions are drawn on both densities. For each direction, the middle panels 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 9.6** (Metric properties of sliced Wasserstein)**.** *For $p \geqslant 1$, $\mathrm{SW}_p$ is a distance on $\mathcal{P}_p(\mathbb{R}^d)$. Moreover, $\mathrm{SW}_p$ metrizes weak convergence together with convergence of the $p$th moment. Finally,*

$$\mathrm{SW}_p(\alpha,\beta) \leqslant \mathcal{W}_p(\alpha,\beta),$$

*and, for $p = 2$ with the uniform probability measure on the sphere,*

$$\mathrm{SW}_2(\alpha,\beta)^2 \leqslant \frac{1}{d}\,\mathcal{W}_2(\alpha,\beta)^2.$$

*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 $\theta$ 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.

The bound $\mathrm{SW}_p \leqslant \mathcal{W}_p$ follows because $P_\theta$ is 1-Lipschitz, so

$$\mathcal{W}_p((P_\theta)_\sharp\alpha,(P_\theta)_\sharp\beta) \leqslant \mathcal{W}_p(\alpha,\beta)$$

for every $\theta$. For $p = 2$, using any coupling $\pi$ between $\alpha$ and $\beta$,

$$\int_{\mathbb{S}^{d-1}} \int |\langle x - y,\, \theta\rangle|^2 \mathrm{d}\pi(x,y)\mathrm{d}\sigma(\theta) = \frac{1}{d}\int \|x-y\|^2 \mathrm{d}\pi(x,y).$$

Optimizing over $\pi$ gives the sharper inequality. The weak-convergence statement follows from the same Cramér–Wold mechanism plus the moment condition: convergence in $\mathrm{SW}_p$ gives convergence of almost all one-dimensional projections and tightness of the $p$th moments; conversely, weak convergence with $p$th-moment convergence implies convergence of projected $\mathcal{W}_p$ distances and dominated convergence on the sphere. □

**Remark 9.7 (Hilbert embedding for $\mathrm{SW}_2$).** 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 $\mathrm{SW}_2$ 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 $\gamma > 0$. Conversely, on compact sets, $\mathcal{W}_p$ can be bounded by a dimension-dependent power of $\mathrm{SW}_p$; such inequalities are weaker than the direct bound of Proposition 9.6 and explain why sliced distances metrize the same topology without being bi-Lipschitz equivalent to $\mathcal{W}_p$ in high dimension [41, 170].

**Definition 9.8** (Max-sliced Wasserstein)**.** The max-sliced distance replaces the average over directions by the most discriminating one:

$$\mathrm{MaxSW}_p(\alpha,\beta) \coloneqq \sup_{\theta\in\mathbb{S}^{d-1}} \mathcal{W}_p((P_\theta)_\sharp\alpha,(P_\theta)_\sharp\beta).$$

It is useful when only a small set of projections carries most of the discrepancy, for instance in generative modeling [79].

**Subspace-sliced variants.** One-dimensional slices are extremely cheap, but they may discard too much geometry in high dimension. A natural compromise is to project onto $k$-dimensional subspaces: the projected OT problems remain lower dimensional, while each projection retains correlations inside a small block of coordinates. Varying $k$ therefore interpolates between ordinary slicing and full OT.

**Definition 9.9** (Subspace-sliced Wasserstein)**.** Fix $1 \leqslant k \leqslant d$. Subspace-sliced variants replace one-dimensional lines by $k$-dimensional orthogonal projections. If $U \in \mathbb{R}^{d\times k}$ satisfies $U^\top U = \mathrm{Id}_k$, then

$$\mathrm{SW}_{p,k}(\alpha,\beta)^p \coloneqq \int \mathcal{W}_p((U^\top)_\sharp\alpha,(U^\top)_\sharp\beta)^p \mathrm{d}U,$$

where $\mathrm{d}U$ denotes the normalized invariant measure on the Stiefel manifold $\mathrm{St}(d,k) = \{U \in \mathbb{R}^{d\times k} : U^\top U = \mathrm{Id}_k\}$, and

$$\mathrm{MaxSW}_{p,k}(\alpha,\beta) \coloneqq \sup_{U^\top U=\mathrm{Id}_k} \mathcal{W}_p((U^\top)_\sharp\alpha,(U^\top)_\sharp\beta).$$

The case $k = 1$ recovers ordinary sliced and max-sliced Wasserstein, while $k = d$ recovers the original Wasserstein distance.

**Proposition 9.10** (Basic bounds for sliced variants)**.** *Let $p \geqslant 1$ and let $\alpha,\beta \in \mathcal{P}_p(\mathbb{R}^d)$. With normalized spherical and Stiefel measures,*

$$\mathrm{SW}_p(\alpha,\beta) \leqslant \mathrm{MaxSW}_p(\alpha,\beta) \leqslant \mathcal{W}_p(\alpha,\beta).$$

*For k-dimensional subspace projections,*

$$\mathrm{SW}_{p,k}(\alpha,\beta) \leqslant \mathrm{MaxSW}_{p,k}(\alpha,\beta) \leqslant \mathcal{W}_p(\alpha,\beta).$$

*Proof.* The first inequality in each line follows because an $L^p$ average over a probability space is bounded by the corresponding supremum. The second inequality follows because orthogonal projections are 1-Lipschitz: pushing any admissible coupling between $\alpha$ and $\beta$ through a projection gives an admissible coupling for the projected measures with no larger transport cost. Optimizing over couplings and then averaging or maximizing over the projection gives the result. □

**Min-sliced lifted transport plans.** The preceding constructions define distances between projected measures. A different use of slicing is to use a projection only as a device for building a feasible high-dimensional transport plan. For equal-weight empirical measures $\alpha = n^{-1}\sum_i \delta_{x_i}$ and $\beta = n^{-1}\sum_i \delta_{y_i}$, sort the projected samples $\langle x_i, \theta\rangle$ and $\langle y_j, \theta\rangle$, and let $\sigma_\theta$ be the monotone matching induced by this sorting. The lifted plan

$$\pi_\theta = \frac{1}{n}\sum_{i=1}^n \delta_{(x_i,y_{\sigma_\theta(i)})}$$

is a genuine coupling between $\alpha$ and $\beta$ in the original space. Min-SWGG-type methods then choose the projection whose lifted plan has the smallest full-dimensional quadratic cost,

$$\mathrm{MSWGG}_2(\alpha,\beta)^2 \coloneqq \min_{\theta\in\mathbb{S}^{d-1}} \int \|x-y\|^2 \mathrm{d}\pi_\theta(x,y).$$

This quantity is not a projected distance; it is a cheap feasible-plan construction. Consequently it gives an upper bound on $\mathcal{W}_2^2(\alpha,\beta)$, and the interest is algorithmic: the plan is obtained by sorting rather than

by solving a high-dimensional linear program. Indeed, each $\pi_\theta$ is an admissible coupling between the original measures, so

$$\mathcal{W}_2^2(\alpha,\beta) \leqslant \int \|x-y\|^2 \mathrm{d}\pi_\theta(x,y), \qquad \mathcal{W}_2^2(\alpha,\beta) \leqslant \mathrm{MSWGG}_2(\alpha,\beta)^2.$$

**Algorithm 9.3** Lifted min-sliced matching

**Input:** Equal-weight point clouds $(x_i)_{i=1}^n$, $(y_i)_{i=1}^n$, finite direction set $\Theta \subset \mathbb{S}^{d-1}$.
**Output:** Feasible coupling $\pi_{\theta^\star}$ induced by the selected projection direction.
**For** each $\theta \in \Theta$ **do**:
  **Let** $\sigma_\theta, \tau_\theta$ be stable sorting permutations of $\langle \theta, x_i\rangle$ and $\langle \theta, y_j\rangle$.
  **Match** $x_{\sigma_\theta(k)}$ to $y_{\tau_\theta(k)}$ for $k = 1, \dots, n$.
  **Store** rank-matching permutation $\rho_\theta = \tau_\theta \circ \sigma_\theta^{-1}$.
  **Evaluate** $E(\theta) = \frac{1}{n}\sum_{i=1}^n \|x_i - y_{\rho_\theta(i)}\|^2$.

**Set** $\theta^\star = \min \operatorname{argmin}_{\theta\in\Theta} E(\theta)$ for the fixed order on $\Theta$. **Return** $\pi_{\theta^\star} = \frac{1}{n}\sum_i \delta_{(x_i, y_{\rho_{\theta^\star}(i)})}$.

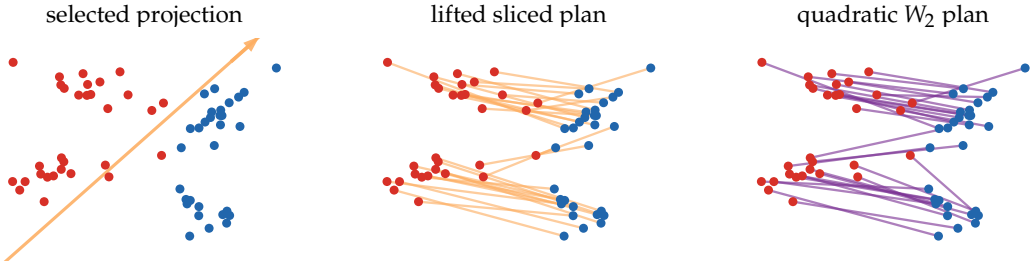



*Figure 9.4:* Lifted min-sliced plan. A one-dimensional direction is selected by a small deterministic sweep, then red and blue atoms are sorted after projection and matched in that order. The middle panel lifts this one-dimensional matching back to the plane; it is a feasible coupling but not the same object as the quadratic $W_2$ matching shown on the right. This illustrates why sliced constructions are computationally light and interpretable, while losing some of the geometry of the full transport problem.

## 9.3 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 $\rho$, typically the uniform law on a convex body or a standard Gaussian.

**Vector quantiles.** Assume that $\rho$ is absolutely continuous. For a target law $\mu$ with finite second moment, its vector quantile relative to $\rho$ is the Brenier map

$$T_\mu = \nabla\varphi_\mu, \qquad (T_\mu)_\sharp\rho = \mu,$$

or equivalently the solution of

$$\min_{T_\sharp\rho=\mu} \int \|x - T(x)\|^2 \mathrm{d}\rho(x).$$

This construction is canonical only after fixing $\rho$: changing the reference law changes the coordinates used to represent $\mu$. Vector quantile regression uses the same idea conditionally, replacing scalar conditional quantiles by conditional Brenier maps and thereby encoding multivariate ranks and depths [50].

**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 $T_\alpha$ be the Brenier map pushing $\rho$ to $\alpha$, understood as an element of $L^2(\rho;\mathbb{R}^d)$ and hence defined only $\rho$-almost everywhere. The linear OT embedding is

$$\alpha \mapsto T_\alpha - \mathrm{Id} \in L^2(\rho;\mathbb{R}^d), \qquad \mathrm{LOT}_\rho(\alpha,\beta) = \|T_\alpha - T_\beta\|_{L^2(\rho)}. \tag{9.6}$$

If one of the two targets equals the reference, the linearized distance is exact: for instance, $\mathrm{LOT}_\rho(\rho,\alpha) = \|T_\alpha - \mathrm{Id}\|_{L^2(\rho)} = \mathcal{W}_2(\rho,\alpha)$. For two arbitrary targets, the coupling $(T_\alpha, T_\beta)_\sharp\rho$ is admissible but not generally optimal, so $\mathrm{LOT}_\rho$ is a tangent-space approximation of the Wasserstein geometry [231]. For a family $(\alpha_s)_s$ with weights $(\lambda_s)_s$, the linearized barycenter is obtained by averaging maps,

$$\bar{T} = \sum_s \lambda_s T_{\alpha_s}, \qquad \bar{\alpha}_{\mathrm{LOT}} = \bar{T}_\sharp\rho.$$

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.

> **Remark 9.11 (Three Hilbertian embeddings of measures).** Several constructions in this text embed measures into Hilbert spaces, but they encode different geometries. Kernel mean embeddings send $\alpha$ to $\int k(x,\cdot)\mathrm{d}\alpha(x)$ in an RKHS and lead to MMD distances; see Section 6.2. 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 9.2. Linear OT sends $\alpha$ to the displacement field $T_\alpha - \mathrm{Id}$ from a fixed reference $\rho$ in $L^2(\rho;\mathbb{R}^d)$. 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.

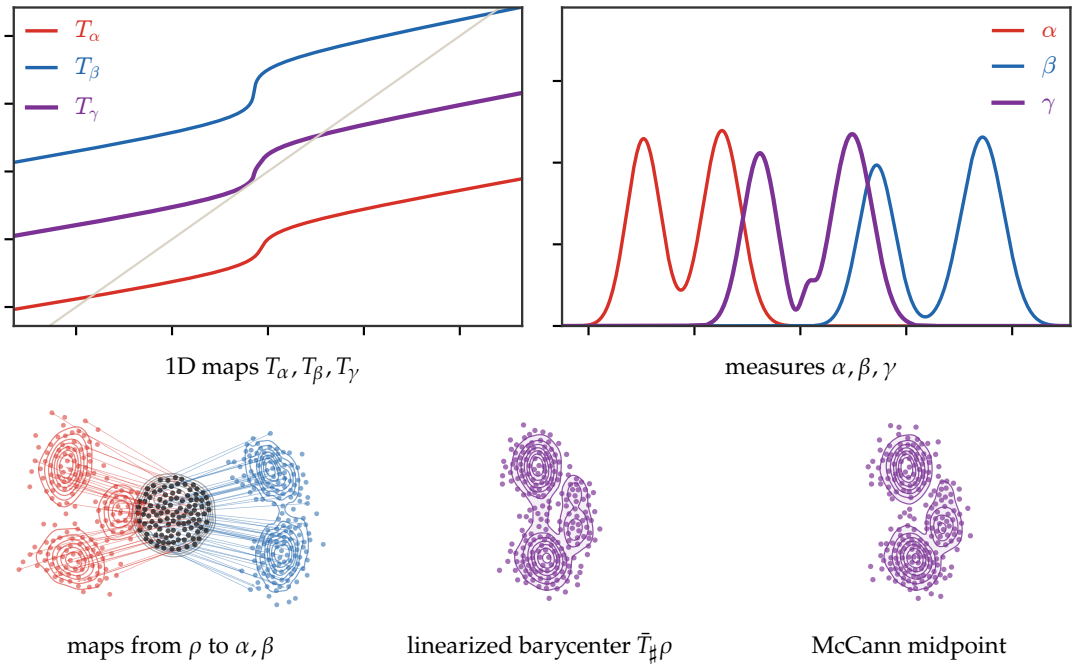



*Figure 9.5:* Linear OT coordinates. Fixing a reference measure $\rho$ turns each target into a map $T_\alpha$ from $\rho$ to $\alpha$, or equivalently into the displacement field $T_\alpha - \mathrm{Id}$. In one dimension this is exactly the quantile parametrization of $\mathcal{W}_2$, so averaging the maps toward a two-component $\alpha$ and a two-component $\beta$ 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 $\beta$ represented by two Gaussian components. The middle purple panel shows the linearized barycenter obtained by averaging the two maps from $\rho$. The right purple panel shows the genuine McCann midpoint between $\alpha$ and $\beta$, obtained by solving a direct OT problem between the two target clouds and interpolating at $t = 1/2$.

**Proposition 9.12** (Local stability of linear OT). *Assume that the measures are supported on a fixed convex compact set, with densities bounded above and below, and that the Brenier maps from $\rho$ are regular. Then, for $\alpha,\beta$ in a sufficiently small regular neighborhood of $\rho$,*

$$\mathcal{W}_2(\alpha,\beta) \leqslant \mathrm{LOT}_\rho(\alpha,\beta) \quad \textit{and} \quad \mathrm{LOT}_\rho(\alpha,\beta) \leqslant C\,\mathcal{W}_2(\alpha,\beta)^\eta$$

*for constants $C > 0$ and $\eta \in (0, 1]$ depending on regularity.*

*Proof.* The first inequality is immediate: $(T_\alpha, T_\beta)_\sharp \rho$ is a feasible coupling between $\alpha$ and $\beta$. The reverse local estimate is a standard stability statement for the Monge–Ampère equation under the stated regularity assumptions: changes in the target measure control changes in the Brenier potential in Hölder norms, hence control $T_\alpha - T_\beta$ in $L^2(\rho)$. In simple one-dimensional settings, quantile functions make this exact with $\eta = 1$. In several dimensions one should not read the statement as a global Lipschitz estimate in $\mathcal{W}_2$. Quantitative stability results for semi-discrete and Monge–Ampère maps give Hölder exponents depending on the dimension, density bounds, support geometry and regularity; see for instance the estimates of Mérigot, Delalande and Chazal [163]. Under stronger smooth perturbations of uniformly convex smooth densities, elliptic regularity can give Lipschitz dependence in stronger function norms, but converting those controls to Wasserstein perturbations generally loses powers. □

## 9.4 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 [179]; spectral gauges provide its convex minimax counterpart and connect to recent spectral-gradient viewpoints such as Muon dynamics [182].

**Definition 9.13** (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

$$0 \preceq M \preceq N \quad \Longrightarrow \quad \gamma(M) \leqslant \gamma(N).$$

The monotonicity condition means that increasing the displacement covariance in Loewner order cannot decrease the transport penalty.

**Definition 9.14** (Spectral Wasserstein distance)**.** Let $\gamma$ 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 $\gamma$ is

$$\mathcal{W}_\gamma(\alpha, \beta)^2 := \inf_{\pi \in \mathcal{U}(\alpha,\beta)} \gamma(M_\pi). \tag{9.7}$$

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

$$\mathcal{W}_{2,A}(\alpha, \beta)^2 := \inf_{\pi \in \mathcal{U}(\alpha,\beta)} \int (x - y)^\top A(x - y) \mathrm{d}\pi(x, y) = \mathcal{W}_2((A^{1/2})_\sharp \alpha, (A^{1/2})_\sharp \beta)^2. \tag{9.8}$$

The polar set of the gauge is

$$\mathcal{B}_\gamma := \{A \succeq 0 \,;\, \mathrm{tr}(AM) \leqslant \gamma(M) \quad \text{for all } M \succeq 0\}, \tag{9.9}$$

so that, for a closed gauge, $\gamma(M) = \sup_{A \in \mathcal{B}_\gamma} \mathrm{tr}(AM)$.

**Proposition 9.15** (Robust representation and metric equivalence)**.** *Assume, for simplicity, that the measures are compactly supported and that $\gamma$ 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*

$$\sqrt{a}\,\mathcal{W}_2(\alpha,\beta) \leqslant \mathcal{W}_\gamma(\alpha,\beta) \leqslant \sqrt{b}\,\mathcal{W}_2(\alpha,\beta).$$

*In particular,* $\mathcal{W}_\gamma$ *is a distance. When* $\gamma$ *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 $\gamma$,

$$\mathcal{W}_\gamma(\alpha,\beta)^2 = \inf_{\pi\in\mathcal{U}(\alpha,\beta)} \sup_{A\in\mathcal{B}_\gamma} \mathrm{tr}(AM_\pi).$$

The coupling set is convex and compact for weak convergence under compact support. Since $\gamma$ 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 $\mathrm{tr}(M) = 1$, so $\mathcal{B}_\gamma$ is convex and compact. The map $(\pi, A) \mapsto \mathrm{tr}(AM_\pi)$ is affine in each variable and continuous. Sion's minimax theorem gives

$$\inf_\pi \sup_{A\in\mathcal{B}_\gamma} \mathrm{tr}(AM_\pi) = \sup_{A\in\mathcal{B}_\gamma} \inf_\pi \mathrm{tr}(AM_\pi) = \sup_{A\in\mathcal{B}_\gamma} \mathcal{W}_{2,A}(\alpha,\beta)^2.$$

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 $a\,\mathrm{tr}(M) \leqslant \gamma(M) \leqslant b\,\mathrm{tr}(M)$ 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 9.16** (Subspace robust Wasserstein)**.** For $1 \leqslant k \leqslant d$, the Paty–Cuturi subspace robust Wasserstein distance is

$$\mathrm{SRW}_{2,k}(\alpha,\beta) := \sup_{U^\top U=\mathrm{Id}_k} \mathcal{W}_2((U^\top)_\sharp\alpha, (U^\top)_\sharp\beta) = \sup_{P^2=P=P^\top,\ \mathrm{tr}(P)=k} \mathcal{W}_{2,P}(\alpha,\beta).$$

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

$$\mathcal{B}_{\gamma_k} = \{A \,;\, 0 \preceq A \preceq \mathrm{Id},\ \mathrm{tr}(A) \leqslant k\}\,.$$

Thus $k = d$ gives $\gamma_d(M) = \mathrm{tr}(M)$ and recovers $\mathcal{W}_2$. The convex hull of rank-$k$ projectors is

$$\{A \,;\, 0 \preceq A \preceq \mathrm{Id},\ \mathrm{tr}(A) = k\}\,,$$

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. For $k = 1$, $\gamma_1(M) = \lambda_{\max}(M)$ and $\mathcal{B}_{\gamma_1} = \{A \succeq 0 \,;\, \mathrm{tr}(A) \leqslant 1\}$. This top-eigenvalue spectral Wasserstein geometry is the case connected to Muon-type spectral dynamics in [182].

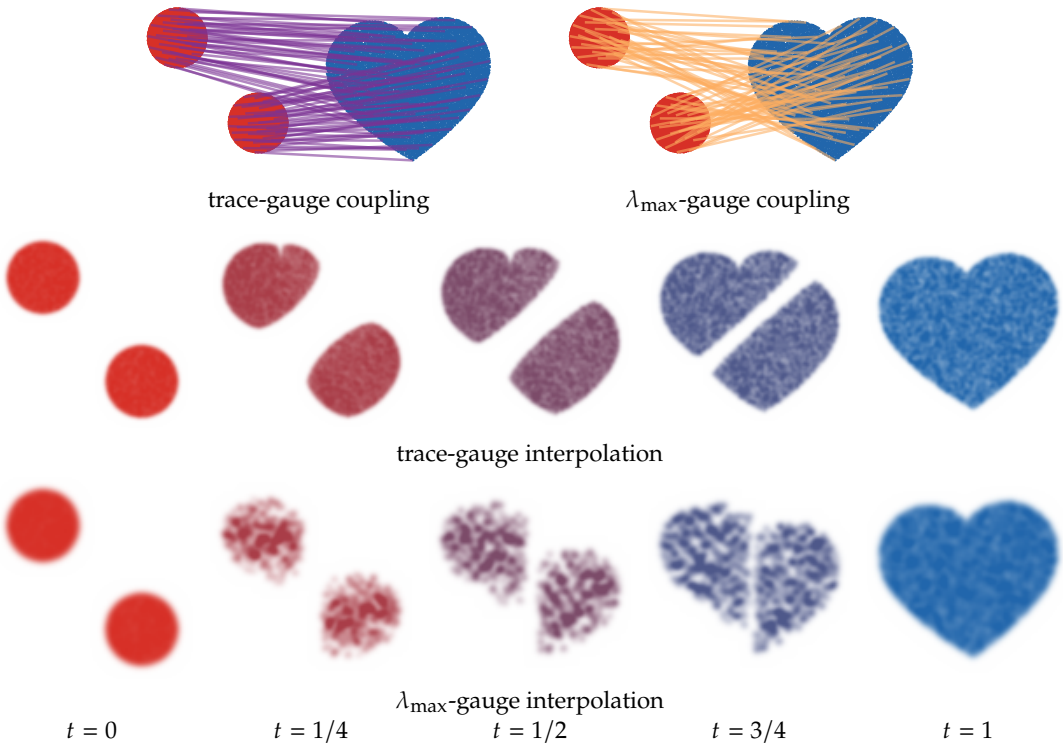



*Figure 9.6:* Trace and spectral gauges for displacement covariances. The trace gauge minimizes the average squared displacement and gives the usual quadratic transport plan from a shifted two-disk source silhouette to a shifted heart-shaped target. The $\lambda_{\max}$ gauge penalizes the worst projected 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 very dense farthest-point silhouette samples and kernel-smoothed lifted plans, with the same convention as Figure 2.2: 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.

CHAPTER 10

# Generalized OT Problems

The second family changes the optimization problem rather than only the ground distance. Barycenters average several measures, multi-marginal OT couples many measures at once, inverse OT learns the cost from observed transport, and weak OT allows randomized conditional responses. These formulations remain close to Kantorovich linear programming, but the object being optimized is richer than a single two-marginal coupling.

## 10.1 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.** For discrete input histograms $\{\mathrm{b}_s\}_{s=1}^S$, with $\mathrm{b}_s \in \Sigma_{n_s}$, and weights $\lambda \in \Sigma_S$, a Wasserstein barycenter can be computed by minimizing

$$\min_{\mathrm{a}\in\Sigma_n} \sum_{s=1}^{S} \lambda_s \mathrm{L}_{\mathrm{C}_s}(\mathrm{a}, \mathrm{b}_s), \tag{10.1}$$

where the cost matrices $\mathrm{C}_s \in \mathbb{R}^{n\times n_s}$ are prescribed.

This barycenter problem was originally introduced by [1] following earlier ideas of [51]. For the quadratic cost on $\mathcal{X} = \mathbb{R}^d$, their theory gives existence and uniqueness when at least one input measure is absolutely continuous, and more generally under hypotheses ensuring that the relevant optimal maps are well defined. Discrete existence, consistency and fixed-point constructions are further studied in [8, 5, 139].

Given a set of input measures $(\beta_s)_s$ defined on some space $\mathcal{X}$, the barycenter problem becomes

$$\min_{\alpha\in\mathcal{M}_+^1(\mathcal{X})} \sum_{s=1}^{S} \lambda_s \mathcal{L}_c(\alpha, \beta_s). \tag{10.2}$$

For $\mathcal{X} = \mathbb{R}^d$ and $c(x, y) = \|x - y\|^2$, if one input measure has a density, then the barycenter is unique [1].

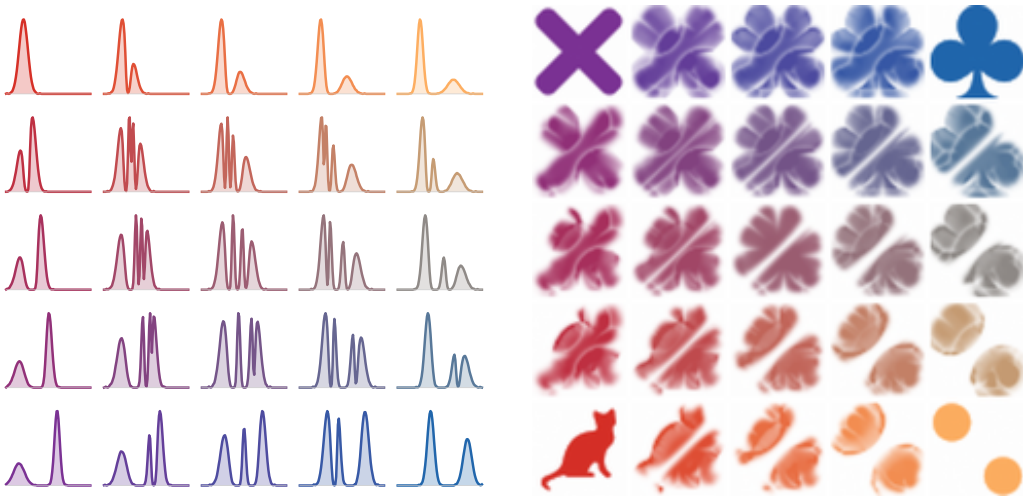
**Example 10.1 (Two measures recover a Wasserstein geodesic).** For $S = 2$, $c(x, y) = \|x - y\|^2$ and weights $(1 - t, t)$, the barycenter is the point at time $t$ on the Wasserstein geodesic between $\beta_0$ and $\beta_1$. If $T$ is the Brenier map from $\beta_0$ to $\beta_1$, this barycenter is $((1 - t)\mathrm{Id} + tT)_\sharp\beta_0$, the McCann interpolation detailed in Section 13.2. If no Monge map is available, the same construction uses an optimal coupling $\pi$ and the interpolation map $(x, y) \mapsto (1 - t)x + ty$, giving $((1 - t)x + ty)_\sharp\pi$.

**Example 10.2 (Dirac inputs recover Fréchet means).** Problem (10.2) generalizes 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 to (10.2) is $\delta_{x^\star}$ where $x^\star$ is a Fréchet mean of the points $(x_s)_s$.

**Remark 10.3 (Mean of a quadratic barycenter).** For $c(x, y) = \|x - y\|^2$, the mean of the barycenter $\alpha^\star$ is necessarily the barycenter of the means,

$$\int_{\mathcal{X}} x \mathrm{d}\alpha^\star(x) = \sum_s \lambda_s \int_{\mathcal{X}} x \mathrm{d}\beta_s(x).$$

Indeed, the squared Wasserstein distance decomposes into a squared distance between means plus a centered Wasserstein term. Minimizing the resulting quadratic function of the barycenter mean gives the displayed identity. If the input measures have compact support, the usual multi-marginal barycentric construction also gives a barycenter supported in the convex hull of the input supports.



one-dimensional quantile barycenters two-dimensional entropic barycenters

*Figure 10.1:* Wasserstein barycenter grids for four corner measures. The left panel uses the one-dimensional formula $Q_{u,v} = \sum_{i,j} \lambda_{ij}(u,v) Q_{ij}$ 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.

The next elementary proposition explains why (10.2) is a convex optimization problem over measures. The difficulty is not convexity, but the fact that the unknown is itself a measure whose support is not known in advance.

**Proposition 10.4** (Convexity of the OT cost)**.** *The map $(\alpha, \beta) \mapsto \mathcal{L}_c(\alpha, \beta)$ is convex.*

*Proof.* Let $(\alpha_0, \beta_0)$ and $(\alpha_1, \beta_1)$ be two pairs of probability measures and let $t \in [0,1]$. For $\eta > 0$, choose couplings $\pi_i \in \mathcal{U}(\alpha_i, \beta_i)$ such that

$$\int c \, \mathrm{d}\pi_i \leqslant \mathcal{L}_c(\alpha_i, \beta_i) + \eta \qquad (i = 0, 1).$$

Then $\pi_t = (1-t)\pi_0 + t\pi_1$ is a coupling between $(1-t)\alpha_0 + t\alpha_1$ and $(1-t)\beta_0 + t\beta_1$. Hence

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

Letting $\eta \to 0$ gives the claim. □

Even when all input measures are discrete, the support of a barycenter is not known a priori. The multi-marginal formulation of Section 10.2 shows that a discrete barycenter can be supported on all weighted averages of one support point from each input. This gives at most $\prod_s n_s$ candidate points if the $s$-th input has $n_s$ atoms, which is prohibitive when the number of inputs is large. A common numerical compromise is therefore to prescribe a smaller support for the barycenter and solve a fixed-support problem.

**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 10.5** (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:*

$$C^{-1}_{\alpha^\star}(r) = \sum_{s=1}^{S} \lambda_s C^{-1}_{\beta_s}(r), \qquad r \in [0,1].$$

*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 (2.18) persists: means average linearly, while covariances average according to the Bures–Wasserstein geometry.

**Example 10.6 (Gaussian inputs remain Gaussian).** The barycenter of Gaussian measures is Gaussian. In one dimension, it is obtained by averaging the means and the standard deviations, so the barycenter variance is the square of this averaged standard deviation. In higher dimensions, the covariance $\mathbf{\Sigma}$ minimizes the Bures objective

$$\mathbf{\Sigma} \mapsto \sum_s \lambda_s \mathcal{B}(\mathbf{\Sigma}, \mathbf{\Sigma}_s)^2,$$

and equivalently solves the fixed-point equation

$$\mathbf{\Sigma} = \sum_s \lambda_s \left( \mathbf{\Sigma}^{1/2} \mathbf{\Sigma}_s \mathbf{\Sigma}^{1/2} \right)^{1/2}.$$

This is the covariance analogue of the usual Euclidean barycenter equation: the mean part averages linearly, while the covariance part averages through the Bures–Wasserstein geometry [5, 29].

**Algorithm 10.1** Gaussian barycenter fixed point

**Input:** Gaussian measures $\mathcal{N}(\mathbf{m}_s, \mathbf{\Sigma}_s)$, weights $\lambda_s$, tolerance tol.
**Output:** Gaussian barycenter $\mathcal{N}(\mathbf{m}, \mathbf{\Sigma})$.
**Set** $\mathbf{m} = \sum_s \lambda_s \mathbf{m}_s$.
**Initialize:** Set $\mathbf{\Sigma}^{(0)} = \sum_s \lambda_s \mathbf{\Sigma}_s$.
**For** $k = 0, 1, \ldots$ **do**:
  $S^{(k)} = \sum_s \lambda_s \left( (\mathbf{\Sigma}^{(k)})^{1/2} \mathbf{\Sigma}_s (\mathbf{\Sigma}^{(k)})^{1/2} \right)^{1/2}$.
  $\mathbf{\Sigma}^{(k+1)} = (\mathbf{\Sigma}^{(k)})^{-1/2} \left( S^{(k)} \right)^2 (\mathbf{\Sigma}^{(k)})^{-1/2}$.
  **If** $\|\mathbf{\Sigma}^{(k+1)} - \mathbf{\Sigma}^{(k)}\| \leqslant$ tol **then**:
    **Set** $\mathbf{\Sigma} = \mathbf{\Sigma}^{(k+1)}$. **Return** $\mathcal{N}(\mathbf{m}, \mathbf{\Sigma})$.

**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 $\alpha$ is unknown. To reduce complexity, one usually fixes a candidate support for the barycenter and solves the discrete problem (10.1); this introduces a discretization error but keeps the number of unknowns manageable.

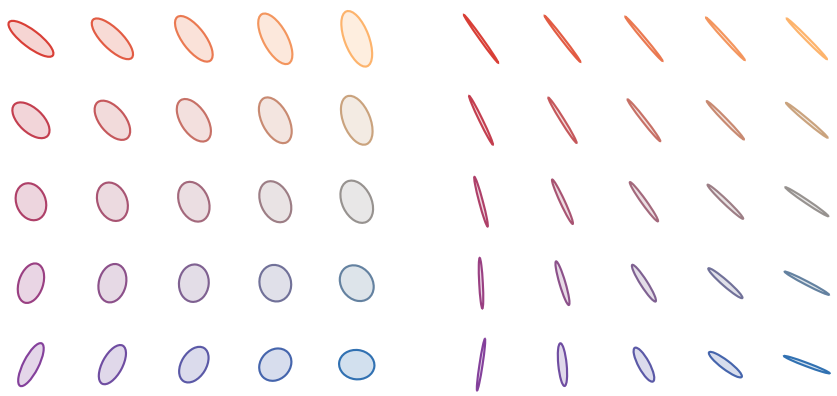
One can then use entropic smoothing and approximate (10.1) by

$$\min_{a \in \Sigma_n} \sum_{s=1}^{S} \lambda_s \mathrm{L}_{C_s}^{\varepsilon}(a, b_s) \tag{10.3}$$

for some $\varepsilon > 0$. This is a smooth convex minimization problem, which can be tackled using gradient descent [72]. An alternative is to use a descent method, typically quasi-Newton, on the semi-dual [73]; this is useful when adding extra regularization on the barycenter, for instance to impose smoothness.

A simple but effective approach, as remarked in [19], rewrites (10.3) as a weighted KL projection problem

$$\min_{(P_s)_s} \left\{ \varepsilon \sum_s \lambda_s \mathrm{KL}(P_s | K_s) \; ; \; \forall s, \; {P_s}^\top \mathbb{1}_n = b_s, \quad P_1 \mathbb{1}_{n_1} = \ldots = P_S \mathbb{1}_{n_S} \right\}, \tag{10.4}$$



moderate anisotropy strong anisotropy

*Figure 10.2:* Bures–Wasserstein barycenters of centered Gaussian covariance matrices. Each panel shows a $5 \times 5$ 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.

where $\mathrm{K}_s \coloneqq e^{-\mathrm{C}_s/\varepsilon}$. The barycenter a is implicitly encoded in the common row marginal

$$\mathrm{a} = \mathrm{P}_1 \mathbb{1}_{n_1} = \ldots = \mathrm{P}_S \mathbb{1}_{n_S}.$$

The optimal couplings solving (10.4) have scaling form

$$\mathrm{P}_s = \mathrm{diag}(\mathrm{u}_s)\mathrm{K}_s \,\mathrm{diag}(\mathrm{v}_s), \tag{10.5}$$

and the generalized Sinkhorn iterations are

$$\forall\, s \in [\![1, S]\!], \qquad \mathrm{v}_s^{(\ell+1)} \coloneqq \frac{\mathrm{b}_s}{\mathrm{K}_s{}^{\top}\mathrm{u}_s^{(\ell)}}, \tag{10.6}$$

$$\forall\, s \in [\![1, S]\!], \qquad \mathrm{u}_s^{(\ell+1)} \coloneqq \frac{\mathrm{a}^{(\ell+1)}}{\mathrm{K}_s\mathrm{v}_s^{(\ell+1)}}, \tag{10.7}$$

$$\text{where} \quad \mathrm{a}^{(\ell+1)} \coloneqq \prod_s (\mathrm{K}_s\mathrm{v}_s^{(\ell+1)})^{\lambda_s}. \tag{10.8}$$

The geometric mean in (10.8) enforces the fact that all couplings share the same barycenter marginal.

**Algorithm 10.2** Entropic barycenter Sinkhorn

**Input:** Costs $\mathrm{C}_s$, target weights $\mathrm{b}_s$, barycenter weights $\lambda_s$, regularization $\varepsilon > 0$, tolerance tol.
**Output:** Barycenter weights a and couplings $\mathrm{P}_s$.
**Initialize:** Set $\mathrm{K}_s = e^{-\mathrm{C}_s/\varepsilon}$, $\mathrm{u}_s^{(0)} = \mathbb{1}_n$ for all $s$, $r_0 = +\infty$, and $k = 0$.
**While** $r_k >$ tol **do**:
  **Set** $k \leftarrow k + 1$.
  **For** each marginal $s$ **do**
    $\mathrm{v}_s^{(k)} = \frac{\mathrm{b}_s}{\mathrm{K}_s{}^{\top}\mathrm{u}_s^{(k-1)}}$.
  **Compute** barycenter marginal: $\mathrm{a}^{(k)} = \prod_s \left(\mathrm{K}_s\mathrm{v}_s^{(k)}\right)^{\lambda_s}$.
  **For** each marginal $s$ **do**
    $\mathrm{u}_s^{(k)} = \frac{\mathrm{a}^{(k)}}{\mathrm{K}_s\mathrm{v}_s^{(k)}}$.
  **Set** $\mathrm{P}_s^{(k)} = \mathrm{diag}(\mathrm{u}_s^{(k)})\mathrm{K}_s \,\mathrm{diag}(\mathrm{v}_s^{(k)})$ for all $s$.
  **Set** $r_k = \max_s \max\{\|\mathrm{P}_s^{(k)}\mathbb{1} - \mathrm{a}^{(k)}\|_1, \|(\mathrm{P}_s^{(k)})^{\top}\mathbb{1} - \mathrm{b}_s\|_1\}$.
**Return** $\mathrm{a}^{(k)}$ and $\mathrm{P}_s^{(k)}$.

**Proposition 10.7** (Dual of entropic barycenters)**.** *The optimal scalings in* (10.5) *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*

$$\max_{(f_s,g_s)_s} \left\{ \sum_s \lambda_s \left( \langle g_s, b_s \rangle - \varepsilon \langle K_s e^{g_s/\varepsilon}, e^{f_s/\varepsilon} \rangle \right) \; ; \; \sum_s \lambda_s f_s = 0 \right\}. \tag{10.9}$$

*Proof.* Introduce Lagrange multipliers in (10.4):

$$\min_{(P_s)_s, a} \max_{(f_s,g_s)_s} \sum_s \lambda_s \Big( \varepsilon \mathrm{KL}(P_s|K_s) + \langle a - P_s \mathbb{1}_{n_s}, f_s \rangle + \langle b_s - P_s^\top \mathbb{1}_n, g_s \rangle \Big).$$

Strong duality holds, so one can exchange the minimum and maximum. The minimization with respect to $a$ gives the constraint $\sum_s \lambda_s f_s = 0$, and the 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

$$\mathrm{KL}^*(U|K) = \sum_{i,j} K_{i,j} \big( e^{U_{i,j}} - 1 \big), \tag{10.10}$$

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. Dropping constants independent of $(f_s, g_s)_s$ gives (10.9). The coordinate maximization in $g_s$ gives (10.6); the block maximization in all $(f_s)_s$ gives the common marginal (10.8) and then (10.7). □

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 [213] for applications to computer graphics and imaging.

## 10.2 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. Given measures $(\alpha_s)_{s=1}^S$ on spaces $(\mathcal{X}_s)_{s=1}^S$ and a cost $c : \mathcal{X}_1 \times \cdots \times \mathcal{X}_S \to \mathbb{R}$, the problem reads

$$\inf_{\pi \in \mathcal{U}(\alpha_1, \ldots, \alpha_S)} \int_{\mathcal{X}_1 \times \cdots \times \mathcal{X}_S} c(x_1, \ldots, x_S) \mathrm{d}\pi(x_1, \ldots, x_S),$$

where $\mathcal{U}(\alpha_1, \ldots, \alpha_S)$ is the set of probability measures whose $s$-th marginal is $\alpha_s$. This is still a linear program in the discrete setting, but its ambient tensor has size $\prod_s n_s$.

**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

$$c_{\mathrm{bar}}(x_1, \ldots, x_S) = \min_{x \in \mathbb{R}^d} \sum_{s=1}^S \lambda_s \|x - x_s\|^2.$$

**Proposition 10.8** (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 $c_{\mathrm{bar}}$, 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 $\alpha$ and couplings $\pi_s \in \mathcal{U}(\alpha, \beta_s)$, glue the couplings along their common $\alpha$ 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 $\beta_s$ 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 10.9** (Gaussian and discrete barycenters). *Quadratic Wasserstein barycenters of Gaussian measures are Gaussian. If the input measures are discrete, then there exists a barycenter supported on the set of weighted averages $\sum_s \lambda_s x_{s,i_s}$ of one support point from each input; in particular, if the $s$-th input has $n_s$ atoms, a barycenter exists with at most $\prod_s n_s$ atoms.*

*Proof.* Let the input Gaussians have means $\mathbf{m}_s$ and covariances $\mathbf{\Sigma}_s$. For any candidate barycenter $\alpha$ with mean $\mathbf{m}$ and covariance $\mathbf{\Sigma}$, Gelbrich's inequality [102], proved later in Theorem 14.18, gives

$$\mathcal{W}_2^2(\alpha, \beta_s) \geqslant \|\mathbf{m} - \mathbf{m}_s\|^2 + \mathcal{B}(\mathbf{\Sigma}, \mathbf{\Sigma}_s)^2,$$

with equality for the Gaussian law with mean $\mathbf{m}$ and covariance $\mathbf{\Sigma}$. Therefore the barycenter objective is bounded below by a function depending only on $(\mathbf{m}, \mathbf{\Sigma})$, and this lower bound is attained by the Gaussian measure with the minimizing mean and covariance. Hence at least one barycenter is Gaussian, and uniqueness in the usual nondegenerate setting gives the Gaussian barycenter mentioned above. For discrete inputs, any multi-marginal optimizer is supported on the finite product of the input supports, and $B$ maps this product to at most $\prod_s n_s$ points. □

**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. The bottleneck is the tensor size $\prod_s n_s$ in the discrete case. 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.

## 10.3 Metric learning and inverse OT

This final section points to inverse problems where the ground cost itself is learned. Such problems are typically bilevel and non-convex, but OT provides useful gradients with respect to the cost.

**Algorithm 10.3** Multi-marginal Sinkhorn

**Input:** Marginals $a_s \in \Sigma_{n_s}$, tensor cost $C$, regularization $\varepsilon > 0$, tolerance tol.
**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 $s$ and residual $r = +\infty$.
**While** $r >$ tol **do**:
  **For** $s = 1, \dots, S$ **do**:
    $(u_s)_i \leftarrow \frac{(a_s)_i}{\sum_{i_1,\dots,i_{s-1},i_{s+1},\dots,i_S} K_{i_1,\dots,i_{s-1},i,i_{s+1},\dots,i_S} \prod_{r\neq s} (u_r)_{i_r}}$.
  **Set** $P_{i_1,\dots,i_S} = K_{i_1,\dots,i_S} \prod_s (u_s)_{i_s}$.
  **Set** $r = \max_s \|(\mathrm{proj}_s)_\sharp P - a_s\|_1$.
**Return** $P$.

**Metric learning and derivatives of OT** OT is convex with respect to the measure and concave with respect to the cost. Ground-metric learning was explicitly studied in [71], and it connects to the broader metric-learning literature [137, 17].

**Proposition 10.10** (Derivative with respect to the cost)**.** *In the discrete setting, assume that the optimal coupling for* $L_C(a, b)$ *is unique and denote it by* $P^\star(C)$*. Then* $C \mapsto L_C(a, b)$ *is differentiable at* $C$ *and*

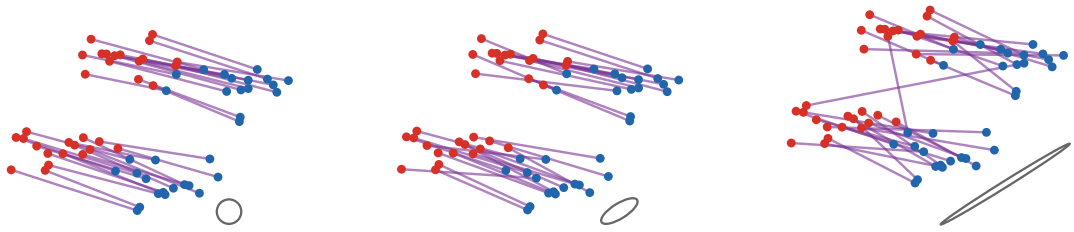
$$\nabla_C L_C(a, b) = P^\star(C).$$

*Proof.* The value is the minimum of affine functions of C,

$$L_C(a, b) = \min_{P \in U(a,b)} \langle C, P \rangle.$$

The envelope theorem, or equivalently Danskin's theorem, states that the subdifferential with respect to C is the convex hull of the optimal couplings. If the optimizer is unique, this subdifferential is the singleton $\{P^\star(C)\}$, hence the value is differentiable with the displayed gradient. □

Thus, if the cost is parameterized as $C_\theta$, gradients of losses involving OT values are obtained by backpropagating through the inner product $\langle P^\star(C_\theta), \partial_\theta C_\theta \rangle$. The difficulty is not differentiating a solved OT problem, but learning a cost for which the resulting matching has the desired semantic behavior; this is a bilevel and usually non-convex optimization problem.



Euclidean cost moderate anisotropy strong anisotropy

*Figure 10.3:* 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.

**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 $(\alpha, \beta)$ 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 potentials $u(x) + v(y)$ to a cost does not change the set of optimal couplings, and many costs can rationalize the same sparse plan.

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 [10]; curvature and identifiability of inverse OT with respect to the cost are studied in [187]. 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 $\widehat{P} \in \mathrm{U}(\mathrm{a}, \mathrm{b})$ and a cost matrix $C$, denote it by

$$\mathcal{L}_{\mathrm{iOT}}(C; \widehat{P}) = \langle C, \widehat{P} \rangle - \min_{P \in \mathrm{U}(\mathrm{a},\mathrm{b})} \langle C, P \rangle.$$

This inverse-OT gap loss is nonnegative and vanishes exactly when $\widehat{P}$ is optimal for $C$.

In practice, one restricts the cost to a finite-dimensional model class, often affine:

$$C_\theta = \sum_{r=1}^{R} \theta_r C^{(r)}, \qquad \theta \in \Theta,$$

where $\Theta$ is convex and the matrices $C^{(r)}$ encode features, graph distances or a Mahalanobis parameterization. This viewpoint appears in low-rank and sparse inverse OT models [83, 9] and in convex formulations for learning OT costs from observed plans [153, 187].

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

$$C(A)_{i,j} = \langle Ax_i, y_j \rangle,$$

so both maps $A \mapsto C(A)$ 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

$$\mathcal{P}_A \coloneqq \operatorname*{argmin}_{P \in \mathrm{U}(\mathbb{1}_n/n, \mathbb{1}_n/n)} \langle C(A), P \rangle.$$

When this face is a singleton, write its element as $P_A$; otherwise $P_A$ denotes a deterministic tie-broken selection. Although $A \mapsto C(A)$ 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 10.4 illustrates this correspondence on the OT4ML point clouds. The construction follows the visual idea of the Python Optimal Transport logo [90]: red source atoms, blue target atoms and straight segments show the selected optimal bijection. The rank-one matrices $A = -e_1 e_1^\top$ and $A = -e_2 e_2^\top$ only score horizontal or vertical correlations. The matrix $A = -I$ gives the usual quadratic $\mathcal{W}_2$ assignment, up to the marginal-only terms discussed below, while $A = +I$ reverses the correlation and produces an anti-$\mathcal{W}_2$ matching.

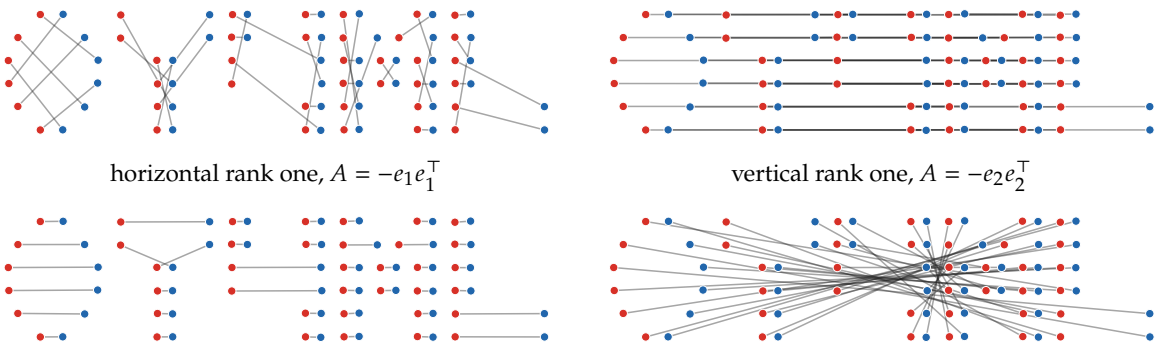



*Figure 10.4:* Forward solutions of the bilinear cost $c_A(x, y) = \langle Ax,\, y\rangle$ on the OT4ML logo point clouds. Each panel solves the equal-weight assignment problem with a different matrix $A$; the source atoms are red, the target atoms are blue, and the gray segments give one deterministic 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 $A_\star = -I$, 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 10.5 generates an observed coupling $\widehat{P}$ from this cost on two empirical mixtures of Gaussians, and then evaluates $\mathcal{L}_{\mathrm{iOT}}(C(A_t); \widehat{P})$ 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, $\widehat{P} \in \mathrm{U}(\mathbb{1}_n/n, \mathbb{1}_n/n) = \mathcal{B}_n/n$ and the plotted loss is the Kantorovich gap

$$\mathcal{L}_{\mathrm{iOT}}(C(A_t); \widehat{P}) = \langle C(A_t),\, \widehat{P}\rangle - \min_{P \in \mathrm{U}(\mathbb{1}_n/n, \mathbb{1}_n/n)} \langle C(A_t),\, P\rangle, \qquad C(A_t)_{i,j} = \langle A_t x_i,\, y_j\rangle.$$

Because $t \mapsto C(A_t)$ 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.

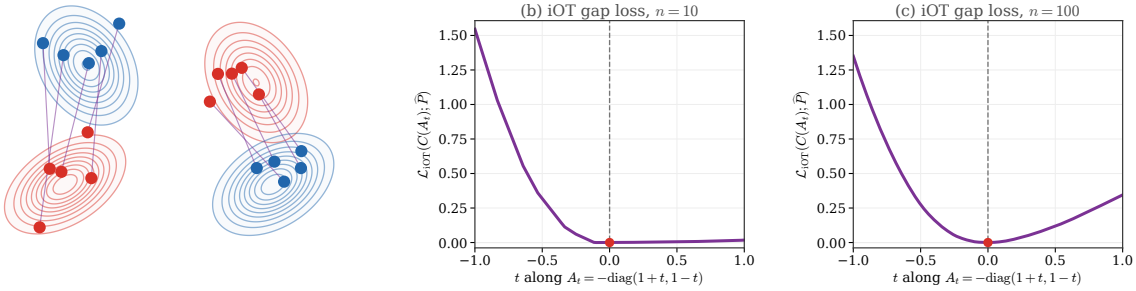
The comparison between $n = 10$ and $n = 100$ illustrates an important statistical effect: as the number of sampled points grows, the flat region of the empirical gap typically shrinks and the loss develops more visible curvature around the generating parameter. This anticipates the population theory of Peyré, Poon and Tron [187]: in the limit $n \to +\infty$, when the limiting Monge map itself has nondegenerate curvature as the cost parameter varies, the iOT loss identifies the cost robustly (up to the usual marginal-only gauge freedoms). In that regime, minimizing the gap is not only a certificate of optimality of the observed transport, but also a stable way to recover the underlying cost.

**Proposition 10.11** (Convex dual-gap formulation of inverse OT)**.** *Let $\widehat{P} \in \mathrm{U}(\mathrm{a}, \mathrm{b})$ be an observed coupling and let $C_\theta$ depend affinely on $\theta \in \Theta$, where $\Theta$ is convex. The condition that $\widehat{P}$ is optimal for the cost $C_\theta$ is equivalent to the existence of dual potentials $(f, g)$ such that*

$$f_i + g_j \leqslant (C_\theta)_{i,j} \quad \text{and} \quad \sum_{i,j} \widehat{P}_{i,j}\big((C_\theta)_{i,j} - f_i - g_j\big) = 0.$$

*Consequently, for a convex regularizer $R$, the noisy inverse problem can be relaxed as the convex program*

$$\min_{\theta \in \Theta, f, g} R(\theta) + \lambda \sum_{i,j} \widehat{P}_{i,j}\big((C_\theta)_{i,j} - f_i - g_j\big) \quad \text{subject to} \quad f_i + g_j \leqslant (C_\theta)_{i,j} \quad \forall i, j. \tag{10.11}$$




observed map, $n = 10$ loss curve, $n = 10$ loss curve, $n = 100$

*Figure 10.5:* 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 $A_\star = -I$, 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 $\mathcal{L}_{\mathrm{iOT}}(C(A_t); \widehat{P})$ 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.

*Proof.* For a fixed cost $C_\theta$, Kantorovich duality gives

$$\min_{P \in \mathrm{U}(\mathrm{a},\mathrm{b})} \langle C_\theta,\, P\rangle = \max_{f_i + g_j \leqslant (C_\theta)_{i,j}} \langle f,\, \mathrm{a}\rangle + \langle g,\, \mathrm{b}\rangle.$$

Since $\widehat{P}$ has marginals $(\mathrm{a}, \mathrm{b})$, every dual feasible pair satisfies

$$\langle C_\theta,\, \widehat{P}\rangle - \langle f,\, \mathrm{a}\rangle - \langle g,\, \mathrm{b}\rangle = \sum_{i,j} \widehat{P}_{i,j}\big((C_\theta)_{i,j} - f_i - g_j\big) \geqslant 0.$$

This nonnegative quantity is exactly the primal-dual gap of $\widehat{P}$. It vanishes if and only if $\widehat{P}$ reaches the dual value and is therefore optimal. If $C_\theta$ is affine and $\Theta$ and $R$ are convex, the constraints and objective in (10.11) are convex, proving the relaxation claim. □

**Algorithm 10.4** Inverse OT by dual-gap fitting

**Input:** Observed plan $\widehat{P} \in \mathrm{U}(\mathrm{a}, \mathrm{b})$, features $C^{(r)}$, feasible set $\Theta$, regularizer $R$.
**Output:** Identified cost $C_{\theta^\star}$ and potentials $(f^\star, g^\star)$.
**Set** parametric cost: $C_\theta = \sum_r \theta_r C^{(r)}$.
**Let** $(\theta^\star, f^\star, g^\star)$ be a minimizer of $\min_{\theta\in\Theta, f, g} R(\theta) + \lambda \sum_{i,j} \widehat{P}_{i,j}\big((C_\theta)_{i,j} - f_i - g_j\big)$
**Subject to** $f_i + g_j \leqslant (C_\theta)_{i,j}$ for all $(i, j)$. **Return** $\theta^\star$, $C_{\theta^\star}$, and $(f^\star, g^\star)$.

The formulation (10.11) is useful because it avoids differentiating through a forward OT solver: it learns a cost by making the observed plan satisfy complementary slackness. In statistical settings, $\widehat{P}$ is only partially observed or noisy, so one adds sparsity, low-rank, smoothness or metric constraints to select a meaningful cost [83, 9]. For entropic OT, the optimality condition becomes smoother:

$$\widehat{P}_{i,j} \approx a_i b_j \exp\left(\frac{f_i + g_j - (C_\theta)_{i,j}}{\varepsilon}\right),$$

which leads to likelihood-based or KL-based convex objectives when $C_\theta$ is affine, and connects inverse OT with generalized Sinkhorn iterations and transport-regularized inverse problems [130, 153]. Neural parameterizations of $C_\theta$ are more flexible but reintroduce non-convexity; the convex formulation above is the clean mathematical baseline.

## 10.4 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 10.12** (Barycentric projection of a coupling)**.** Let $\alpha, \beta \in \mathcal{P}_2(\mathbb{R}^d)$ and let $\pi \in \mathcal{U}(\alpha, \beta)$. Disintegrate $\pi$ with respect to its first marginal as $\pi(\mathrm{d}x, \mathrm{d}y) = \pi_x(\mathrm{d}y)\alpha(\mathrm{d}x)$. The barycentric projection of $\pi$ is the map

$$\bar{T}_\pi(x) := \int_{\mathbb{R}^d} y \mathrm{d}\pi_x(y), \qquad \bar{\beta}_\pi := (\bar{T}_\pi)_\sharp \alpha. \tag{10.12}$$

The projected target $\bar{\beta}_\pi$ records the distribution of conditional means, not the full second marginal. Thus it is generally different from $\beta$; 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 [7, Chap. 7].

**Proposition 10.13** (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 (10.12). Then $(\mathrm{Id}, \bar{T}_\pi)_\sharp \alpha$ is an optimal coupling between $\alpha$ and $\bar{\beta}_\pi$. Equivalently, $\bar{T}_\pi$ is a quadratic optimal transport map from $\alpha$ to the projected target $\bar{\beta}_\pi$.*

*Proof.* By Theorem 3.28, $\pi$ is concentrated on a $c$-cyclically monotone set $\Gamma$ 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 $\alpha$-negligible set, $\pi_x$ is supported on the section

$$\Gamma_x = \{ y \,;\, (x, y) \in \Gamma \}$$

for $\alpha$-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 $\alpha$ and $\bar{\beta}_\pi$. □

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 (10.12). The framework was introduced through general transport costs and weak transport inequalities in [110]; existence, duality and optimality conditions on Polish spaces are developed in [15]. For a weak cost $C : \mathcal{X} \times \mathcal{P}(\mathcal{Y}) \to \mathbb{R} \cup \{+\infty\}$, the weak OT value is

$$\mathrm{WOT}_C(\alpha, \beta) := \inf_{\pi \in \mathcal{U}(\alpha, \beta)} \int C(x, \pi_x) \mathrm{d}\alpha(x). \tag{10.13}$$

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 $\nu$.

**Proposition 10.14** (Weak Kantorovich duality)**.** *Assume that $\mathcal{X}, \mathcal{Y}$ are compact metric spaces and that $C(x, \nu)$ is lower semicontinuous, bounded from below and convex in $\nu$, with the standard qualification assumptions ensuring Fenchel–Rockafellar duality. 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 $\pi$ 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 $\alpha$, the second term becomes $\int g\mathrm{d}\beta$ because the second marginal of $\pi$ is $\beta$. This proves weak duality.

For the reverse inequality, consider the convex minimization over probability kernels $x \mapsto \pi_x$ with the affine constraint $\int \pi_x \mathrm{d}\alpha(x) = \beta$. Fenchel–Rockafellar duality gives a continuous Lagrange multiplier $g$ for this marginal constraint. Minimizing the Lagrangian over each conditional law gives exactly the pointwise term $g^C(x)$, while the multiplier contributes $\int g\mathrm{d}\beta$. The compactness, lower semicontinuity, convexity and qualification assumptions ensure no duality gap. This is the weak-cost analogue of Proposition 4.2. □

**Proposition 10.15** (Barycentric weak transport is weaker than $\mathcal{W}_2$)**.** *Let $\alpha,\beta \in \mathcal{P}_2(\mathbb{R}^d)$ and define*

$$C_{\mathrm{bar}}(x,\nu) = \|x - \int y\mathrm{d}\nu(y)\|^2.$$

*Equivalently, for a coupling $\pi$, the integrand is $\|x - \bar{T}_\pi(x)\|^2$. Then*

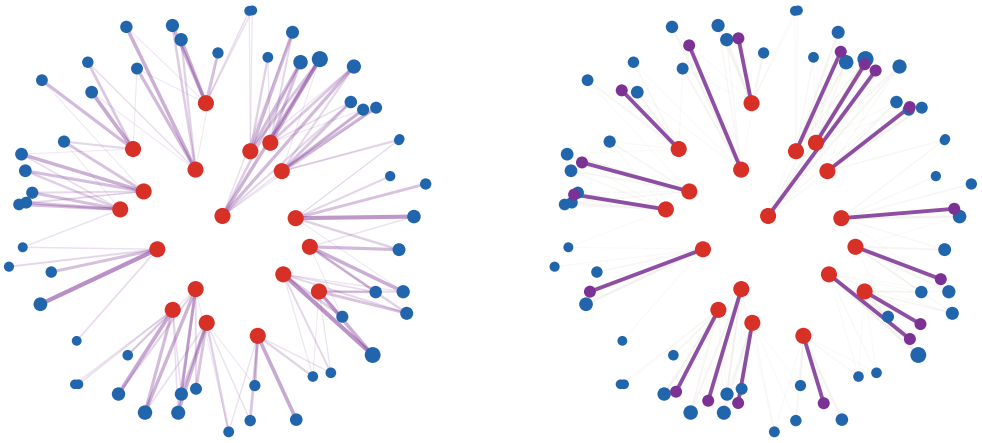
$$\mathcal{W}_{C_{\mathrm{bar}}}(\alpha,\beta) \leqslant \mathcal{W}_2^2(\alpha,\beta).$$

*Proof.* Let $\pi$ 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 $\int C_{\mathrm{bar}}(x,\pi_x)\mathrm{d}\alpha(x) \leqslant \int \|x-y\|^2\mathrm{d}\pi(x,y)$. Taking the infimum over $\pi$ proves the claim. □

The barycentric cost is the canonical example to keep in mind: admissibility still constrains the full conditional laws to have second marginal $\beta$, 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.



conditionals $\pi_x$ barycentric projection $\bar{T}_\pi$

*Figure 10.6:* 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 $\pi_x$ 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.

CHAPTER 11

# Beyond Comparing Measures

The last group 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 also has to encode structure carried by the support, the fibers or the non-commutative state space.

## 11.1 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, where the fiber remains linear and commutative but the channels may interact.

**Positive vector-valued measures.**

**Definition 11.1** (Positive vector-valued measure)**.** A positive $\mathbb{R}^m_+$-valued measure on $X$ is a tuple

$$\mu = (\mu^1, \dots, \mu^m) \in \mathcal{M}_+(X; \mathbb{R}^m_+),$$

where each component $\mu^k$ 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 $\mu_0^k(X) = \mu_1^k(X)$ for every $k$. The natural vector-valued extension of OT therefore starts from the positive cone $\mathbb{R}^m_+$.

To keep the notation readable, first assume 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 $\ell$. The conservative vector transport cost associated with an action density $\Phi$ is

$$\mathcal{W}^2_\Phi(\mu_0, \mu_1) := \inf_{u,V} \int_0^1 \int_X \Phi(u_t(x), V_t(x))\, \mathrm{d}x\, \mathrm{d}t \tag{11.1}$$

subject to the endpoint constraints $u_0 \mathrm{d}x = \mu_0$, $u_1 \mathrm{d}x = \mu_1$ and the componentwise continuity equation

$$\partial_t u_t + \nabla_x \cdot V_t = 0, \qquad (\nabla_x \cdot V_t)^k = \sum_{\ell=1}^d \partial_{x_\ell} V^k_{t,\ell}. \tag{11.2}$$

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 11.2 (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 $V_\ell$ belongs to the range of $\mathsf{M}(u)$. One chooses

$\mathsf{M}$ so that this matrix perspective is convex and one-homogeneous in $(u, V)$; this holds for the linear positive mobilities below. 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 11.3** (Diagonal positive vector Benamou–Brenier)**.** *Assume that $\mu_0^k, \mu_1^k \in \mathcal{M}_+(\mathcal{X})$ have the same mass $m_k$ for every $k$. For the diagonal mobility $\mathsf{M}_{\mathrm{diag}}$, the value of (11.1) is*

$$\mathcal{W}^2_{\mathrm{diag}}(\mu_0, \mu_1) = \sum_{k:m_k>0} m_k \, \mathcal{W}_2^2\left(\frac{\mu_0^k}{m_k}, \frac{\mu_1^k}{m_k}\right),$$

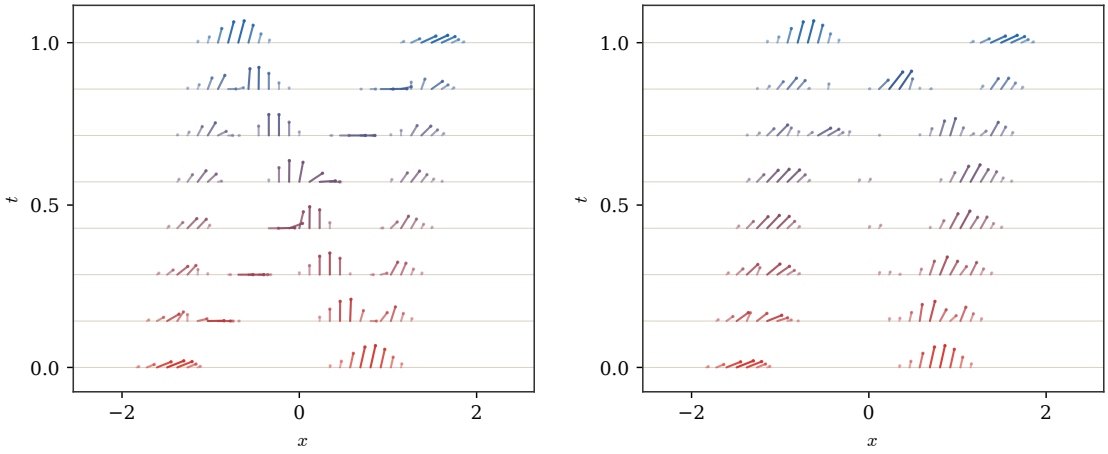
*with the convention that zero-mass channels contribute zero.*

*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 (11.2) 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 $m_k \, \mathcal{W}_2^2(\mu_0^k/m_k, \mu_1^k/m_k)$. 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 [155, 156]; related nonlinear mobility distances and gradient structures appear in [80, 165]. Figure 11.1 contrasts the exact diagonal case $\kappa = 0$, where each positive channel is transported by its quantile map, with a large-$\kappa$ 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 11.1:* 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 11.3 gives two independent scalar quantile geodesics. Right: a large-$\kappa$ 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 11.4** (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 an element

$$\mathcal{A} \in \mathcal{M}_+(\mathcal{X}; \mathbb{S}^m_+).$$

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 model minimizes the matrix-perspective action

$$\mathcal{W}^2_{\mathrm{mat}}(\mathcal{A}_0, \mathcal{A}_1) := \inf_{A,P} \int_0^1 \int_{\mathcal{X}} \sum_{\ell=1}^{d} \operatorname{tr}\left(P_{t,\ell}^\top A_t^\dagger P_{t,\ell}\right)\mathrm{d}x\mathrm{d}t \tag{11.3}$$

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

$$\partial_t A_t + \nabla_x \cdot P_t = 0, \qquad \nabla_x \cdot P_t = \sum_{\ell=1}^{d} \partial_{x_\ell} P_{t,\ell}. \tag{11.4}$$

Here $A^\dagger$ denotes the Moore–Penrose inverse, with the usual perspective convention: the action is finite only when the columns of each $P_{t,\ell}$ belong to the range of $A_t$. The map $(A, P) \mapsto \operatorname{tr}(P^\top A^\dagger P)$ is the matrix fractional function; it is jointly convex on $A \succeq 0$. 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 11.5** (Diagonal matrix subproblem)**.** *Assume that the endpoints are diagonal in a fixed orthonormal basis,*

$$\mathcal{A}_i = \operatorname{diag}(\mu_i^1, \dots, \mu_i^m), \qquad i = 0, 1,$$

*and that* $\mu_0^k(\mathcal{X}) = \mu_1^k(\mathcal{X}) = m_k$ *for every* $k$*. If one restricts the admissible curves in* (11.3) *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*

$$\sum_{k:m_k>0} m_k \, \mathcal{W}_2^2\left(\frac{\mu_0^k}{m_k}, \frac{\mu_1^k}{m_k}\right),$$

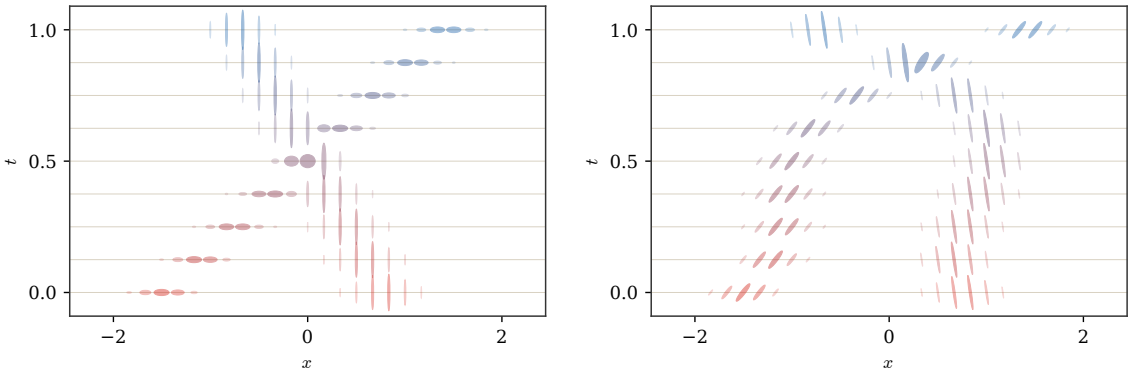
*with zero contribution from zero-mass channels. Thus the commuting matrix submodel is exactly the diagonal positive vector-valued Benamou–Brenier model of Proposition* 11.3*.*

*Proof.* The continuity equation (11.4) 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 [173, 127, 174]; dynamic versions and related non-commutative geometries appear in [61, 59, 49, 184]. Figure 11.2 shows the analogous independent/coupled contrast for positive $2 \times 2$ 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 11.2:* Positive $2 \times 2$ 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.

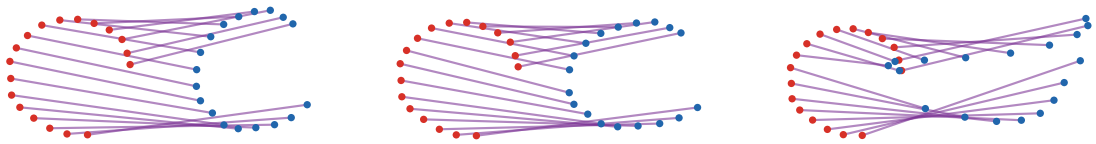
## 11.2 Gromov–Wasserstein

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. The Gromov–Wasserstein problem reads

$$\mathrm{GW}((a, D), (b, D'))^p := \min_{P\in U(a,b)} \mathcal{E}_{D,D'}(P) := \sum_{i,j,i',j'} \Delta(D_{i,i'}, D'_{j,j'})^p P_{i,j} P_{i',j'}, \tag{11.5}$$

where $p \geqslant 1$ and $\Delta$ is a distance on $\mathbb{R}$, typically $\Delta(u, v) = |u - v|$. This is a non-convex quadratic problem over the transport polytope. In the uniform case with the additional hard-assignment constraint $m = n$ and P a permutation matrix, it becomes a Quadratic Assignment Problem (QAP) [150]; this restricted graph-matching form is already NP-hard in full generality. The relaxed coupling formulation used in GW can therefore be read as a soft graph-matching model [152].



isometric copy mild deformation strong deformation

*Figure 11.3:* 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 the matrices $D, D'$ are genuine distance matrices, the general construction below shows that GW satisfies the triangle inequality and defines a distance between metric spaces equipped with a probability distribution, up to measure-preserving isometries. This distance was introduced and studied in detail by Memoli in [160]. An in-depth mathematical exposition (in particular, its geodesic structure and gradient flows) is given in [221]. See also [203] for applications in computer vision. Its relation to Hausdorff and Gromov–Hausdorff distances is discussed at the end of this section.

**General setting.**

**Definition 11.6** (Metric-measure space)**.** A metric-measure space is a triple

$$\mathbb{X} = (\mathcal{X}, d_{\mathcal{X}}, \alpha),$$

where $(\mathcal{X}, d_{\mathcal{X}})$ is a metric space and $\alpha$ is a probability measure on $\mathcal{X}$.

The general setting corresponds to computing couplings between metric-measure spaces $\mathbb{X} = (\mathcal{X}, d_{\mathcal{X}}, \alpha)$ and $\mathbb{Y} = (\mathcal{Y}, d_{\mathcal{Y}}, \beta)$, where the distance and the measure are both part of the data. The natural setting is that of Polish metric spaces; compactness is often assumed in this section to avoid existence and integrability issues. One defines

$$\mathcal{GW}(\mathbb{X}, \mathbb{Y})^p := \min_{\pi\in\mathcal{U}(\alpha,\beta)} \int_{\mathcal{X}^2\times\mathcal{Y}^2} \Delta(d_{\mathcal{X}}(x, x'), d_{\mathcal{Y}}(y, y'))^p \mathrm{d}\pi(x, y)\mathrm{d}\pi(x', y'). \tag{11.6}$$

**Proposition 11.7** (Euclidean GW is controlled by Wasserstein)**.** *Let $\alpha, \beta$ be probability measures on $\mathbb{R}^d$, equipped with the Euclidean distance, and take $\Delta(u, v) = |u - v|$ in (11.6). Then*

$$\mathcal{GW}((\mathbb{R}^d, \|\cdot\|, \alpha), (\mathbb{R}^d, \|\cdot\|, \beta)) \leqslant 2\mathcal{W}_p(\alpha, \beta).$$

*Proof.* Let $\pi$ be any coupling between $\alpha$ and $\beta$. For two independent pairs $(X, Y), (X', Y') \sim \pi$, the reverse triangle inequality gives

$$\big|\|X - X'\| - \|Y - Y'\|\big| \leqslant \|X - Y\| + \|X' - Y'\|.$$

Taking the $L^p$ norm and using Minkowski gives a bound by $2(\int \|x - y\|^p \mathrm{d}\pi)^{1/p}$. Optimizing over $\pi$ proves the claim. □

To turn GW from a distortion score into a metric statement, one must quotient out the relabelings that preserve both distances and mass.

**Definition 11.8** (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 measurable map $\varphi : \operatorname{supp}(\alpha) \to \operatorname{supp}(\beta)$ such that $\varphi_\sharp \alpha = \beta$, $\varphi(\operatorname{supp}(\alpha)) = \operatorname{supp}(\beta)$, and

$$d_{\mathcal{Y}}(\varphi(x), \varphi(x')) = d_{\mathcal{X}}(x, x')$$

for all $x, x' \in \operatorname{supp}(\alpha)$.

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 11.9** (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.*

*Proof.* If $\mathcal{GW}(\mathbb{X}, \mathbb{Y}) = 0$ and $\pi$ 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 $\operatorname{supp}(\pi)^2$. We show that $\mathbb{X}$ and $\mathbb{Y}$ are isometric by showing that both are isometric to the support space $(\operatorname{supp}(\pi), d_\pi, \pi)$, 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 $((x, y), (x', y')) \in \operatorname{supp}(\pi)^2$,

$$d_{\mathcal{X}}(\psi(x, y), \psi(x', y')) = d_{\mathcal{X}}(x, x') = d_{\mathcal{Y}}(y, y') = d_\pi((x, y), (x', y')),$$

so $\psi$ is an isometry and therefore injective. To see surjectivity onto $\operatorname{supp}(\alpha)$, take $x \in \operatorname{supp}(\alpha)$. Since $\psi_\sharp \pi = \alpha$, there is a sequence $(x_k, y_k) \in \operatorname{supp}(\pi)$ with $x_k \to x$. The equality of distances on $\operatorname{supp}(\pi)$ makes $(y_k)_k$ Cauchy, and compactness gives a convergent subsequence with limit $(x, y) \in \operatorname{supp}(\pi)$. The same argument for the second projection shows that the support space is also isometric to $\mathbb{Y}$.

For the triangle inequality, let $\pi$ be an optimal coupling between $\mathbb{X}$ and $\mathbb{Y}$, and $\xi$ an optimal coupling between $\mathbb{Y}$ and $\mathbb{Z} = (Z, d_Z, \gamma)$. By the gluing lemma, take $\sigma$ on $\mathcal{X} \times \mathcal{Y} \times Z$ whose $(\mathcal{X}, \mathcal{Y})$ and $(\mathcal{Y}, Z)$ marginals are $\pi$ and $\xi$. 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 $\rho$ 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. Symmetry and non-negativity are immediate. □

The metric structure also gives geodesics. Sturm's construction is useful conceptually because it allows one to speak about 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 [221].

**Proposition 11.10** (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 (11.6), and let $\pi^\star$ be an optimal coupling. Define, on $Z = \mathcal{X}_0 \times \mathcal{X}_1$,*

$$d_t((x_0, x_1), (x_0', x_1')) \coloneqq (1 - t) d_{\mathcal{X}_0}(x_0, x_0') + t\, d_{\mathcal{X}_1}(x_1, x_1'), \qquad \mathbb{X}_t = (Z, d_t, \pi^\star).$$

*At $t = 0$ and $t = 1$, and possibly in degenerate cases, one quotients $Z$ by the zero-distance relation associated with $d_t$. Then $t \mapsto \mathbb{X}_t$ is a constant-speed geodesic:*

$$\mathcal{GW}(\mathbb{X}_s, \mathbb{X}_t) = |t - s| \, \mathcal{GW}(\mathbb{X}_0, \mathbb{X}_1) \qquad \forall s, t \in [0, 1].$$

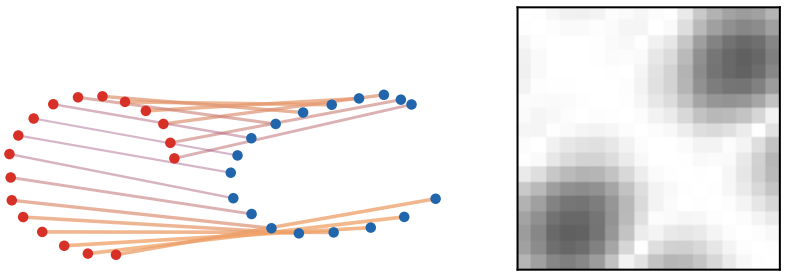
*Proof.* Write $D = \mathcal{GW}(\mathbb{X}_0, \mathbb{X}_1)$. For $s < t$, couple $\mathbb{X}_s$ and $\mathbb{X}_t$ by the diagonal coupling induced by the identity on $Z$ and the measure $\pi^\star$. For two independent points $z = (x_0, x_1)$ and $z' = (x_0', x_1')$ sampled from $\pi^\star$,

$$d_t(z, z') - d_s(z, z') = (t - s)\big(d_{\mathcal{X}_1}(x_1, x_1') - d_{\mathcal{X}_0}(x_0, x_0')\big).$$

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. □



hard correspondence $\sigma$ residual $|d_{\mathcal{X}} - d_{\mathcal{Y}} \circ \sigma|$

*Figure 11.4:* 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.

We now record the profile lower bound used above as a useful initialization principle for the non-convex solver.

**Proposition 11.11** (Mémoli profile lower bound)**.** *Let $\mathbb{X} = (\mathcal{X}, d_{\mathcal{X}}, \alpha)$ and $\mathbb{Y} = (\mathcal{Y}, d_{\mathcal{Y}}, \beta)$ be compact metric-measure spaces and take $\Delta(u, v) = |u - v|$ in (11.6), with the same exponent $p \geqslant 1$. For each $x \in \mathcal{X}$ and $y \in \mathcal{Y}$, define the distance-profile measures on $\mathbb{R}_+$ by*

$$\alpha_x := (d_{\mathcal{X}}(x, \cdot))_\sharp \alpha, \qquad \beta_y := (d_{\mathcal{Y}}(y, \cdot))_\sharp \beta.$$

*Let $\mathsf{E}_{\mathbb{X}} = (x \mapsto \alpha_x)_\sharp \alpha$ and $\mathsf{E}_{\mathbb{Y}} = (y \mapsto \beta_y)_\sharp \beta$, which are probability measures on $\mathcal{P}(\mathbb{R}_+)$. Then*

$$\mathcal{W}_p\big(\mathsf{E}_{\mathbb{X}}, \mathsf{E}_{\mathbb{Y}}\big) \leqslant \mathcal{GW}(\mathbb{X}, \mathbb{Y}).$$

*Here the left-hand distance is taken on the space $\mathcal{P}(\mathbb{R}_+)$ of profile measures. Its ground cost is the one-dimensional Wasserstein distance $\mathcal{W}_p$.*

*Proof.* Fix any $\pi \in \mathcal{U}(\alpha, \beta)$. It induces a coupling $(x, y) \mapsto (\alpha_x, \beta_y)$ between $\mathsf{E}_{\mathbb{X}}$ and $\mathsf{E}_{\mathbb{Y}}$, hence

$$\mathcal{W}_p\big(\mathsf{E}_{\mathbb{X}}, \mathsf{E}_{\mathbb{Y}}\big)^p \leqslant \int_{\mathcal{X} \times \mathcal{Y}} \mathcal{W}_p(\alpha_x, \beta_y)^p \, \mathrm{d}\pi(x, y).$$

For fixed $(x, y)$, the map $(x', y') \mapsto (d_{\mathcal{X}}(x, x'), d_{\mathcal{Y}}(y, y'))$ pushes the same coupling $\pi$ to a coupling between $\alpha_x$ and $\beta_y$. Therefore

$$\mathcal{W}_p(\alpha_x, \beta_y)^p \leqslant \int_{\mathcal{X} \times \mathcal{Y}} |d_{\mathcal{X}}(x, x') - d_{\mathcal{Y}}(y, y')|^p \, \mathrm{d}\pi(x', y').$$

Integrating in $(x, y)$ gives

$$\mathcal{W}_p\big(\mathsf{E}_{\mathbb{X}}, \mathsf{E}_{\mathbb{Y}}\big)^p \leqslant \int_{\mathcal{X}^2 \times \mathcal{Y}^2} |d_{\mathcal{X}}(x, x') - d_{\mathcal{Y}}(y, y')|^p \, \mathrm{d}\pi(x, y) \mathrm{d}\pi(x', y').$$

Taking the infimum over $\pi$ and then the $p$-th root proves the claim. □

This lower bound is useful computationally because the profile cost matrix $C_{ij} = \mathcal{W}_p(\alpha_{x_i}, \beta_{y_j})^p$ 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.

**Entropic regularization and iterative solver.** For the common squared distortion $\Delta(u, v)^2 = (u - v)^2$, one often computes a stationary point of the entropic relaxation

$$\min_{P \in U(a,b)} \mathcal{E}_{D,D'}(P) - \varepsilon H(P). \tag{11.7}$$

Although the objective is non-convex, successive linearizations lead to a practical mirror-descent scheme [186]. Up to an irrelevant global factor in the gradient, one alternates

$$P^{(\ell+1)} := \min_{P \in U(a,b)} \langle P, C^{(\ell)} \rangle - \varepsilon H(P), \qquad C^{(\ell)} := D^{\odot 2} a \, \mathbb{1}_m^\top + \mathbb{1}_n (D'^{\odot 2} b)^\top - 2 D \, P^{(\ell)} \, D'^\top. \tag{11.8}$$

Each update is an ordinary entropic OT problem and can therefore be solved with Sinkhorn iterations. This is the standard entropic GW solver used to compute soft maps between domains; it improves scalability and smooths the landscape, but it does not remove the non-convexity of the GW objective.

**Algorithm 11.1** Entropic Gromov–Wasserstein linearization

**Input:** Metric matrices $D, D'$, weights $a, b$, regularization $\varepsilon > 0$, tolerance tol.
**Output:** Approximate entropic GW coupling $P \in U(a, b)$.
**Initialize:** Set $P^{(0)} = a \otimes b$.
**For** $k = 0, 1, \ldots$ **do**:
  $C^{(k)} = D^{\odot 2} a \, \mathbb{1}_m^\top + \mathbb{1}_n (D'^{\odot 2} b)^\top - 2 D \, P^{(k)} \, D'^\top$.
  **Solve** entropic OT subproblem: $P^{(k+1)} = \operatorname{argmin}_{P \in U(a,b)} \langle P, C^{(k)} \rangle - \varepsilon H(P)$.
  **If** $\|P^{(k+1)} - P^{(k)}\| \leqslant$ tol **then**:
    **Return** $P^{(k+1)}$.

**Adding features.** Fused Gromov–Wasserstein augments the structural term with a feature transport cost [224]. In the discrete case, given a cross-feature cost $M \in \mathbb{R}^{n \times m}$ and a parameter $\lambda \in [0, 1]$, one minimizes

$$\mathrm{FGW}_{\lambda,p}((a, D), (b, D'))^p := \min_{P \in U(a,b)} (1 - \lambda) \sum_{i,j} M_{ij} P_{ij} + \lambda \sum_{i,j,i',j'} \Delta(D_{ii'}, D'_{jj'})^p P_{ij} P_{i'j'}.$$

The first term compares node attributes in the usual OT sense, while the second compares the intrinsic geometry. The endpoints $\lambda = 0$ and $\lambda = 1$ recover feature-only OT and pure GW respectively; intermediate values trade attribute matching against structural matching. This is useful when the two spaces have both distances and features, and these two sources of information may disagree.

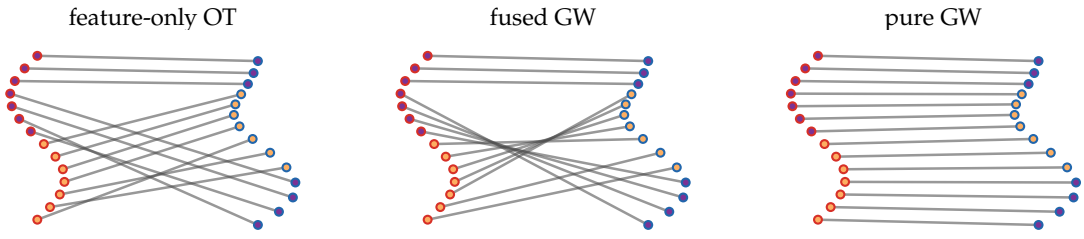



*Figure 11.5:* Feature information and intrinsic geometry in fused Gromov–Wasserstein. Small inner disks encode binary node features. Feature-only OT follows the attributes even when this crosses the shape structure, pure GW follows the intrinsic ordering, and fused GW balances the feature term with the pairwise-distance distortion.

**Hausdorff and Gromov–Hausdorff viewpoints.** If $A, B$ are compact subsets of a common metric space $(Z, d_Z)$, their Hausdorff distance is

$$d_{\mathrm{H}}^Z(A, B) = \max \left\{ \sup_{a \in A} \inf_{b \in B} d_Z(a, b), \sup_{b \in B} \inf_{a \in A} d_Z(a, b) \right\}.$$

The Gromov–Hausdorff distance removes the common ambient space by minimizing this quantity over all isometric embeddings into a third space:

$$d_{\mathrm{GH}}(\mathcal{X},\mathcal{Y}) = \inf_{Z,\varphi,\psi} d_{\mathrm{H}}^{Z}(\varphi(\mathcal{X}),\psi(\mathcal{Y})).$$

Equivalently, it is half the minimal distortion of a correspondence between $\mathcal{X}$ and $\mathcal{Y}$ [114, 159]. This is a worst-case set distance: every point must be matched with small distortion. Gromov–Wasserstein replaces correspondences by probability couplings and worst-case distortion by averaged distortion. It is therefore better adapted to noisy sampled shapes and weighted graphs, but it can ignore small sets of mass that would dominate the Hausdorff distance.

## 11.3 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 [173, 61, 59, 184, 108, 55].

**Finite-dimensional states and couplings.**

**Definition 11.12** (Hermitian and density matrices)**.** Let $\mathbb{H}_n$ be the real vector space of $n \times n$ Hermitian matrices,

$$\mathbb{H}_n^+ = \{A \in \mathbb{H}_n \,;\, A \succeq 0\}, \qquad \mathbb{H}_n^{+,1} = \left\{A \in \mathbb{H}_n^+ \,;\, \operatorname{tr}(A) = 1\right\}.$$

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

$$\operatorname{tr}(F\ \operatorname{Tr}_B T) = \operatorname{tr}((F \otimes \mathrm{Id}_m)T), \qquad \operatorname{tr}(G\ \operatorname{Tr}_A T) = \operatorname{tr}((\mathrm{Id}_n \otimes G)T), \tag{11.9}$$

for all $F \in \mathbb{H}_n$ and $G \in \mathbb{H}_m$. Thus $\operatorname{Tr}_B(T) \in \mathbb{H}_n^+$ and $\operatorname{Tr}_A(T) \in \mathbb{H}_m^+$ 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 11.13** (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

$$\mathrm{QOT}_C(A,B) \coloneqq \min_{T \in \mathbb{H}_{nm}^+} \{\operatorname{tr}(CT) \,;\, \operatorname{Tr}_B(T) = A,\ \operatorname{Tr}_A(T) = B\}. \tag{11.10}$$

**Example 11.14 (Classical diagonal case).** If $A$, $B$, $C$ and $T$ are all diagonal in fixed bases, then $A$ and $B$ are probability vectors, $T$ is a nonnegative matrix and the partial-trace constraints reduce to the usual row and column sum constraints. Hence classical Kantorovich OT is the diagonal, commutative subcase of (11.10). The genuinely quantum feature is that $T$ may contain off-diagonal coherences and entanglement.

**Proposition 11.15** (Quantum Kantorovich duality)**.** *For $A \in \mathbb{H}_n^{+,1}$ and $B \in \mathbb{H}_m^{+,1}$, the dual of (11.10) is*

$$\mathrm{QOT}_C(A,B) = \max_{F \in \mathbb{H}_n,\ G \in \mathbb{H}_m} \{\operatorname{tr}(FA) + \operatorname{tr}(GB) :\ F \otimes \mathrm{Id}_m + \mathrm{Id}_n \otimes G \preceq C\}. \tag{11.11}$$

*If $A$ and $B$ are positive definite, strong duality follows directly from Slater's condition; the semidefinite case follows by restriction to the supports of $A$ and $B$ or by approximation.*

*Proof.* Introduce Hermitian Lagrange multipliers $F$ and $G$ for the two marginal constraints. The Lagrangian is

$$\operatorname{tr}(CT) + \operatorname{tr}(F(A - \operatorname{Tr}_B T)) + \operatorname{tr}(G(B - \operatorname{Tr}_A T)) = \operatorname{tr}(FA) + \operatorname{tr}(GB) + \operatorname{tr}((C - F \otimes \mathrm{Id}_m - \mathrm{Id}_n \otimes G)T),$$

where (11.9) 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. The general finite-dimensional semidefinite case is obtained by approximation $A_\delta = (1-\delta)A + \delta \mathrm{Id}_n/n$, $B_\delta = (1-\delta)B + \delta \mathrm{Id}_m/m$ and by compactness, or equivalently by reducing to the supports of $A$ and $B$. □

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 $\mathrm{tr}(A) = \mathrm{tr}(B) = 1$.

**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 11.16** (von Neumann quantum entropy)**.** For a density matrix or positive semidefinite matrix $T$, the shifted von Neumann entropy functional used here is

$$H(T) = \mathrm{tr}\big(T(\log T - \mathrm{Id})\big), \qquad \nabla H(T) = \log T,$$

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 $-\mathrm{tr}(T \log T)$ by a sign and an additive constant.

For $\varepsilon > 0$ define

$$\mathrm{QOT}_C^\varepsilon(A, B) = \min_{T \succeq 0} \{\mathrm{tr}(CT) + \varepsilon H(T) : \mathrm{Tr}_B(T) = A, \mathrm{Tr}_A(T) = B\}. \tag{11.12}$$

This is the non-commutative analogue of entropic OT [184, 55]: the Shannon entropy of a coupling is replaced by the trace entropy of a density matrix.

**Proposition 11.17** (Entropic quantum OT duality)**.** *Assume $A \succ 0$, $B \succ 0$ and $\varepsilon > 0$. Then (11.12) has a unique positive minimizer. Its dual is*

$$\mathrm{QOT}_C^\varepsilon(A, B) = \max_{F \in \mathbb{H}_n,\, G \in \mathbb{H}_m} \left\{ \mathrm{tr}(FA) + \mathrm{tr}(GB) - \varepsilon\ \mathrm{tr}\exp\left(\frac{F \otimes \mathrm{Id}_m + \mathrm{Id}_n \otimes G - C}{\varepsilon}\right)\right\}. \tag{11.13}$$

*At optimality, primal and dual variables are linked by the Gibbs formula*

$$T_e(F, G) = \exp\left(\frac{F \otimes \mathrm{Id}_m + \mathrm{Id}_n \otimes G - C}{\varepsilon}\right), \tag{11.14}$$

*with $\mathrm{Tr}_B(T_e) = A$ and $\mathrm{Tr}_A(T_e) = B$.*

*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. Slater's condition justifies the Lagrange dual computation. The Fenchel identity

$$\sup_{T \succeq 0}\ \mathrm{tr}(YT) - \varepsilon H(T) = \varepsilon\ \mathrm{tr}\exp(Y/\varepsilon)$$

is the matrix analogue of the scalar exponential conjugacy. Applying it to the Lagrangian of (11.12), with $Y = F \otimes \mathrm{Id}_m + \mathrm{Id}_n \otimes G - C$, gives (11.13). The stationarity condition of this Fenchel identity gives (11.14); differentiating the dual objective with respect to $F$ and $G$ yields the two marginal equations. □

Writing $K = \exp(-C/\varepsilon)$, the objective in (11.12) differs by a constant from $\varepsilon$ times the quantum KL divergence

$$D_H(T|K) = \mathrm{tr}\big(T(\log T - \log K) - T + K\big).$$

The exact quantum analogue of Sinkhorn is an implicit alternating Bregman projection scheme onto the affine marginal sets

$$\mathcal{M}_A = \{T \succeq 0 : \mathrm{Tr}_B(T) = A\}, \qquad \mathcal{M}_B = \{T \succeq 0 : \mathrm{Tr}_A(T) = B\}.$$

**Proposition 11.18** (Exact Bregman projections)**.** *Assume $A, B > 0$ and let $K = \exp(-C/\varepsilon)$. The minimizer of (11.12) 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 (11.13).*

*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 $\Lambda$. 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 (11.13) 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 11.2 records this exact but implicit non-commutative analogue of Sinkhorn.

**Algorithm 11.2** Exact quantum Bregman projections

**Input:** Density matrices $A, B$, cost $C$, regularization $\varepsilon > 0$, tolerance tol.
**Output:** Quantum entropic coupling $T$ with partial traces $A$ and $B$.
**Initialize:** Set Hermitian potentials $F^{(0)} = 0$ and $G^{(0)} = 0$.
**For** $k = 0, 1, \ldots$ **do**:
  $T^{(k)} = T_e(F^{(k)}, G^{(k)}) = \exp\left(\frac{F^{(k)} \otimes \mathrm{Id}_m + \mathrm{Id}_n \otimes G^{(k)} - C}{\varepsilon}\right)$.
  **Solve** $A$-projection equation: $\operatorname{Tr}_B T_e(F^+, G^{(k)}) = A$,
  **Set** $F^{(k+1)} = F^+$.
  **Solve** $B$-projection equation: $\operatorname{Tr}_A T_e(F^{(k+1)}, G^+) = B$,
  **Set** $G^{(k+1)} = G^+$.
  **If** both partial-trace residuals are at most tol **then**:
    **Return** $T_e(F^{(k+1)}, G^{(k+1)})$.

**Gurvits scaling and quantum Sinkhorn.** The algorithm often called quantum Sinkhorn comes from the operator-scaling literature of Gurvits and subsequent developments [115, 116, 105, 101]. It replaces the true Gibbs coupling (11.14) by the symmetric factorization

$$T_s(F, G) = \exp\left(\frac{Z}{2\varepsilon}\right) \exp(-C/\varepsilon) \exp\left(\frac{Z}{2\varepsilon}\right) = (U \otimes V)K(U \otimes V), \qquad Z = F \otimes \mathrm{Id}_m + \mathrm{Id}_n \otimes G, \tag{11.15}$$

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

$$\begin{aligned} R_V &= \mathcal{K}(V^2), \quad & U &\leftarrow R_V^{-1/2}\big(R_V^{1/2} A R_V^{1/2}\big)^{1/2} R_V^{-1/2}, \\ S_U &= \mathcal{K}^\star(U^2), \quad & V &\leftarrow S_U^{-1/2}\big(S_U^{1/2} B S_U^{1/2}\big)^{1/2} S_U^{-1/2}. \end{aligned} \tag{11.16}$$

These inverse square roots are well-defined when $K \succ 0$ and $U, V, A, B \succ 0$. This is Gurvits/operator

**Algorithm 11.3** Gurvits/operator scaling for quantum Sinkhorn

**Input:** Positive marginals $A, B$, positive kernel operator $K$, maps $\mathcal{K}, \mathcal{K}^\star$, tolerance tol.
**Output:** Symmetrically scaled coupling $T_s$.
**Initialize:** Set $U = \mathrm{Id}_n$ and $V = \mathrm{Id}_m$.
**Set** residual $r = +\infty$.
**While** $r >$ tol **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 $B$.
**Return** $T_s$.

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 11.19 (Gurvits scaling is not the exact Bregman scheme).** It is important not to identify (11.16) with the exact Bregman scheme for (11.12). 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 (11.13). 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 11.20 (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 (11.9) 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.

CHAPTER 12

# 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. These ideas prepare the gradient-flow and generative-model chapters that follow.

## 12.1 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

$$\frac{\mathrm{d}x(t)}{\mathrm{d}t} = v_t(x(t)), \tag{12.1}$$

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 (12.1).

In the Eulerian description, the same motion is written directly on the evolving measure: the particle ODE becomes the PDE

$$\frac{\partial \alpha_t}{\partial t} + \operatorname{div}(v_t \alpha_t) = 0. \tag{12.2}$$

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 $\alpha_t$ has a smooth density. For general measures, and in particular for empirical measures, it is understood in the weak sense: for any smooth test function $(t, x) \mapsto \varphi(t, x)$ compactly supported in time,

$$\int_0^1 \int_{\mathbb{R}^d} \left(\partial_t \varphi(t, x) + \langle v_t(x),\, \nabla_x \varphi(t, x)\rangle\right) \mathrm{d}\alpha_t(x)\mathrm{d}t = 0. \tag{12.3}$$

This equation is obtained from (12.2) 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 (12.1).

**Proposition 12.1** (Lagrangian flows solve the continuity equation)**.** *Consider a smooth flow $T_t : \mathbb{R}^d \to \mathbb{R}^d$ and define $\alpha_t = (T_t)_\sharp \alpha_0$. Define the Eulerian velocity field by*

$$v_t(T_t(y)) = \partial_t T_t(y).$$

*Then $(\alpha_t, v_t)$ solves the continuity equation in the weak sense (12.3). 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 $\varphi(t, x)$ be a smooth test function vanishing at $t = 0$ and $t = 1$. 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 $\varphi$ gives (12.3). □

**From measure evolutions to vector fields.** For a given evolution $(\alpha_t)_t$, there are typically infinitely many velocity fields $v_t$ satisfying

$$\partial_t \alpha_t + \operatorname{div}(\alpha_t v_t) = 0. \tag{12.4}$$

This non-uniqueness comes from the kernel of the weighted divergence. The linear space of vector fields that leave a measure $\alpha$ invariant is

$$\mathcal{H}_\alpha = \{v \;;\; \operatorname{div}(\alpha v) = 0\}\,.$$

It is usually non-trivial: for instance, if $\alpha$ 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. A simple choice, introduced by Dacorogna and Moser [74], is to impose that the flux $\alpha_t v_t$ is a gradient field, leading formally to

$$v_t = -\frac{1}{\alpha_t} \nabla \Delta^{-1}(\partial_t \alpha_t), \tag{12.5}$$

with suitable boundary conditions, for instance vanishing at infinity. This formula is useful conceptually but delicate when $\alpha_t$ vanishes, and it does not generally produce a gradient velocity field.

**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:

$$\min_v \frac{1}{2} \int_0^1 \int_{\mathbb{R}^d} \|v_t(x)\|^2 \,\mathrm{d}\alpha_t(x)\mathrm{d}t \quad \text{subject to} \quad \partial_t \alpha_t + \operatorname{div}(\alpha_t v_t) = 0. \tag{12.6}$$

**Proposition 12.2** (Least-square velocities are gradients)**.** *Assume that $\alpha_t = \rho_t \,\mathrm{d}x$ is a smooth positive density curve and that boundary terms vanish. The minimizer of (12.6), if it exists, is a gradient field*

$$v_t = \nabla \varphi_t,$$

*where $\varphi_t$, unique up to an additive constant, solves the weighted Poisson equation*

$$-\operatorname{div}(\rho_t \nabla \varphi_t) = \partial_t \rho_t, \qquad v_t = -\nabla \Delta_{\alpha_t}^{-1}(\partial_t \alpha_t), \quad \Delta_{\alpha_t} \varphi = \operatorname{div}(\alpha_t \nabla \varphi). \tag{12.7}$$

*Proof.* Introduce a Lagrange multiplier $\varphi_t$ 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 (12.7). 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.

**Algorithm 12.1** Least-square velocity reconstruction

**Input:** Smooth positive density curve $(\rho_t)_{t\in[0,1]}$ and boundary conditions.
**Output:** Minimal-energy velocity field $v_t$ realizing the curve.
**For** each time $t$ **do**:
  **Compute** $\partial_t \rho_t$.
  **Solve** weighted Poisson equation: $-\operatorname{div}(\rho_t \nabla \varphi_t) = \partial_t \rho_t, \qquad \int \varphi_t \rho_t = 0.$
  **Set** $v_t = \nabla \varphi_t$.

**Return** $(v_t)_{t\in[0,1]}$.

## 12.2 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.

Instead of assuming that a whole curve $(\alpha_t)_{t\in[0,1]}$ is prescribed, one only fixes its endpoints $\alpha_0$ and $\alpha_1$ and minimizes the least-square energy (12.6). The theorem of Benamou and Brenier states that this geodesic energy is exactly the squared Wasserstein distance [18].

**Theorem 12.3** (Benamou–Brenier). *For probability measures $\alpha_0, \alpha_1 \in \mathcal{P}_2(\mathbb{R}^d)$,*

$$\mathcal{W}_2^2(\alpha_0, \alpha_1) = \inf_{(\alpha_t, v_t)} \int_0^1 \int_{\mathbb{R}^d} \|v_t(x)\|^2 \, \mathrm{d}\alpha_t(x)\mathrm{d}t, \tag{12.8}$$

*where the infimum is over $(\alpha_t, v_t)$ 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 $\alpha_0$ has a density and $T$ is the optimal Monge map $T_\sharp \alpha_0 = \alpha_1$, the minimizer is*

$$\alpha_t = ((1-t)\mathrm{Id} + tT)_\sharp \alpha_0, \qquad v_t\big((1-t)x + tT(x)\big) = T(x) - x. \tag{12.9}$$

*Proof.* For the inequality "dynamic $\leqslant$ static", assume first that a Monge map $T$ exists and define $(\alpha_t, v_t)$ by (12.9). 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 $\pi$: 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 $\alpha_0$, 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. □

Although (12.8) is not jointly convex in $(\alpha_t, v_t)$, it becomes convex after replacing velocities by the momentum measure $m_t = v_t \alpha_t$ and using the perspective action. In the absolutely continuous case $\alpha_t = \rho_t \, \mathrm{d}x$ and $m_t(x) = \rho_t(x) v_t(x)$, this reads

$$\mathcal{W}_2^2(\alpha_0, \alpha_1) = \inf_{\substack{\partial_t \rho_t + \operatorname{div} m_t = 0 \\ \rho_{t=0}\mathrm{d}x = \alpha_0, \ \rho_{t=1}\mathrm{d}x = \alpha_1}} \int_0^1 \int_{\mathbb{R}^d} \frac{\|m_t(x)\|^2}{\rho_t(x)} \, \mathrm{d}x \, \mathrm{d}t, \tag{12.10}$$

with the usual convention that the integrand is 0 when $(\rho_t, m_t) = (0, 0)$ and $+\infty$ when $\rho_t = 0$ but $m_t \neq 0$. For singular endpoints or curves, the same statement is interpreted with vector-valued momentum measures and the corresponding recession convention. This convex reformulation enables geodesic interpolation by convex optimization once the domain is discretized.

**Proximal splitting.** The convex momentum formulation also explains the original Benamou–Brenier solver. Papadakis, Peyré and Oudet [178] showed that, after discretization, the ALG2 scheme is a Douglas–Rachford splitting, equivalently ADMM on the Fenchel–Rockafellar dual. Suppressing discretization indices, write $U = (\rho, m)$ and introduce the perspective integrand

$$J(\rho, m) = \begin{cases} \|m\|^2/\rho, & \rho > 0, \\ 0, & (\rho, m) = (0, 0), \\ +\infty, & \text{otherwise.} \end{cases}$$

Let

$$\mathcal{F}(U) = \int_0^1 \int_{\mathbb{R}^d} J(\rho_t(x), m_t(x))\, \mathrm{d}x\, \mathrm{d}t, \qquad C = \{(\rho, m)\ ;\ \partial_t \rho + \operatorname{div} m = 0,\ \rho_0 \mathrm{d}x = \alpha_0,\ \rho_1 \mathrm{d}x = \alpha_1\},$$

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

$$\operatorname{prox}_{\tau\mathcal{F}}(\bar{U}) = \operatorname{argmin}_U \frac{1}{2}\|U - \bar{U}\|_{\mathrm{H}}^2 + \tau\mathcal{F}(U), \qquad \operatorname{prox}_{\tau\mathcal{G}}(\bar{U}) = \operatorname{argmin}_{U \in C} \frac{1}{2}\|U - \bar{U}\|_{\mathrm{H}}^2.$$

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:

$$\begin{aligned} U^{k+1} &= \operatorname{prox}_{\tau\mathcal{F}}(Z^k), \\ \widetilde{U}^{k+1} &= \operatorname{prox}_{\tau\mathcal{G}}(2U^{k+1} - Z^k), \\ Z^{k+1} &= Z^k + \widetilde{U}^{k+1} - U^{k+1}. \end{aligned}$$

Equivalently one may swap the roles of $\mathcal{F}$ and $\mathcal{G}$. At convergence, the two shadow sequences $U^k$ and $\widetilde{U}^k$ 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 [178, 86].

**Algorithm 12.2** Douglas–Rachford for dynamic Benamou–Brenier

**Input:** Functionals $\mathcal{F}, \mathcal{G} = \iota_C$, proximal parameter $\tau > 0$, initial field $Z^0$.
**Output:** Discrete density-momentum field $U^\star$.
**For** $k = 0, 1, \dots$ **do**:
  $U^{k+1} = \operatorname{prox}_{\tau\mathcal{F}}(Z^k)$.
  **Project** reflected point: $\widetilde{U}^{k+1} = \operatorname{prox}_{\tau\mathcal{G}}(2U^{k+1} - Z^k) = \operatorname{Proj}_C(2U^{k+1} - Z^k)$.
  **Update** $Z^{k+1} = Z^k + \widetilde{U}^{k+1} - U^{k+1}$.
  **If** $\|U^{k+1} - \widetilde{U}^{k+1}\| \leqslant \mathrm{tol}$ **then**:
    **Return** $U^{k+1}$.

density sequence midpoint velocities

*Figure 12.1:* 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 $Y$.

**Remark 12.4 (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

$$P_t : \mathcal{S} \to \mathbb{R}^d, \qquad P_t(\gamma) = \gamma(t).$$

The Benamou–Brenier cost admits the equivalent formulation

$$\mathcal{W}_2^2(\alpha_0,\alpha_1) = \inf_{M \in \mathcal{P}(\mathcal{S})} \left\{ \int_{\mathcal{S}} \int_0^1 \|\dot\gamma(t)\|^2 \mathrm{d}t \, \mathrm{d}M(\gamma) \; ; \; (P_0)_\sharp M = \alpha_0, \; (P_1)_\sharp M = \alpha_1 \right\}.$$

If $\alpha_0$ has a density, the minimizer $M^*$ is unique. Its time marginals reproduce the optimal curve: $\alpha_t = (P_t)_\sharp M^*$ for all $t$. Furthermore, for a.e. $t$, denoting $Q_t(\gamma) = \dot\gamma(t)$ on absolutely continuous paths, the conditional law of the velocity is deterministic:

$$(P_t, Q_t)_\sharp M^*(\mathrm{d}x, \mathrm{d}q) = \alpha_t(\mathrm{d}x)\delta_{v_t^*(x)}(\mathrm{d}q),$$

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 $\alpha_t$-a.e. spatial point.

**Extensions of the dynamic formulation.** The same variational grammar extends beyond the quadratic Wasserstein distance. One changes either the kinetic exponent, the mobility or the balance equation, while keeping a continuity-type constraint and a convex perspective action.

**Remark 12.5 (Generalized Benamou–Brenier distances).** The dynamic formulation is not specific to $\mathcal{W}_2$. For measures with finite $p$-th moments and $p > 1$, one has the analogous action formula

$$\mathcal{W}_p^p(\alpha_0,\alpha_1) = \inf_{\substack{\partial_t \alpha_t + \nabla\cdot(\alpha_t v_t) = 0 \\ \alpha_{t=0}=\alpha_0,\ \alpha_{t=1}=\alpha_1}} \int_0^1 \int_{\mathbb{R}^d} |v_t(x)|^p \, \mathrm{d}\alpha_t(x)\, \mathrm{d}t.$$

When $\alpha_t = \rho_t \, \mathrm{d}x$ and $m_t = \rho_t v_t$, this becomes the convex perspective action

$$\int_0^1 \int_{\mathbb{R}^d} \frac{|m_t(x)|^p}{\rho_t(x)^{p-1}} \, \mathrm{d}x \, \mathrm{d}t, \qquad \partial_t \rho_t + \nabla\cdot m_t = 0,$$

with the usual convention that the integrand is 0 if $(\rho, m) = (0,0)$ and $+\infty$ if $\rho = 0$ but $m \neq 0$.

A second class of variants changes the mobility of the medium: the quadratic action $|m|^2/\rho$ is replaced by $|m|^2/\theta(\rho)$ for a suitable concave mobility $\theta$. Under appropriate structural assumptions, this produces transport metrics adapted to nonlinear diffusions and finite-volume discretizations [80]. On finite graphs and Markov chains, the analogous action uses an edge mobility, often the logarithmic mean of the endpoint densities, and leads to discrete Wasserstein geometries [154, 165]. These extensions keep the same variational grammar as Benamou–Brenier: a continuity-type constraint, an action density, and geodesics obtained by minimizing an integrated kinetic cost.

**Dynamic unbalanced OT.** 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 the action penalizes both spatial motion and growth. This dynamic formulation underlies the Hellinger–

Kantorovich and Wasserstein–Fisher–Rao metrics [145, 66]; its equivalence with static entropy-transport and cone formulations is developed in [146, 65]. A representative quadratic action 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 same perspective convention as above. Equivalently, writing $m_t = \rho_t v_t$ and $s_t = \rho_t g_t$, one minimizes $\int_0^1 \int (|v_t|^2 + \kappa^2 g_t^2) \mathrm{d}\rho_t \mathrm{d}t$ under $\partial_t \rho_t + \nabla\cdot(\rho_t v_t) = g_t \rho_t$. The parameter $\kappa$ fixes the relative cost of reaction and transport; changing it rescales the radial/angular balance in the associated cone metric. For measure-valued triples, the action is understood in the lower-semicontinuous perspective sense

$$\mathcal{A}_\kappa(\rho, m, s) := \int \left( \frac{\|\dot m\|^2}{\dot\rho} + \kappa^2 \frac{\dot s^2}{\dot\rho} \right) \mathrm{d}\lambda, \qquad (\dot\rho, \dot m, \dot s) = \left( \frac{\mathrm{d}\rho}{\mathrm{d}\lambda}, \frac{\mathrm{d}m}{\mathrm{d}\lambda}, \frac{\mathrm{d}s}{\mathrm{d}\lambda} \right),$$

where $\lambda$ dominates $\rho$ and the total variations of $m$ and $s$, and the value is independent of this choice. The convention is $0/0 = 0$ and $a/0 = +\infty$ for $a > 0$, so finite action forces both the flux and source to be absolutely continuous with respect to the transported mass.

**Proposition 12.6** (Static/dynamic equivalence for unbalanced OT)**.** *Fix the action above and let* $\mathrm{CW}_\kappa$ *be the cone value of Theorem 9.4 with the cone metric normalized to the same growth scale* $\kappa$*. For nonnegative finite measures* $\alpha_0, \alpha_1$ *on* $\mathbb{R}^d$*, the dynamic value*

$$\mathrm{WFR}_\kappa^2(\alpha_0, \alpha_1) := \inf_{\substack{\partial_t \rho_t + \nabla\cdot m_t = s_t \\ \rho_0 = \alpha_0,\ \rho_1 = \alpha_1}} \int_0^1 \mathcal{A}_\kappa(\rho_t, m_t, s_t)\, \mathrm{d}t \tag{12.11}$$

*equals the static cone formulation* $\mathrm{CW}_\kappa(\alpha_0, \alpha_1)$*. Hence the static unbalanced problem and the balance-equation least-action problem define the same geodesic distance.*

*Proof.* The cone construction turns variation of mass into radial motion and spatial transport into angular motion on $\mathfrak{C}[\mathbb{R}^d]$. Applying the Benamou–Brenier theorem on the cone to the lifted endpoint measures gives a dynamic least-action problem on $\mathfrak{C}[\mathbb{R}^d]$ whose static value is the cone value $\mathrm{CW}_\kappa$ of Theorem 9.4. This is the standard static/dynamic identification for the Hellinger–Kantorovich and Wasserstein–Fisher–Rao metrics [145, 146, 66, 65].

Projecting a cone curve back to the base space with the weight $r^2$ produces a measure curve $\rho_t$, a spatial flux $m_t$ and a source term $s_t$ satisfying the balance equation. With the matching normalization of the cone metric, the cone kinetic energy decomposes exactly into the perspective action $\mathcal{A}_\kappa$ in (12.11). Conversely, any finite-action triple $(\rho_t, m_t, s_t)$ can be lifted to a cone curve whose radial velocity realizes the growth term and whose spatial velocity realizes the transport term, with the same action after relaxation. The two infima are therefore equal; lower semicontinuity gives the general finite-measure statement from the smooth positive case. □

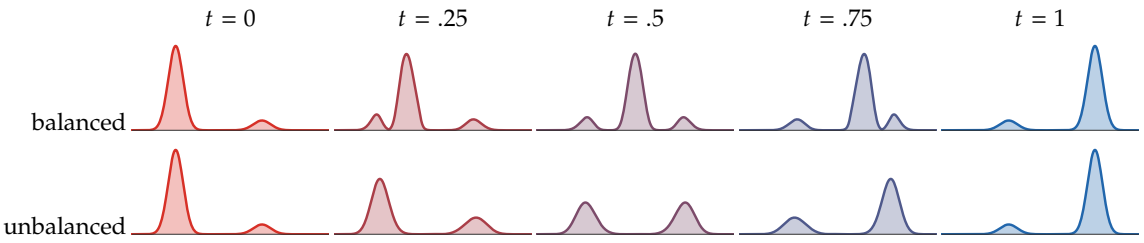



*Figure 12.2:* Balanced and unbalanced Sinkhorn-barycenter interpolations between two one-dimensional Gaussian mixtures with swapped modal masses. The balanced row conserves total mass, so excess mass from the dominant left mode must move along the line toward the dominant right target mode, producing 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.

CHAPTER 13

# Wasserstein Gradient Flows

Once $\mathcal{W}_2$ is a dynamic metric, one can run gradient descent directly on the space of measures. This chapter derives the formal Wasserstein gradient, explains the JKO minimizing-movement scheme, records the role of geodesic convexity in convergence, and then applies the same calculus to mean-field neural-network training.

## 13.1 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 of minimizing movement over a metric space is to construct a discrete-time evolution using an implicit Euler scheme:

$$\alpha_{t+\tau} := \arg\min_{\alpha} \frac{1}{2\tau}\mathcal{W}_2(\alpha_t, \alpha)^2 + f(\alpha). \tag{13.1}$$

**Algorithm 13.1** JKO minimizing movement

**Input:** Energy $f$, initial measure $\alpha^0$, time step $\tau > 0$, number of steps $K$.
**Output:** Discrete gradient-flow trajectory $(\alpha^k)_{k=0}^K$.
**For** $k = 0, \dots, K-1$ **do**:
  $\alpha^{k+1} \in \operatorname{argmin}_{\alpha \in \mathcal{P}_2(\mathbb{R}^d)} \frac{1}{2\tau}\mathcal{W}_2^2(\alpha^k, \alpha) + f(\alpha)$.
  **Set** $\alpha_t^\tau = \alpha^k$ for $t \in [k\tau, (k+1)\tau)$.
**Return** $(\alpha^k)_{k=0}^K$ and $\alpha_t^\tau$.

**Euclidean gradient flows.** If we restrict (13.1) to finite dimensions and assume $\alpha_t = \delta_{x(t)}$ and $\alpha = \delta_x$ (single Dirac measures), this matches the implicit Euler scheme:

$$x(t+\tau) := \arg\min_x \frac{1}{2\tau}\|x - x(t)\|^2 + h(x),$$

where $h(x) = f(\delta_x)$. Its solution is formally given 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:

$$\dot{x}(t) = -\nabla h(x(t)). \tag{13.2}$$

**Wasserstein gradient formula.** The implicit Euler scheme has the advantage that it does not require $h$ or $f$ to be smooth. For $f$, this is crucial to handle evolutions over measures that may have densities, atoms or other singular parts.

As $\tau \to 0$, under certain conditions on $f$, (13.1) defines a continuous evolution $t \mapsto \alpha_t$. As discussed earlier, this evolution can be described as a Lagrangian evolution (12.1). We use the following first-variation convention: for any $\beta \in \mathcal{P}(\mathbb{R}^d)$ and the signed zero-mass perturbation $\rho = \beta - \alpha$,

$$f((1-\tau)\alpha + \tau\beta) = f(\alpha + \tau\rho) = f(\alpha) + \tau \int [\delta f(\alpha)](x)\mathrm{d}\rho(x) + o(\tau).$$

The key infinitesimal object is the vector field that represents this differential in the Wasserstein metric.

**Definition 13.1** (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 $\alpha$ is the gradient vector field

$$\nabla_{\mathcal{W}} f(\alpha) = \nabla_x \delta f(\alpha).$$

The associated formal gradient flow is the continuity equation

$$\frac{\partial \alpha_t}{\partial t} + \operatorname{div}(-\nabla_{\mathcal{W}} f(\alpha_t)\alpha_t) = 0. \tag{13.3}$$

The following proposition explains why this vector field is the Riemannian gradient for the $L^2(\alpha)$ metric on velocities.

**Proposition 13.2** (Formal Wasserstein gradient)**.** *Assume that $f$ admits a smooth first variation $\delta f(\alpha)$ and that $\alpha$ has a smooth positive density. For infinitesimal perturbations generated by a velocity field $v$ through $(\mathrm{Id} + \tau v)_\sharp \alpha$, 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*

$$\nabla_{\mathcal{W}} f(\alpha) = \nabla \delta f(\alpha).$$

*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

$$-\int \delta f(\alpha) \operatorname{div}(\alpha v)\mathrm{d}x = \int \langle \nabla \delta f(\alpha),\, v\rangle \mathrm{d}\alpha.$$

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 [176]. Rigorous metric-space treatments and numerical JKO schemes can be found in [7, 20, 180, 98].

**From the JKO step to the velocity field.** A first-order expansion of the JKO step explains why (13.3) uses the vector field $\nabla_{\mathcal{W}} f(\alpha)$. Write (13.1) as a minimization over displacement fields $v$ such that $\alpha = (\mathrm{Id} + \tau v)_\sharp \alpha_t$:

$$\min_v \frac{1}{2\tau}\tau^2 \|v\|^2_{L^2(\alpha_t)} + f((\mathrm{Id} + \tau v)_\sharp \alpha_t).$$

Then we perform a first-order Taylor expansion of this formulation using

$$(\mathrm{Id} + \tau v)_\sharp \alpha_t = \alpha_t - \tau \operatorname{div}(v\alpha_t) + o(\tau)$$

$$f((\mathrm{Id} + \tau v)_\sharp \alpha_t) = f(\alpha_t) - \tau \int \delta f(\alpha_t) \operatorname{div}(v\alpha_t)\mathrm{d}x + o(\tau)$$

$$= f(\alpha_t) + \tau \int \langle \nabla_x \delta f(\alpha_t)(x), v(x)\rangle \mathrm{d}\alpha_t(x) + o(\tau)$$

to obtain the following first-order expansion in $\tau$ of the problem minimized in (13.1)

$$\min_v f(\alpha_t) + \tau \int \left[\frac{1}{2}\|v(x)\|^2 + \langle \nabla_{\mathcal{W}} f(\alpha_t)(x), v(x)\rangle\right] \mathrm{d}\alpha_t(x) + o(\tau).$$

The pointwise minimizer is $v = -\nabla_{\mathcal{W}} f(\alpha_t)$, which gives the velocity in the continuity equation. We now detail examples of such Wasserstein gradient flows.

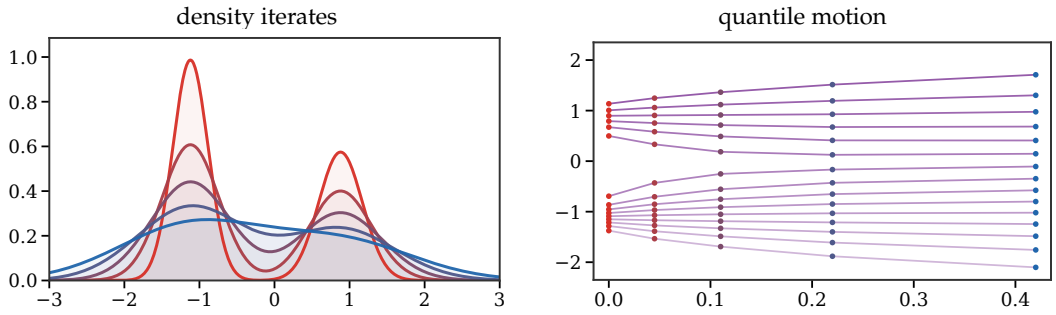



*Figure 13.1:* JKO minimizing movements for the entropy flow in one dimension. The left panel displays successive implicit-Euler minimizers for the heat equation, colored from red to blue. The right panel tracks inverse CDF values $Q_t(s) = F_t^{-1}(s)$ for selected probability levels $s$, giving a Lagrangian view of the proximal movement in Wasserstein space.

**Discrete evolutions.** If $f(\alpha)$ can be evaluated on discrete distributions and $\nabla_W$ is continuous in this case, the flow (13.3) maintains the number of Dirac masses, $\alpha_t = \frac{1}{n}\sum_i \delta_{x_i(t)}$. The particles $X(t) := (x_i(t))_i$ evolve according to a system of coupled ODEs:

$$\dot{x}_i(t) = -n\nabla_{x_i} F(X(t)), \tag{13.4}$$

where $F(X) := f\left(\frac{1}{n}\sum_i \delta_{x_i}\right)$ and the factor $n$ comes from the empirical Wasserstein metric $\frac{1}{n}\sum_i \|\dot{x}_i\|^2$.

**Algorithm 13.2** Empirical Wasserstein particle descent

**Input:** Particles $X^0 = (x_1^0, \dots, x_n^0)$, functional $f$, step size $h$, tolerance tol.
**Output:** Particle trajectory $(X^k)_k$ and empirical measures.
**Define** $F(X) = f\left(\frac{1}{n}\sum_{i=1}^n \delta_{x_i}\right)$.
**For** $k = 0, 1, \dots$ **do**:
  **For** $i = 1, \dots, n$ **do**
    $g_i^k = n\nabla_{x_i} F(X^k), \qquad x_i^{k+1} = x_i^k - h\, g_i^k.$
  **If** $\max_i \|x_i^{k+1} - x_i^k\| \leqslant$ tol **then**:
    **Return** $\frac{1}{n}\sum_i \delta_{x_i^{k+1}}$.

**Linear Functionals.** The first benchmark is a functional whose first variation does not depend on the current measure.

**Example 13.3 (Linear potentials generate independent particles).** Take a linear functional

$$f(\alpha) = \int h(x)\mathrm{d}\alpha(x). \tag{13.5}$$

Then $\delta f(\alpha) = h$ is independent of $\alpha$, and the flow (13.3) 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 (13.2).

**Shannon Neg-Entropy.** Entropy gives the opposite benchmark: the flow is no longer a deterministic push-forward of particles, but a diffusion of density.

**Example 13.4 (Entropy generates heat and porous-medium flows).** The canonical density-dependent functional is the Shannon neg-entropy

$$f(\alpha) = \int \log\left(\frac{\mathrm{d}\alpha}{\mathrm{d}x}(x)\right)\mathrm{d}\alpha(x). \tag{13.6}$$

Here, $\delta f(\alpha) = \log(\mathrm{d}\alpha/\mathrm{d}x)$ up to an additive constant, so $\nabla_W f(\alpha) = \nabla\alpha/\alpha$ (often called the score). The flow

(13.3) becomes the heat equation

$$\partial_t \alpha_t = \Delta \alpha_t.$$

Other entropy functionals lead to nonlinear diffusion equations; finite-volume and particle discretizations are discussed in [53, 107, 154, 87]. For a generalized entropy

$$f(\alpha) = \int g\left(\frac{\mathrm{d}\alpha}{\mathrm{d}x}\right)\mathrm{d}x, \tag{13.7}$$

with a scalar convex function $g$, one obtains nonlinear diffusions in the smooth-density regime:

$$\frac{\partial \alpha_t}{\partial t} = \Delta(P(\alpha_t)),$$

where the pressure $P$ satisfies $P'(s) = sg''(s)$. For example, $g(s) = s\log(s)$ gives $P(s) = s$ and recovers (13.6), 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.

The preceding examples are also governed by a precise geodesic-convexity criterion. A celebrated theorem by McCann [157] states that an internal energy of the form (13.7), for $g : \mathbb{R}^+ \to \mathbb{R} \cup \{+\infty\}$ with $g(0) = 0$, is geodesically convex on $\mathcal{P}(\mathbb{R}^d)$ when $g$ is convex and the map $r \mapsto r^d g(r^{-d})$ is convex and nonincreasing on $(0, +\infty)$. Examples of such functions are $g(s) = s^q$ for $q > 1$ and Shannon entropy $g(s) = s\log(s)$. By contrast, $g(s) = -\log(s)$, associated with the reverse KL divergence, does not satisfy this displacement-convexity criterion.

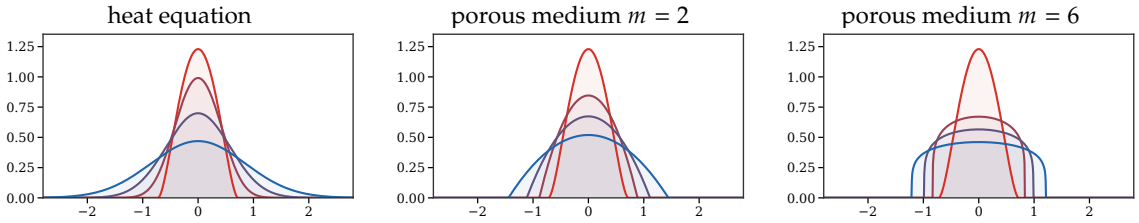



*Figure 13.2:* Entropy-driven Wasserstein gradient flows from the same compact initial density. The heat flow is generated by Shannon entropy $g(\rho) = \rho\log\rho$ and instantly develops Gaussian tails. The porous-medium flows use the power entropy $g(\rho) = \rho^m/(m-1)$, hence $\partial_t\rho = \Delta(\rho^m)$: 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.

**Interaction Energies.** In a similar spirit, to obtain nonlinear evolutions without requiring the measure to have density, one can consider

$$f(\alpha) := \iint k(x,y)\mathrm{d}\alpha(x)\mathrm{d}\alpha(y). \tag{13.8}$$

For a symmetric kernel $k$:

$$\delta f(\alpha)(x) = 2\int k(x,y)\mathrm{d}\alpha(y), \quad \nabla_W f(\alpha)(x) = 2\int \nabla_x k(x,y)\mathrm{d}\alpha(y).$$

For $\alpha_0 = \frac{1}{n}\sum_i \delta_{x_i}$, the flow (13.3) 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 $\beta$, 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 corresponding simulation loop is Algorithm 13.3.

**Algorithm 13.3** MMD particle flow against a teacher law

**Input:** Initial particles $(x_i^0)_{i=1}^n$, teacher law $\beta$ or teacher samples $(y_b)_{b=1}^B$, kernel $k$, step size $h$.
**Output:** Particle trajectory targeting $\beta$.
**For** $k = 0, 1, \ldots$ **do**:
  **For** $i = 1, \ldots, n$ **do**
    **Set** self-interaction $r_i^k = -\frac{2}{n}\sum_{j=1}^n \nabla_x k(x_i^k, x_j^k)$.
    **If** $\beta$ is available analytically **then**:
      **Set** teacher attraction $a_i^k = 2\int \nabla_x k(x_i^k, y)\mathrm{d}\beta(y)$.
    **If** only samples $(y_b)_{b=1}^B$ are available **then**:
      **Set** $a_i^k = \frac{2}{B}\sum_{b=1}^B \nabla_x k(x_i^k, y_b)$.
    **Set** velocity $v_i^k = r_i^k + a_i^k$.
    **Update** $x_i^{k+1} = x_i^k + h\, v_i^k$.
**Return** $(x_i^k)_{i,k}$.

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 $\beta$ 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 $\beta$, and small particle systems may cover the target modes poorly. Figure 13.3 illustrates this finite-particle effect.

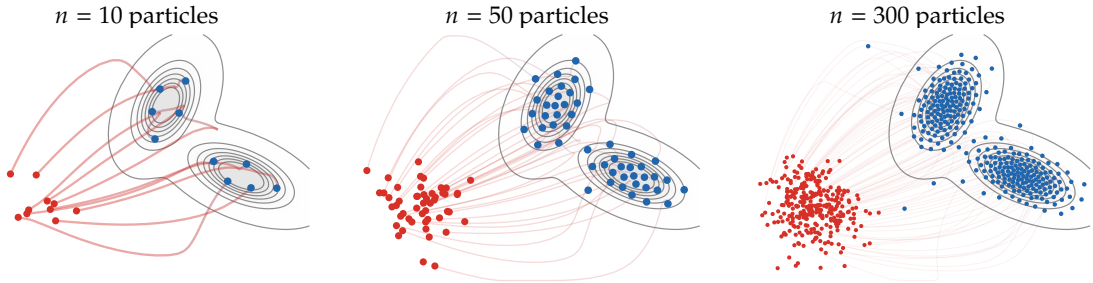



*Figure 13.3:* Particle count in the deterministic Wasserstein gradient flow of the squared MMD-type discrepancy to a smooth two-Gaussian teacher distribution, using here the energy-distance kernel $k(x,y) = -\|x-y\|$. The teacher itself is shown only through true density contours, while red dots are 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.

**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

$$\partial_t \rho_t = -\operatorname{div}(b\rho_t) + \sigma^2 \Delta \rho_t.$$

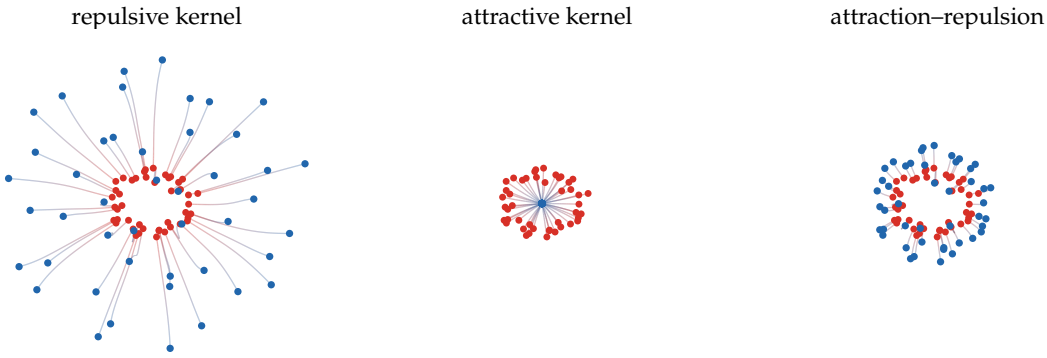



*Figure 13.4:* Interaction-energy particle flows for three choices of $k$. A positive Gaussian kernel $k(x, y) = \exp(-\|x - y\|^2/(2\sigma^2))$ produces short-range repulsion under Wasserstein descent; changing its sign produces attraction and collapse; adding a quadratic long-range attraction to the 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 (13.4).

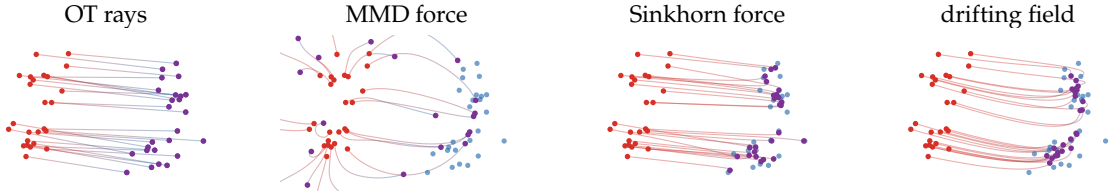



*Figure 13.5:* Particle trajectories induced by different discrepancy geometries. The red particles and blue target cloud are the same in all panels. Straight OT displacement produces rays from an optimal matching; an MMD-type witness field gives smoother nonlocal forces; the Sinkhorn-divergence force 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.

**Example 13.5 (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

$$\rho \mapsto \int V\rho \,\mathrm{d}x + \sigma^2 \int \rho \log \rho \,\mathrm{d}x.$$

The mean-field case is different: the drift is recomputed from the current empirical distribution of all particles,

$$\mathrm{d}X_i^n(t) = b(X_i^n(t), \mu_t^n)\mathrm{d}t + \sqrt{2}\sigma \mathrm{d}B_i(t), \qquad \mu_t^n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i^n(t)}.$$

For finite $n$, the empirical law $\mu_t^n$ 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 $\mu_t^n$ 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

$$\mathcal{E}(\rho) + \sigma^2 \int \rho \log \rho \,\mathrm{d}x.$$

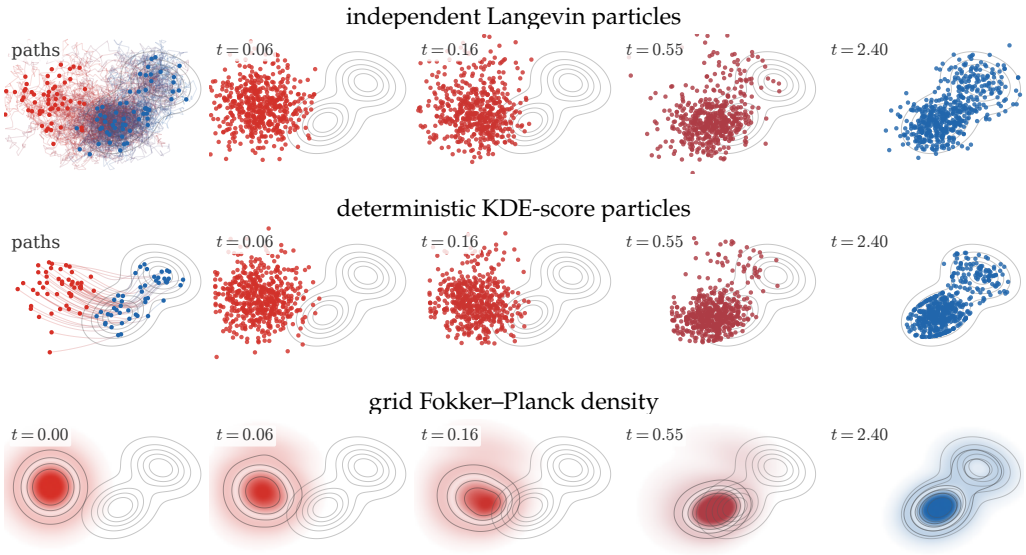



*Figure 13.6:* Three numerical representations of the same entropy-regularized Wasserstein gradient flow of $\mathrm{KL}(\rho|\beta)$, where $\beta$ is a two-Gaussian target shifted to the right of an initially isotropic Gaussian density. The first row simulates independent Langevin particles and displays a thinned set of trajectories in the left panel. The second row evolves many deterministic particles with velocity $\tau(\nabla \log \beta - \nabla \log \rho_t)$, estimating $\nabla \log \rho_t$ by a sharper kernel-density score; 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.

## 13.2 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 $\alpha_0$ and $\alpha_1$ is obtained, as in Definition 3.35, 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 13.6** (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 $\lambda$-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 13.7** (Basic geodesically convex energies)**.** *The following formal statements hold on $\mathcal{P}_2(\mathbb{R}^d)$.*

1. *If $h$ is convex, then $\alpha \mapsto \int h \mathrm{d}\alpha$ is geodesically convex; if $h$ is $\lambda$-strongly convex, it is $\lambda$-geodesically convex.*
2. *If $W(x-y)$ is convex as a function of the displacement, then $\alpha \mapsto \frac{1}{2}\iint W(x-y)\mathrm{d}\alpha(x)\mathrm{d}\alpha(y)$ is geodesically convex.*
3. *Shannon entropy $\alpha \mapsto \int \rho \log \rho \, \mathrm{d}x$ is geodesically convex.*
4. *The relative entropy $\mathrm{KL}(\alpha|\gamma)$ with $\mathrm{d}\gamma = e^{-V}\mathrm{d}x/Z$ is $\lambda$-geodesically convex when $V$ is $\lambda$-strongly convex.*

*Proof.* Along a Monge geodesic $X_t = (1-t)X_0 + tX_1$, convexity of $h$ gives $h(X_t) \leqslant (1-t)h(X_0) + t h(X_1)$, and strong convexity gives the additional quadratic term; integrating proves the first claim. The

interaction claim follows similarly by applying convexity of $W$ to pairwise differences $X_t - X'_t = (1-t)(X_0 - X'_0) + t(X_1 - X'_1)$ and integrating over two independent copies. The entropy claim is McCann's displacement convexity theorem; at the density level it follows from the concavity of the Jacobian determinant under the interpolation of optimal maps. Finally, $\mathrm{KL}(\alpha|\gamma) = \int \rho \log \rho \, \mathrm{d}x + \int V \, \mathrm{d}\alpha + \text{constant}$, so it is the sum of displacement-convex entropy and a $\lambda$-geodesically convex linear potential. □

**Convergence of the flow.** In general, analyzing (13.3) is delicate. The cleanest case is when $f$ is geodesically convex in the sense above. This condition is the Wasserstein analogue of convexity in Euclidean gradient descent.

**Proposition 13.8** (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*

$$\partial_t \alpha_t + \operatorname{div}(\alpha_t v_t) = 0, \qquad v_t = -\nabla_{\mathcal{W}} f(\alpha_t).$$

*Then*

$$\frac{\mathrm{d}}{\mathrm{d}t} f(\alpha_t) = -\int \|\nabla_{\mathcal{W}} f(\alpha_t)(x)\|^2 \, \mathrm{d}\alpha_t(x) \leqslant 0.$$

*If $T_t$ is the optimal map from $\alpha_t$ 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*

$$f(\alpha_t) - f(\alpha^\star) \leqslant \frac{\mathcal{W}_2^2(\alpha_0, \alpha^\star)}{2t}.$$

*If $f$ is $\lambda$-geodesically convex with $\lambda > 0$, then*

$$f(\alpha_t) - f(\alpha^\star) \leqslant e^{-2\lambda t} \big( f(\alpha_0) - f(\alpha^\star) \big).$$

*Proof.* The chain rule and Proposition 13.2 give

$$\frac{\mathrm{d}}{\mathrm{d}t} f(\alpha_t) = \int \langle \nabla_{\mathcal{W}} f(\alpha_t)(x), \, v_t(x) \rangle \, \mathrm{d}\alpha_t(x) = -\int \|\nabla_{\mathcal{W}} f(\alpha_t)(x)\|^2 \, \mathrm{d}\alpha_t(x).$$

Geodesic convexity along the geodesic $((1-s)\mathrm{Id} + sT_t)_\sharp \alpha_t$ gives

$$f(\alpha^\star) - f(\alpha_t) \geqslant \int \langle \nabla_{\mathcal{W}} f(\alpha_t)(x), \, T_t(x) - x \rangle \, \mathrm{d}\alpha_t(x).$$

Since $v_t = -\nabla_{\mathcal{W}} f(\alpha_t)$, 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

$$t \big( f(\alpha_t) - f(\alpha^\star) \big) \leqslant \int_0^t \big( f(\alpha_s) - f(\alpha^\star) \big) \, \mathrm{d}s \leqslant \frac{1}{2} \mathcal{W}_2^2(\alpha_0, \alpha^\star).$$

If $f$ is $\lambda$-geodesically convex, the Wasserstein analogue of strong convexity gives the slope inequality

$$\int \|\nabla_{\mathcal{W}} f(\alpha_t)\|^2 \mathrm{d}\alpha_t \geqslant 2\lambda \big( f(\alpha_t) - f(\alpha^\star) \big).$$

Combining it with the energy dissipation identity yields

$$\frac{\mathrm{d}}{\mathrm{d}t} \big( f(\alpha_t) - f(\alpha^\star) \big) \leqslant -2\lambda \big( f(\alpha_t) - f(\alpha^\star) \big),$$

and Gronwall's lemma gives the exponential rate. □

**Proposition 13.9** (Convex examples covered by the theory)**.** *The hypotheses of Proposition 13.8 are satisfied in the following standard cases, at least at the formal smooth level used in this section.*

1. *For the linear energy $f(\alpha) = \int h \mathrm{d}\alpha$, geodesic convexity holds when $h$ is convex. If $h$ is $\lambda$-strongly convex, then $f$ is $\lambda$-geodesically convex and the flow enjoys the exponential rate of Proposition 13.8.*
2. *For the interaction energy $f(\alpha) = \frac{1}{2}\iint W(x-y)\mathrm{d}\alpha(x)\mathrm{d}\alpha(y)$, geodesic convexity holds when $W$ is convex and even. This covers repulsive or attractive pairwise models whose displacement cost has no non-convex wells.*
3. *The Shannon entropy $f(\alpha) = \int \rho \log \rho \, \mathrm{d}x$ and, more generally, McCann displacement-convex internal energies generate diffusion-type Wasserstein gradient flows.*
4. *If $\gamma = Z^{-1}e^{-V}\mathrm{d}x$ and $V$ is $\lambda$-strongly convex, then the relative entropy $\mathrm{KL}(\alpha|\gamma)$ is $\lambda$-geodesically convex. Its flow is the Fokker–Planck equation with invariant law $\gamma$.*

*Proof.* Let $(\alpha_t)_t$ be the McCann interpolation between $\alpha_0$ and $\alpha_1$, written with an optimal coupling as $X_t = (1-t)X_0 + tX_1$. For a linear energy, Jensen's inequality gives

$$h(X_t) \leqslant (1-t)h(X_0) + t h(X_1),$$

and the strong convexity version gives the additional term $-\frac{\lambda}{2}t(1-t)\|X_0 - X_1\|^2$. Integrating over the optimal coupling proves geodesic convexity and $\lambda$-geodesic convexity.

For interaction energies, use two independent copies of the optimal coupling. The pairwise displacement evolves as

$$X_t - X_t' = (1-t)(X_0 - X_0') + t(X_1 - X_1').$$

Convexity of $W$ gives the convexity inequality after integration over the product coupling. Evenness of $W$ ensures that the interaction is symmetric in the two particles and matches the usual factor $1/2$ in (13.8).

The entropy claim is McCann's displacement-convexity theorem. For smooth positive densities and Brenier maps, it follows from the change-of-variables formula and the concavity of the determinant along positive matrices; the general statement is obtained by approximation. Finally,

$$\mathrm{KL}(\alpha|\gamma) = \int \rho \log \rho \, \mathrm{d}x + \int V \mathrm{d}\alpha + \log Z,$$

so it is the sum of the displacement-convex entropy and the $\lambda$-geodesically convex linear potential generated by $V$. Proposition 13.8 then applies to all four cases. □

**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 3.5. 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 $\mathrm{Ric}_g$ 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 13.10** (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 $\lambda$-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 [69, 229]; it is a central theme of the optimal-transport approach to curvature in Villani's monograph [226]. Lott–Villani and Sturm then used the same entropy-convexity principle to define synthetic lower Ricci curvature bounds on metric-measure spaces [151, 219, 220]. 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.

## 13.3 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.

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 $\sigma$, 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 $\alpha$ to be a continuous distribution of infinitely many neurons.

Let $\rho$ be a probability distribution on data-label pairs $(z, y) \in \mathbb{R}^d \times \mathbb{R}^{d'}$. The population risk is

$$f(\alpha) = \int \ell(G_\alpha(z), y)\mathrm{d}\rho(z, y),$$

and the empirical risk is the special case $\rho = \rho_N := N^{-1}\sum_{k=1}^{N} \delta_{(z_k, y_k)}$. Since $\alpha \mapsto G_\alpha$ is linear, $f$ is convex as a function of $\alpha$ 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 (13.3).

Assume that $\ell$ is differentiable in its first variable. The first variation is

$$\delta f(\alpha)(x) = \int \langle \nabla_1 \ell(G_\alpha(z), y),\, \psi(x, z)\rangle \mathrm{d}\rho(z, y), \tag{13.9}$$

and the Wasserstein gradient in parameter space is

$$\nabla_W f(\alpha)(x) = \nabla_x \delta f(\alpha)(x) = \int [D_x \psi(x, z)]^\top \nabla_1 \ell(G_\alpha(z), y)\mathrm{d}\rho(z, y).$$

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:

$$f(\alpha) = \frac{1}{2}\iint k(x, x')\mathrm{d}\alpha(x)\mathrm{d}\alpha(x') + \int g(x)\mathrm{d}\alpha(x) + \frac{1}{2}\int \|y\|^2 \mathrm{d}\rho(z, y), \tag{13.10}$$

with

$$k(x, x') = \int \langle \psi(x, z),\, \psi(x', z)\rangle \mathrm{d}\rho(z, y), \qquad g(x) = -\int \langle y,\, \psi(x, z)\rangle \mathrm{d}\rho(z, y). \tag{13.11}$$

Thus

$$\delta f(\alpha)(x) = \int k(x, x')\mathrm{d}\alpha(x') + g(x), \qquad \nabla_W f(\alpha)(x) = \int \nabla_x k(x, x')\mathrm{d}\alpha(x') + \nabla_x g(x).$$

These kernels are generally not convex in the particle variable, so the geodesic-convex convergence theory above does not apply directly.

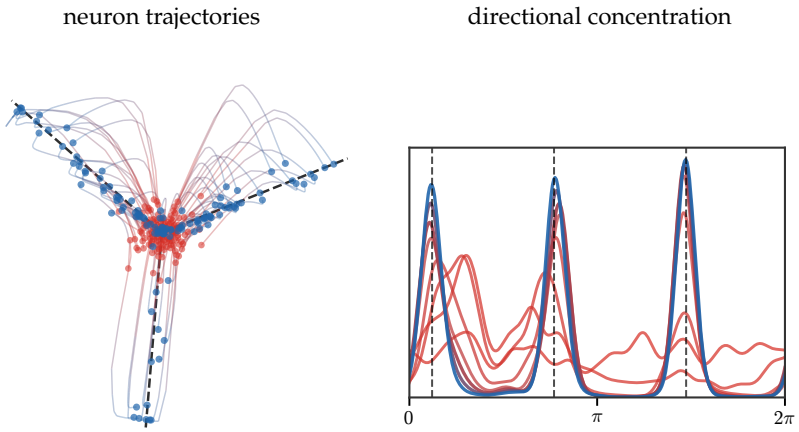



*Figure 13.7:* Mean-field training of a homogeneous two-layer model as transport in neuron space. The left panel shows the Wasserstein particle gradient flow in the reduced homogeneous coordinates $(|u|v_1, |u|v_2)$, with black dashed rays marking the teacher directions. The right panel shows the weighted 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 $W_2$ 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 13.11** (Affine convexity and stationary densities)**.** *Let $F = Q + \int V\,\mathrm{d}\alpha + C$ be as above, and assume that $Q$ is classically convex. Suppose that a Wasserstein gradient flow for $F$ converges to a measure $\alpha_\infty = \rho_\infty \mathrm{d}x$. Assume also the standard regularity needed to pass to the limit in the first variation, and assume that the support and positivity of $\rho_\infty$ allow the stationary condition to be tested against all admissible zero-mass density perturbations. In the form needed here, assume that this stationarity yields, for every competitor $\beta$, the variational inequality*

$$\int \delta F(\alpha_\infty)(x)\mathrm{d}(\beta - \alpha_\infty)(x) \geqslant 0.$$

*Then $\alpha_\infty$ is a global minimizer of $F$.*

*Proof sketch.* The dissipation identity for the gradient flow gives stationarity of the limit: formally, after passing to the limit,

$$\int \|\nabla \delta F(\alpha_\infty)\|^2 \mathrm{d}\alpha_\infty = 0.$$

Without such a support and positivity assumption, this identity only controls the first variation on the region explored by the limit. The density hypothesis allows one to test against sufficiently many signed density perturbations of total mass zero. By approximation and the assumed regularity, this yields the displayed first-order variational inequality for arbitrary competitors $\beta$. Classical convexity of $F$ in the affine variable $\alpha$ then gives the usual subgradient inequality

$$F(\beta) \geqslant F(\alpha_\infty) + \int \delta F(\alpha_\infty)\mathrm{d}(\beta - \alpha_\infty) \geqslant F(\alpha_\infty).$$

Thus no competitor has smaller energy. For square-loss two-layer mean-field models, (13.10) is exactly of this quadratic-plus-linear form, and positive semidefiniteness of the induced kernel $k$ is the classical convexity assumption. □

The mean-field description of two-layer training was developed in several works, including [63, 158]. The distinctive contribution of Chizat and Bach is a global-convergence analysis for positively homogeneous networks without adding an explicit regularizer or relying on noisy SGD to create a Laplacian term. The following formal statement isolates the core mechanism and ignores the technical issues due to ReLU non-smoothness, support propagation and compactness.

**Proposition 13.12** (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 $\alpha$ 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 nonzero direction $\omega$, the support of $\alpha$ intersects the ray $\{\lambda\omega : \lambda > 0\}$. Then $\alpha$ is a global minimizer of $f$ over the mean-field model class.*

*Proof.* Write

$$h_\alpha(x) = \delta f(\alpha)(x) = \langle \nabla J(G_\alpha), \psi(x, \cdot) \rangle_\rho .$$

By two-homogeneity of $\psi$, one has $h_\alpha(\lambda x) = \lambda^2 h_\alpha(x)$. Normalize a nonzero direction $\omega$ and choose $r_\omega > 0$ with $r_\omega \omega \in \mathrm{supp}(\alpha)$. Stationarity gives a zero radial derivative at this point:

$$0 = \frac{\mathrm{d}}{\mathrm{d}r} h_\alpha(r\omega)\Big|_{r=r_\omega} = 2 r_\omega h_\alpha(\omega).$$

Hence $h_\alpha(\omega) = 0$ for every direction $\omega$, and by homogeneity $h_\alpha(x) = 0$ for every $x$.

For any competitor $\beta$, convexity of $J$ gives

$$f(\beta) - f(\alpha) \geqslant \int h_\alpha(x) \mathrm{d}(\beta - \alpha)(x) = 0.$$

Thus no competitor has smaller risk. The rigorous theorem replaces the full directional support assumption by propagation and overparameterization hypotheses ensuring that a negative descent direction would be present in the support and would contradict stationarity. □

CHAPTER 14

# 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.

## 14.1 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 $\alpha$ (typically an isotropic Gaussian) and the target data distribution $\beta$. 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 [181]. The diffusion branch builds on score matching [124], denoising score matching [227], nonequilibrium noising chains [212], denoising diffusion probabilistic models [122], score-based generative modeling [214], and the continuous-time score-SDE/probability-flow formulation [215]. The deterministic regression branch was introduced, essentially in parallel, under three closely related names: flow matching [147], rectified flow [148], and stochastic interpolants [2]. 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 $\alpha_t$ and then finding $v_t$ using the least-squares formula (12.7). 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 $\alpha$ and $\beta$.

**Stochastic interpolant.** We assume that $\alpha_t$ is defined via a "projection" (in a loose sense) of a latent distribution $\pi \in \mathcal{P}(\mathbb{R}^{d'})$, using an operator $P_t : \mathbb{R}^{d'} \to \mathbb{R}^d$ where $d' \gg d$, i.e.

$$\forall t \in [0,1], \quad \alpha_t := (P_t)_\sharp \pi. \tag{14.1}$$

The basic two-endpoint construction already covers most flow-matching paths used in practice.

**Example 14.1 (Linear two-endpoint stochastic interpolants).** Set $d' = 2d$, write $(x, y) \in \mathbb{R}^d \times \mathbb{R}^d$, and choose $P_0(x, y) = x$ and $P_1(x, y) = y$. If $\pi$ has marginals $(\alpha_0, \alpha_1)$, then $\alpha_t = (P_t)_\sharp \pi$ interpolates between the two endpoint laws. The simplest choices are the independent coupling $\pi = \alpha_0 \otimes \alpha_1$ and the straight path

$$P_t(x, y) = (1 - t)x + ty.$$

With this linear path and an arbitrary coupling $\pi$, 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 [147, 148].

More complex constructions are possible when sampling from $\pi$ remains simple; stochastic interpolants add latent variables or noise, connecting deterministic flows, probability-flow ODEs and diffusion SDEs [2].

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 $\alpha_t$ consists of $n \times m$ Dirac masses

$$\alpha_t = \frac{1}{nm}\sum_{i,j} \delta_{P_t(x_i,y_j)}.$$

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.

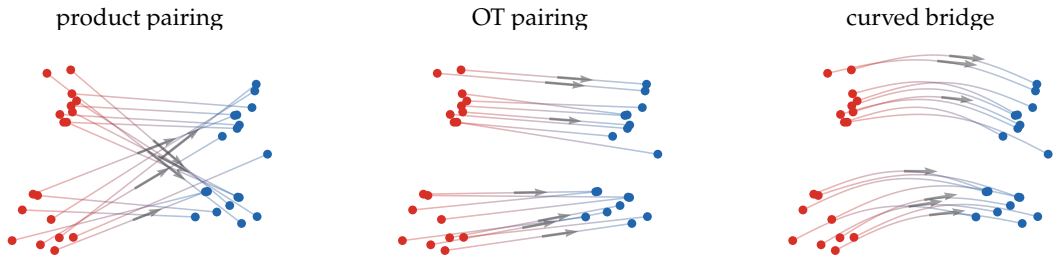



*Figure 14.1:* Flow matching interpolants between the same empirical source and target measures. A product-style random pairing produces crossing paths, an OT pairing gives direct displacement rays, and a curved bridge changes the path geometry while keeping the same endpoints. Gray arrows mark representative midpoint velocities $\partial_t P_t$.

**Flow matching formula.** This interpolation is not directly useful for sampling from $\beta$, but it can be used to define a flow field $v_t$ so that the Eulerian advection equation (12.2) holds. This flow field is computed by solving an unconstrained least-squares problem, or equivalently, it is a conditional expectation.

**Proposition 14.2** (Flow matching vector field)**.** *For each fixed $t$, assume $\partial_t P_t \in L^2(\pi; \mathbb{R}^d)$. The solution of the flow-matching problem over measurable fields $v_t : \mathbb{R}^d \to \mathbb{R}^d$*

$$\min_{v_t} \int_{\mathbb{R}^{d'}} \|v_t(P_t(u)) - [\partial_t P_t](u)\|^2 \, \mathrm{d}\pi(u). \tag{14.2}$$

*Equivalently, the minimizer is characterized $\alpha_t$-almost everywhere by the conditional expectation*

$$v_t(z) = \mathbb{E}_{u\sim\pi}\big([\partial_t P_t](u) \,\big|\, z = P_t(u)\big). \tag{14.3}$$

*Then the pair $(\alpha_t, v_t)$ satisfies the continuity equation (12.2).*

*Proof.* We first recall the two equivalent ways of writing the interpolated measure. Formally, one may write

$$\alpha_t(z) = \int_{\mathbb{R}^{d'}} \delta(z - P_t(u)) \, \mathrm{d}\pi(u),$$

while the rigorous meaning is that, for every smooth test function $\varphi$,

$$\int_{\mathbb{R}^d} \varphi(z) \, \mathrm{d}\alpha_t(z) = \int_{\mathbb{R}^{d'}} \varphi(P_t(u)) \, \mathrm{d}\pi(u). \tag{14.4}$$

The minimizer in (14.2) is the orthogonal projection in $L^2(\pi; \mathbb{R}^d)$ of the latent velocity $\partial_t P_t(u)$ onto the closed subspace of functions that depend on $u$ only through $P_t(u)$. This projection is the conditional expectation (14.3). Formally, this can be read as

$$v_t(z) = \frac{1}{\alpha_t(z)} \int_{\mathbb{R}^{d'}} \delta(z - P_t(u))[\partial_t P_t](u) \, \mathrm{d}\pi(u),$$

and rigorously it means that, for every smooth test vector field $m$,

$$\int \langle m(z), \, v_t(z)\rangle \, \mathrm{d}\alpha_t(z) = \int \langle m(P_t(u)), \, [\partial_t P_t](u)\rangle \, \mathrm{d}\pi(u). \tag{14.5}$$

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 $\varphi$,

$$\frac{\mathrm{d}}{\mathrm{d}t} \int \varphi(z)\, \mathrm{d}\alpha_t(z) - \int \langle v_t(z),\, \nabla\varphi(z)\rangle\, \mathrm{d}\alpha_t(z) = 0. \tag{14.6}$$

Using (14.4) and differentiating under the integral sign gives

$$\frac{\mathrm{d}}{\mathrm{d}t} \int \varphi(z)\mathrm{d}\alpha_t(z) = \int \langle \nabla\varphi(P_t(u)),\, [\partial_t P_t](u)\rangle \mathrm{d}\pi(u). \tag{14.7}$$

On the other hand, applying (14.5) with $m = \nabla\varphi$ gives

$$\int \langle v_t(z),\, \nabla\varphi(z)\rangle\, \mathrm{d}\alpha_t(z) = \int \langle \nabla\varphi(P_t(u)),\, [\partial_t P_t](u)\rangle \mathrm{d}\pi(u). \tag{14.8}$$

Comparing (14.7) and (14.8) yields (14.6), which is the desired continuity equation. □

The conditional expectation in (14.3) has a simple measure-theoretic meaning. Let $\alpha_t = (P_t)_\sharp \pi$ and define the vector-valued measure $m_t$ on $\mathbb{R}^d$ by

$$\int_{\mathbb{R}^d} \langle \psi(z),\, \mathrm{d}m_t(z)\rangle \coloneqq \int_{\mathbb{R}^{d'}} \langle \psi(P_t(u)),\, [\partial_t P_t](u)\rangle \mathrm{d}\pi(u)$$

for every bounded continuous vector field $\psi$. Since $\alpha_t(A) = 0$ implies $\pi(P_t^{-1}(A)) = 0$, one has $m_t \ll \alpha_t$. The Radon–Nikodym decomposition of $m_t$ with respect to $\alpha_t$ is therefore

$$\mathrm{d}m_t(z) = v_t(z)\mathrm{d}\alpha_t(z), \qquad v_t = \frac{\mathrm{d}m_t}{\mathrm{d}\alpha_t}.$$

In the language of Lebesgue decomposition, the flux measure $m_t$ has only an absolutely continuous part with respect to $\alpha_t$ and no singular part; the conditional expectation is precisely this density. Equivalently, disintegrating $\pi$ with respect to the map $P_t$ 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 $\{u : P_t(u) = z\}$, and

$$v_t(z) = \int_{\{P_t(u)=z\}} [\partial_t P_t](u)\mathrm{d}\pi_{t,z}(u).$$

Thus the solution of (14.2) is the conditional expectation of the velocities $\partial_t P_t$: 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 (14.2).

**Algorithm 14.1** Flow matching regression and sampling

**Input:** Interpolant $P_t(u)$, training source $u \sim \pi$, parametrized field $v_\theta(t,z)$, training steps $N$.
**Output:** Learned sampler $X_0 \mapsto X_1$.
**Training:**
**For** $q = 1, \dots, N$ **do**:
  **Draw** $t_q \sim \mathrm{Unif}(0,1)$ and $u_q \sim \pi$.
  **Set** $z_q = P_{t_q}(u_q)$ and $w_q = \partial_t P_t(u_q)|_{t=t_q}$.
  **Update** $\theta$ by one stochastic-gradient step on $\|v_\theta(t_q, z_q) - w_q\|^2$.

**Sampling:**
**Draw** $X_0 \sim \alpha_0$.
**Integrate** $\dot{X}_t = v_\theta(t, X_t)$, $t \in [0,1]$. **Return** $X_1$.

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 (14.1) is represented by a deterministic flow rather than by the original coupling. The sampling procedure consists in first drawing $X_0 \sim \alpha$, and then integrating 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 $\alpha_1 = \beta$.

**Connection with diffusion models.** In the special case where $P_t(x, y) = (1-t)x + ty$ is a linear interpolation and $\pi = \alpha \otimes \beta$, the curve $\alpha_t$ is a convolution of rescaled versions of $\alpha_0$ and $\alpha_1$. The flow-matching problem (14.2) 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 [215]. 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 [147, 148, 2]. The next two propositions are written in the noising direction, from a data law $\alpha$ 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 (12.7). The regression (14.2) is computationally simpler because the continuity equation has already been enforced by the chosen interpolant. To prove this, we rely on Tweedie's formula, which expresses the optimal Gaussian denoiser through the score, i.e. the gradient of the log-density.

**Proposition 14.3** (Tweedie identity). *Let $W$ be a random vector in $\mathbb{R}^d$ with density $\beta$. For $\sigma > 0$, observe*

$$Z = W + \sigma\, \varepsilon, \quad \text{where } \varepsilon \sim \mathcal{N}(0, I_d) \text{ is independent of } W.$$

*Denote by*

$$\beta_\sigma = \beta * \mathcal{N}\big(0, \sigma^2 I_d\big)$$

*the density of $Z$. Then*

$$\mathbb{E}\big[\, W \mid Z = z \,\big] = z + \sigma^2 \nabla \log \beta_\sigma(z) \qquad \text{for all } z \in \mathbb{R}^d.$$

*Proof.* Bayes' rule gives the conditional density $p_{W|Z}(w \mid z) = \dfrac{\beta(w)\, \varphi_\sigma(z-w)}{\beta_\sigma(z)}$ with $\varphi_\sigma$ the $\mathcal{N}(0, \sigma^2 I_d)$ density. Hence

$$\mathbb{E}[W \mid Z = z] = \frac{1}{\beta_\sigma(z)} \int_{\mathbb{R}^d} w\, \beta(w)\, \varphi_\sigma(z-w)\, \mathrm{d}w.$$

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

$$\nabla_z \beta_\sigma(z) = \int \beta(w)\, \nabla_z \varphi_\sigma(z-w)\, \mathrm{d}w = -\sigma^{-2}\Big(z - \mathbb{E}[W \mid Z = z]\Big)\, \beta_\sigma(z).$$

Rearranging finishes the proof. □

**Proposition 14.4** (Gaussian-endpoint flow-matching field). *Let $X \sim \alpha$ and $Y \sim \mathcal{N}(0, I_d)$ be independent. For $t \in (0,1)$ set*

$$Z_t = (1-t)\, X + t\, Y, \qquad \alpha_t = \mathrm{Law}(Z_t).$$

*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*

$$v^\star(x,t) = -\frac{1}{1-t}\, x - \frac{t}{1-t} \nabla \log \alpha_t(x) \qquad (x \in \mathbb{R}^d,\ t \in (0,1)).$$

*In particular, for each $t \in (0,1)$ this field is a gradient field,*

$$v^\star(\cdot, t) = -\nabla\left( \frac{\|\cdot\|^2}{2(1-t)} + \frac{t}{1-t} \log \alpha_t \right).$$

*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 14.3. 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 14.3 to $\mathbb{E}[W \mid Z_t = z]$ and noting $\mathbb{E}[Y \mid Z_t = z] = -\,t\, \nabla \log \alpha_t(z)$ gives the claimed formula. □

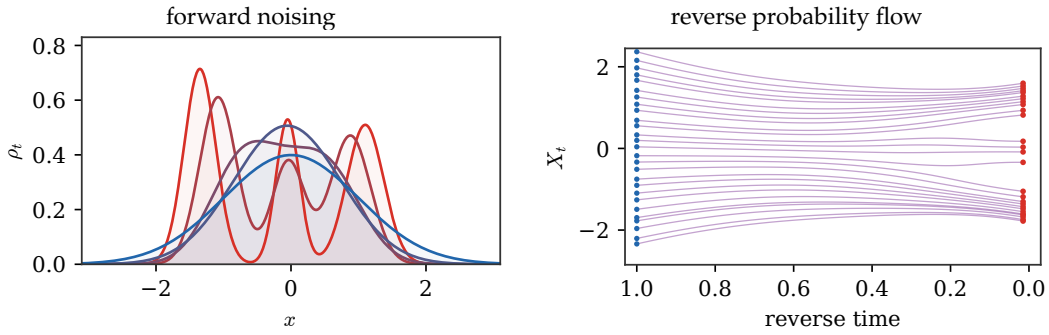



*Figure 14.2:* One-dimensional diffusion bridge for a Gaussian-mixture data law. The forward path $Z_t = (1-t)X + tY$ smooths the red data density toward a blue Gaussian endpoint. Reversing the probability-flow ODE transports a denser set of blue noise samples back toward the data modes, making the splitting of trajectories across mixture components visible.

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 the plots below, 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 14.4 gives $v_t(z) = -(z + t\sigma^2 s_t(z))/(1-t)$.

**Algorithm 14.2** Exact probability-flow sampling for a Gaussian mixture

**Input:** Gaussian-mixture data law, schedule $(a_t, b_t)$, noise level $\sigma$, number of samples $R$.
**Output:** Backward samples $(Z_0^{(r)})_r$.
**Define** the noising variable: $Z_t = a_t X + b_t Y, \qquad Y \sim \mathcal{N}(0, \sigma^2\mathrm{Id})$.
**Compute** closed-form mixture density $p_t$ and score $s_t = \nabla \log p_t$.
**Set** probability-flow velocity: $v_t(z) = \frac{a_t'}{a_t} z + \left(\frac{a_t' b_t^2}{a_t} - b_t' b_t\right) \sigma^2 s_t(z)$.
**For** $r = 1, \ldots, R$ **do**:
  **Draw** $Z_1^{(r)}$ from the Gaussian endpoint.
  **Integrate** $\dot{Z}_t^{(r)} = v_t(Z_t^{(r)})$ backward from $t = 1$ to $t = 0$.
**Return** $(Z_0^{(r)})_r$.

**When is the induced map optimal?** Integrating the learned velocity gives a deterministic map from $\alpha_0$ to $\alpha_1$, 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 [121].

**Proposition 14.5** (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*

$$\Sigma_t = (1-t)^2\Sigma_0 + t^2\Sigma_1. \tag{14.9}$$

*Then the exact flow-matching velocity is affine, $v_t(z) = A_t z$, with*

$$A_t = \big(t\Sigma_1 - (1-t)\Sigma_0\big)\Sigma_t^{-1}. \tag{14.10}$$

*The induced flow map $T_t^{\mathrm{FM}}$ from $\alpha_0$ to $\alpha_t$ is*

$$T_t^{\mathrm{FM}} = \Sigma_0^{1/2}\Big((1-t)^2\mathrm{Id} + t^2\Sigma_0^{-1/2}\Sigma_1\Sigma_0^{-1/2}\Big)^{1/2}\Sigma_0^{-1/2}. \tag{14.11}$$

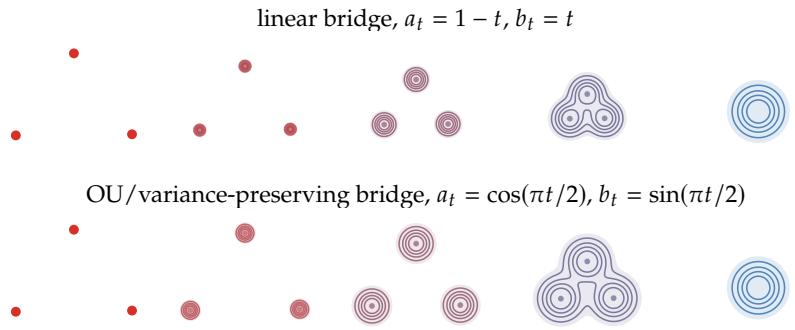



*Figure 14.3:* Two-dimensional noising paths from three Dirac masses to a single Gaussian. The top row shows the linear interpolation $Z_t = (1-t)X + tY$, whose component centers move linearly toward the origin and whose covariance grows like $(t\sigma)^2\mathrm{Id}$. The bottom row uses the variance-preserving 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.

*In particular,*

$$T_1^{\mathrm{FM}} = \Sigma_0^{1/2}\big(\Sigma_0^{-1/2}\Sigma_1\Sigma_0^{-1/2}\big)^{1/2}\Sigma_0^{-1/2}. \tag{14.12}$$

*This terminal map coincides with the quadratic optimal transport map*

$$T^{\mathrm{OT}} = \Sigma_0^{-1/2}\big(\Sigma_0^{1/2}\Sigma_1\Sigma_0^{1/2}\big)^{1/2}\Sigma_0^{-1/2} \tag{14.13}$$

*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 (14.10). 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

$$\widetilde{\Sigma}_t = (1-t)^2\mathrm{Id} + t^2 C.$$

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 (14.11), and $t = 1$ gives (14.12).

Both $T_1^{\mathrm{FM}}$ and $T^{\mathrm{OT}}$ 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 $T_1^{\mathrm{FM}}$ is symmetric positive definite. The map $T_1^{\mathrm{FM}}$ 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 $T_1^{\mathrm{FM}}$ is equivalent to $\Sigma_0 C^{1/2} = C^{1/2}\Sigma_0$, hence to $\Sigma_0 C = C\Sigma_0$ 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 $\Sigma_0$ and $\Sigma_1$ commute, they are orthogonally co-diagonalizable, and both (14.12) and (14.13) 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 14.6 (Changing the bridge speed does not restore optimality).** The same terminal map (14.12) is obtained for any scalar schedule $Z_t = a_t X_0 + b_t X_1$ with the same endpoints, because after whitening the covariance path remains $a_t^2 \mathrm{Id} + b_t^2 C$. Thus changing the speed of a 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 [138]. 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 follows the reverse field $z + \sigma^2 \nabla \log p_\tau(z)$ as $\tau$ decreases. This is the noising law used in the left panel of Figure 14.4; the 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 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 $\tau = -\log \cos(\pi t/2)$. Figure 14.5 therefore compares OU with a genuinely different scalar bridge,

$$a_t = (1 - t)(1 - 2t), \qquad b_t = t,$$

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.

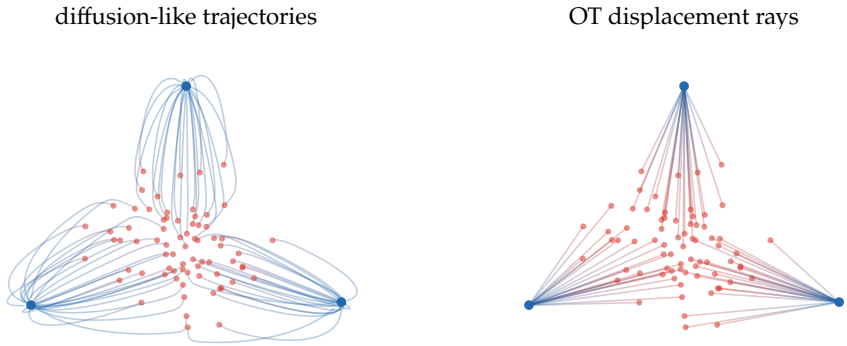



*Figure 14.4:* Diffusion-style sampling trajectories compared with OT rays in the three-Dirac setting of Figure 14.3. Red particles are sampled from the centered Gaussian endpoint and transported toward the three blue atoms. The left panel integrates the reverse probability-flow ODE for the variance-preserving 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.

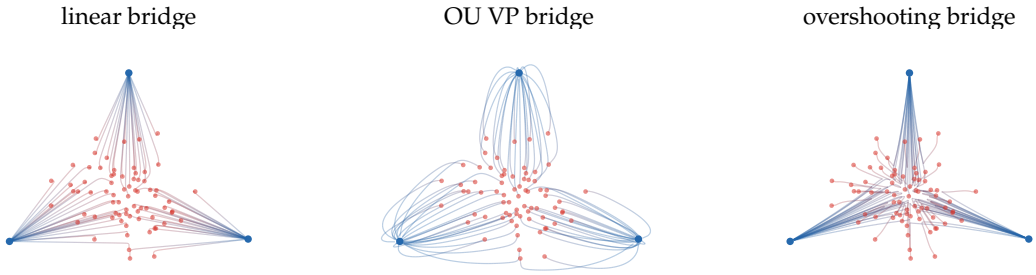



*Figure 14.5:* Effect of the interpolant on the exact reverse flow for the same three-Dirac target and the same Gaussian endpoint. The linear bridge $a_t = 1 - t$, $b_t = t$ produces almost radial curves. The variance-preserving OU bridge $a_\tau = e^{-\tau}$, $b_\tau = \sqrt{1 - e^{-2\tau}}$ changes 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.

## 14.2 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.

**Training a one-step flow.** Let $\zeta$ be a simple latent distribution and let $\alpha_\theta = (G_\theta)_\sharp \zeta$ be the model distribution. Assume that the target data distribution is $\beta$. A 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 7.8. The associated formal descent is

$$\partial_t \mu_t + \mathrm{div}(\mu_t w_t) = 0, \qquad w_t(x) = -\nabla \delta_\alpha \mathcal{E}_\beta(\mu_t)(x). \tag{14.14}$$

Instead of integrating (14.14) at inference time, one fits a parametric residual field $U_\eta$ along the current model distribution:

$$\min_\eta \int_0^1 \int \|U_\eta(t, x) - w_t(x)\|^2 \, \mathrm{d}\mu_t(x) \, \mathrm{d}t. \tag{14.15}$$

In a particle or generator implementation, the learned residual is then used to update the current generator by

$$\alpha_\theta^+ = (\mathrm{Id} + \tau U_\eta)_\sharp \alpha_\theta, \qquad \text{or equivalently} \qquad G_\theta^+(z) = G_\theta(z) + \tau U_\eta(G_\theta(z)).$$

**Algorithm 14.3** One-step Wasserstein-flow generator update

**Input:** Generator $G_{\theta_k}$, latent law $\zeta$, data law $\beta$, numerical descent-field oracle $W_\beta$, step size $\tau$, batch size $B$.
**Output:** Updated generator $G_{\theta_{k+1}}$.
**Draw** $z_b \sim \zeta$ for $b = 1, \dots, B$.
**Set** $x_b = G_{\theta_k}(z_b)$.
**Set** $w_k(x) = W_\beta[\alpha_{\theta_k}](x)$, where $W_\beta[\alpha] = -\nabla \delta_\alpha \mathcal{E}_\beta(\alpha)$.
**Set** $\eta_k$ by minimizing the empirical least-squares loss: $\frac{1}{B}\sum_{b=1}^{B} \|U_\eta(x_b) - w_k(x_b)\|^2$.
**Update by composition:** $G_{\theta_{k+1}}(z) = G_{\theta_k}(z) + \tau U_{\eta_k}(G_{\theta_k}(z))$. **Return** $G_{\theta_{k+1}}$.

After many training updates, the accumulated 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 [117], while drifting methods evolve the generated distribution during training through a fitted vector field and also admit one-step inference [78]. The gradient-flow interpretation of drifting models, and its relation to KL, MMD, sliced-Wasserstein and Sinkhorn-type discrepancies, is analyzed in [113, 120]. These ideas are also connected to the Sinkhorn-type normalization dynamics used to model attention in Sinkformers [201].

**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 $L^2(\mu_t)$. A simple continuous version uses a positive kernel $K_\varepsilon(x, y)$ and defines, for any measure $\nu$,

$$B_\varepsilon[\nu](x) := \frac{\int (y - x) K_\varepsilon(x, y)\, \mathrm{d}\nu(y)}{\int K_\varepsilon(x, y)\, \mathrm{d}\nu(y)}. \tag{14.16}$$

For the Gaussian kernel $K_\varepsilon(x, y) = \exp(-\|x - y\|^2/(2\varepsilon))$, this normalized field is a score of a smoothed density:

$$B_\varepsilon[\nu](x) = \varepsilon \nabla \log\left( \int K_\varepsilon(x, y)\, \mathrm{d}\nu(y) \right). \tag{14.17}$$

The drifting velocity is then

$$u_t(x) = B_\varepsilon[\beta](x) - B_\varepsilon[\mu_t](x) = \varepsilon \nabla \log \frac{\int K_\varepsilon(x, y)\, \mathrm{d}\beta(y)}{\int K_\varepsilon(x, y)\, \mathrm{d}\mu_t(y)}. \tag{14.18}$$

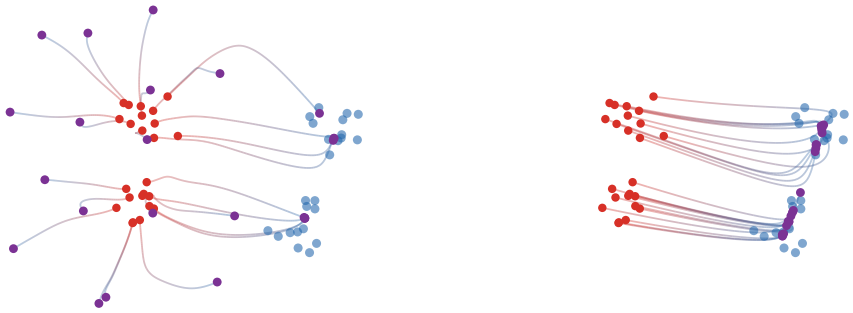
The first term pulls samples toward data, while the second term corrects self-attraction and prevents all particles from collapsing onto the same high-density region. Sinkhorn drifting replaces these one-sided kernel normalizations by two-sided entropic OT couplings, so that the cross and self terms are normalized by Sinkhorn scaling rather than by a single denominator [120].

**Algorithm 14.4** Self-corrected drifting particle update

**Input:** Particles $x_i^k$ for $\mu_k$, data samples $(y_b)_{b=1}^B$ from $\beta$, kernel scale $\varepsilon$, step $h$.
**Output:** Updated particles $x_i^{k+1}$.
**For** each particle $i$ **do**:
  **Set** $Z_{\beta,i} = \sum_{b=1}^{B} K_\varepsilon(x_i^k, y_b)$ and $b_i^k = Z_{\beta,i}^{-1} \sum_{b=1}^{B} (y_b - x_i^k) K_\varepsilon(x_i^k, y_b)$.
  **Set** $Z_{\mu,i} = \sum_{j=1}^{n} K_\varepsilon(x_i^k, x_j^k)$ and $m_i^k = Z_{\mu,i}^{-1} \sum_{j=1}^{n} (x_j^k - x_i^k) K_\varepsilon(x_i^k, x_j^k)$.
  **Set** $u_i^k = b_i^k - m_i^k$.
  **Update** $x_i^{k+1} = x_i^k + h\, u_i^k$.
**Return** $(x_i^{k+1})_i$.

**Proposition 14.7** (Drifting as a time-dependent Wasserstein gradient). *Let $\mu_t$ be a smooth curve of positive densities and let $u_t = \nabla \varphi_t$ be a smooth time-dependent gradient field. Define the semi-relaxed functional*

$$\mathcal{R}_t(\alpha | \mu_t) := -\int \varphi_t(x)\, \mathrm{d}\alpha(x) + \int \varphi_t(x)\, \mathrm{d}\mu_t(x). \tag{14.19}$$



raw kernel drift self-corrected drift

*Figure 14.6:* 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 $B_\varepsilon[\beta] - B_\varepsilon[\mu_t]$, so a longer integration brings particles to the blue modes while repelling them from their own current concentration.

*Here $\mu_t$ and $\varphi_t$ are frozen when taking the first variation with respect to the first argument $\alpha$. Then the continuity equation*

$$\partial_t \mu_t + \operatorname{div}(\mu_t u_t) = 0$$

*is the formal Wasserstein gradient descent of the time-dependent functional $\alpha \mapsto \mathcal{R}_t(\alpha|\mu_t)$.*

*Proof.* Since $\mu_t$ and $\varphi_t$ are fixed in the variation with respect to $\alpha$, the first variation is

$$\delta_\alpha \mathcal{R}_t(\alpha|\mu_t)(x) = -\varphi_t(x).$$

By Proposition 13.2,

$$\nabla_{\mathcal{W}} \mathcal{R}_t(\alpha|\mu_t) = \nabla \delta_\alpha \mathcal{R}_t(\alpha|\mu_t) = -\nabla\varphi_t = -u_t.$$

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. □

**Example 14.8 (Kernel drifting as a semi-relaxed divergence).** For the Gaussian-kernel drift (14.18), set

$$\varphi_t(x) = \varepsilon \log \frac{\int K_\varepsilon(x,y)\,\mathrm{d}\beta(y)}{\int K_\varepsilon(x,y)\,\mathrm{d}\mu_t(y)}.$$

Then $u_t = \nabla\varphi_t$, so Proposition 14.7 shows that kernel drifting is the Wasserstein gradient descent of

$$\mathcal{R}_t^{\mathrm{drift}}(\alpha|\mu_t) = \varepsilon \int \log \frac{\int K_\varepsilon(x,y)\,\mathrm{d}\mu_t(y)}{\int K_\varepsilon(x,y)\,\mathrm{d}\beta(y)}\,\mathrm{d}\alpha(x) + \text{constant}.$$

It is "semi-relaxed" because the current model $\mu_t$ is used to build the potential, but it is not varied inside the denominator when computing the first variation in $\alpha$.

**Remark 14.9 (General fields and projection onto gradients).** A general regressed field $b_t$ is not necessarily a Wasserstein gradient, since Wasserstein tangent vectors are represented by gradient fields modulo $L^2(\mu_t)$-null directions. The gradient component is obtained by the weighted projection

$$\nabla\varphi_t = \operatorname*{argmin}_{\nabla\varphi} \int \|\nabla\varphi(x) - b_t(x)\|^2\,\mathrm{d}\mu_t(x).$$

One may first normalize $b_t$ pointwise, for instance by $b_t/(\|b_t\| + \eta)$, or globally by $\|b_t\|_{L^2(\mu_t)}$, before this projection. Proposition 14.7 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.

## 14.3 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.

Transformers were introduced as sequence-to-sequence architectures driven by self-attention [223] and have since become a central architecture for language and vision models [45, 81]. Their distinctive feature is that each token is updated by a data-dependent average of all other tokens. This makes an attention layer permutation-equivariant before positional encoding, context dependent after conditioning on the input sequence, and naturally compatible with a measure viewpoint in which a prompt is regarded as an empirical distribution of tokens.

The mathematical limit used below concerns depth rather than model scale: one lets the number of residual attention layers grow while each layer makes a small update, as in continuous-depth neural networks [58]. For attention, the resulting velocity is nonlinear in the current token law because it is normalized by the whole context. This measure-theoretic view appears in the analysis of attention as a Lipschitz or interacting-particle operator [230, 106], in the Sinkhorn-normalized dynamics of Sinkformers [201], and in recent well-posedness and mean-field-limit results for several attention mechanisms [54]. It also separates the infinite-depth limit studied here from the token-limit question, where one controls how a finite empirical context approximates its limiting attention operator [36].

We now consider very deep transformers, focusing on a single-head attention mechanism for simplicity while ignoring MLP layers, layer normalization, causality, and masking. This stripped-down framework is best suited to modeling encoders and vision transformers; the references above indicate which parts of this simplified picture extend to richer attention mechanisms.

**Attention as a context-dependent velocity.** After tokenization, embedding, and positional encoding, each input (from a set of tokens) is represented as a point cloud $(x_i)_{i=1}^n$ of $n$ points in the space of vectorized tokens. An attention layer with skip connection and rescaling by $1/T$ (where $T$ is the depth) defines a transformation of the tokens:

$$x_i \mapsto x_i + \frac{1}{T} \sum_j \frac{e^{\langle Qx_i, Kx_j \rangle} V x_j}{\sum_\ell e^{\langle Qx_i, Kx_\ell \rangle}},$$

where $\theta = (K, Q, V)$ are the parameters of the attention layer, represented by three matrices.

**Token measure evolution.** To handle an arbitrary number of tokens, we define $\alpha = \frac{1}{n} \sum_i \delta_{x_i}$ as the empirical measure of tokens and rewrite the transformer mapping as:

$$x_i \mapsto x_i + \frac{1}{T} \Gamma_\theta[\alpha](x_i),$$

where

$$\Gamma_\theta[\alpha](x) := \frac{\int e^{\langle Qx, Ky \rangle} V y \, \mathrm{d}\alpha(y)}{\int e^{\langle Qx, Kz \rangle} \, \mathrm{d}\alpha(z)}.$$

At the level of the token distribution, the layer pushes $\alpha$ forward by the "in-context" mapping $\Gamma_{\theta_t}[\alpha]$, which depends on the context $\alpha$, the tokens, and the depth-dependent parameters $\theta_t$. Denoting $t \in [0, 1]$ as the depth and $\tau = 1/T$ as the step size, this gives:

$$\alpha_{t+\tau} = (\mathrm{Id} + \tau \Gamma_{\theta_t}[\alpha_t])_\sharp \alpha_t.$$

As $\tau \to 0$, this converges formally to the conservation equation

$$\partial_t \alpha_t + \mathrm{div}(\alpha_t \Gamma_{\theta_t}[\alpha_t]) = 0.$$

**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

$$\Phi_\alpha(x) = \int \exp(\langle Qx, \, Ky \rangle) \mathrm{d}\alpha(y), \qquad U_\alpha(x) = \log \Phi_\alpha(x).$$

**Algorithm 14.5** Residual attention depth evolution

**Input:** Tokens $(x_i^0)_{i=1}^n$, depth $T$, layer parameters $(Q_k, K_k, V_k)$.
**Output:** Final token measure $\alpha_T$.
**Initialize:** $\alpha_0 = \frac{1}{n}\sum_{i=1}^n \delta_{x_i^0}$, $\tau = 1/T$.
**For** $k = 0, \dots, T-1$ **do**:
  **For** $i = 1, \dots, n$ **do**
    $\Gamma_{\theta_k}[\alpha_k](x_i^k) = \frac{\sum_j \exp(\langle Q_k x_i^k, K_k x_j^k\rangle) V_k x_j^k}{\sum_j \exp(\langle Q_k x_i^k, K_k x_j^k\rangle)}$.
    $x_i^{k+1} = x_i^k + \tau\, \Gamma_{\theta_k}[\alpha_k](x_i^k)$.
  **Set** $\alpha_{k+1} = (\mathrm{Id} + \tau\Gamma_{\theta_k}[\alpha_k])_\sharp \alpha_k$.
**Return** $\alpha_T$.

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 is an instantaneous gradient in $x$. It is not, however, the gradient of the first variation of a fixed functional of $\alpha$, because the potential $U_\alpha$ itself depends on the current measure through the same attention normalization. Thus the PDE is generally a transportation dynamics, not a Wasserstein gradient flow. Special variants recover additional structure: Sinkhorn attention can be interpreted through doubly stochastic normalization and Wasserstein-type gradient flows [201, 54], 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.

## 14.4 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 [181] 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.

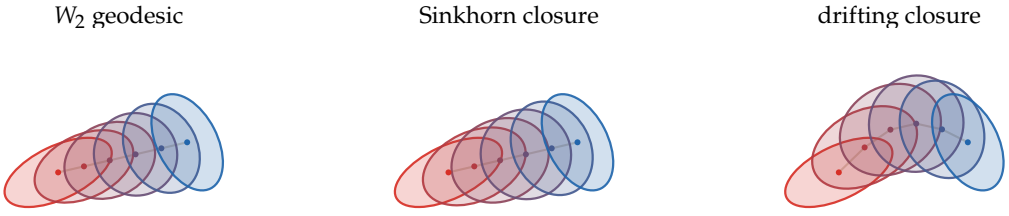



*Figure 14.7:* Gaussian closures of transport dynamics between two overlapping anisotropic Gaussians. The left panel is the exact $W_2$ Gaussian geodesic. The middle panel shows a regularized Sinkhorn-style closure, where the same mean path is accompanied by inflated intermediate covariances. The right panel shows a drifting-style closure with a curved mean path and moment-matched covariance ellipses. These finite-dimensional pictures keep only means and covariances, and therefore discard higher-order shape information that would be created by a genuinely nonlinear velocity field.

**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 14.10** (Affine velocities preserve Gaussianity)**.** *Let $\alpha_t = \mathcal{N}(\mathbf{m}_t, \mathbf{\Sigma}_t)$, with $\mathbf{\Sigma}_t$ positive definite, solve the continuity equation with an affine velocity*

$$v_t(x) = b_t + A_t(x - \mathbf{m}_t).$$

*Then $\alpha_t$ remains Gaussian and its moments solve*

$$\dot{\mathbf{m}}_t = b_t, \qquad \dot{\mathbf{\Sigma}}_t = A_t\mathbf{\Sigma}_t + \mathbf{\Sigma}_t A_t^\top.$$

*Conversely, any smooth Gaussian curve with positive definite covariance can be generated by such an affine velocity. If one wants the velocity to be a Wasserstein tangent gradient, one chooses the unique symmetric solution of the Lyapunov equation*

$$A_t\mathbf{\Sigma}_t + \mathbf{\Sigma}_t A_t = \dot{\mathbf{\Sigma}}_t.$$

*Proof.* Let $X_t$ follow the characteristic ODE $\dot{X}_t = b_t + A_t(X_t - \mathbf{m}_t)$. This linear ODE maps Gaussian random variables to Gaussian random variables. Taking expectation gives $\dot{\mathbf{m}}_t = b_t$. Writing $\tilde{X}_t = X_t - \mathbf{m}_t$, one has $\dot{\tilde{X}}_t = A_t\tilde{X}_t$, hence

$$\dot{\mathbf{\Sigma}}_t = \frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}(\tilde{X}_t\tilde{X}_t^\top) = A_t\mathbf{\Sigma}_t + \mathbf{\Sigma}_t A_t^\top.$$

For the converse, set $b_t = \dot{\mathbf{m}}_t$ and choose any matrix $A_t$ satisfying the covariance equation. Since $\mathbf{\Sigma}_t$ is positive definite, the Lyapunov map $A \mapsto A\mathbf{\Sigma}_t + \mathbf{\Sigma}_t A$ 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 $x \mapsto \langle b_t,\, x\rangle + \langle A_t(x - \mathbf{m}_t),\, x - \mathbf{m}_t\rangle/2$. □

**Constrained evolution on the Gaussian manifold.** For non-affine velocities, the finite-dimensional substitute is to project the Wasserstein dynamics onto the Gaussian manifold.

Let

$$\mathcal{G} = \{\mathcal{N}(\mathbf{m}, \mathbf{\Sigma}) : \mathbf{m} \in \mathbb{R}^d,\ \mathbf{\Sigma} \succ 0\}$$

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

$$\alpha^{k+1} \in \operatorname*{argmin}_{\alpha\in\mathcal{M}} \frac{1}{2\tau}\mathcal{W}_2^2(\alpha, \alpha^k) + f(\alpha).$$

For $\mathcal{M} = \mathcal{G}$, tangent velocities are affine gradient fields $v(x) = b + A(x - \mathbf{m})$ with $A = A^\top$. The constrained gradient is therefore the $L^2(\mathcal{N}(\mathbf{m}, \mathbf{\Sigma}))$ projection of the ambient Wasserstein gradient onto this finite-dimensional affine space, whenever the ambient gradient exists.

**Proposition 14.11** (Gaussian-constrained Wasserstein gradients)**.** *Let $f$ be a smooth functional and assume that its restriction to nondegenerate Gaussian measures can be written as*

$$f(\mathcal{N}(\mathbf{m}, \mathbf{\Sigma})) = F(\mathbf{m}, \mathbf{\Sigma}).$$

*Then the Wasserstein gradient constrained to the Gaussian family is the affine vector field*

$$v_F(x) = \nabla_{\mathbf{m}}F(\mathbf{m}, \mathbf{\Sigma}) + 2\nabla_{\mathbf{\Sigma}}F(\mathbf{m}, \mathbf{\Sigma})(x - \mathbf{m}),$$

*where $\nabla_{\mathbf{\Sigma}}F$ denotes the symmetric matrix derivative. Equivalently, $v_F$ is the $L^2(\mathcal{N}(\mathbf{m}, \mathbf{\Sigma}))$ 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*

$$\dot{\mathbf{m}}_t = -\nabla_{\mathbf{m}}F(\mathbf{m}_t, \mathbf{\Sigma}_t), \qquad \dot{\mathbf{\Sigma}}_t = -2\big(\mathbf{\Sigma}_t\nabla_{\mathbf{\Sigma}}F(\mathbf{m}_t, \mathbf{\Sigma}_t) + \nabla_{\mathbf{\Sigma}}F(\mathbf{m}_t, \mathbf{\Sigma}_t)\mathbf{\Sigma}_t\big), \tag{14.20}$$

*and the descent velocity is affine.*

*Proof.* Test the functional along a Gaussian tangent vector, represented by an affine gradient field

$$v(x) = b + A(x - \mathbf{m})$$

with $A$ symmetric. The induced first-order variations are $\dot{\mathbf{m}} = b$ and $\dot{\mathbf{\Sigma}} = A\mathbf{\Sigma} + \mathbf{\Sigma}A$. Therefore

$$\mathrm{d}F(\mathbf{m}, \mathbf{\Sigma})[b, A\mathbf{\Sigma} + \mathbf{\Sigma}A] = \langle \nabla_{\mathbf{m}} F,\, b\rangle + \operatorname{tr}(\nabla_{\mathbf{\Sigma}} F(A\mathbf{\Sigma} + \mathbf{\Sigma}A))\,.$$

Since $A$, $\mathbf{\Sigma}$ and $\nabla_{\mathbf{\Sigma}} F$ are symmetric, the second term equals

$$2\operatorname{tr}(\nabla_{\mathbf{\Sigma}} F\, A\mathbf{\Sigma}) = \int \langle 2\nabla_{\mathbf{\Sigma}} F(x - \mathbf{m}),\, A(x - \mathbf{m})\rangle \mathrm{d}\mathcal{N}(\mathbf{m}, \mathbf{\Sigma})(x).$$

Together with the mean term, this gives

$$\mathrm{d}F(\mathbf{m}, \mathbf{\Sigma})[\dot{\mathbf{m}}, \dot{\mathbf{\Sigma}}] = \int \langle v_F(x),\, v(x)\rangle \mathrm{d}\mathcal{N}(\mathbf{m}, \mathbf{\Sigma})(x)$$

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 14.10 gives (14.20). □

This proposition should be read as the organizing rule for Gaussian closures: once the scalar energy has been reduced to a function of $(\mathbf{m}, \mathbf{\Sigma})$, its constrained Wasserstein gradient is automatically affine and the covariance follows the Bures-type ODE (14.20). When the first variation of $f$ is quadratic, this constrained gradient coincides with the full Wasserstein gradient.

**Gaussian-preserving gradient flows.** The next examples show that many familiar energies already have affine Wasserstein gradients on Gaussian inputs, so their full flow remains inside the Gaussian family.

**Example 14.12 (Gaussian energies and affine gradients).** Proposition 14.11 turns many standard energies into explicit affine fields:

- *Quadratic potential energy.* If

  $$f(\alpha) = \int \Big(\frac{1}{2} x^\top H x + \langle \ell,\, x\rangle\Big) \mathrm{d}\alpha(x), \qquad H = H^\top,$$

  then

  $$\nabla_{\mathcal{W}} f(\alpha)(x) = Hx + \ell = (H\mathbf{m} + \ell) + H(x - \mathbf{m}).$$

  This is the Gaussian form of transport under a quadratic confinement.

- *Quadratic interaction energy.* If

  $$f(\alpha) = \frac{1}{4} \iint (x - y)^\top G(x - y) \mathrm{d}\alpha(x) \mathrm{d}\alpha(y), \qquad G = G^\top,$$

  then $F(\mathbf{m}, \mathbf{\Sigma}) = \frac{1}{2}\operatorname{tr}(G\mathbf{\Sigma})$ and

  $$\nabla_{\mathcal{W}} f(\alpha)(x) = G(x - \mathbf{m}).$$

  The mean is unchanged and the covariance contracts or expands according to the signs of $G$.

- *Relative entropy to a Gaussian.* For $\bar{\alpha} = \mathcal{N}(\bar{\mathbf{m}}, \bar{\mathbf{\Sigma}})$,

  $$f(\alpha) = \mathrm{KL}(\alpha | \bar{\alpha})$$

  has

  $$\nabla_{\mathcal{W}} f(\alpha)(x) = \bar{\mathbf{\Sigma}}^{-1}(\mathbf{m} - \bar{\mathbf{m}}) + (\bar{\mathbf{\Sigma}}^{-1} - \mathbf{\Sigma}^{-1})(x - \mathbf{m}).$$

  The descent equations are the Ornstein–Uhlenbeck moment equations

  $$\dot{\mathbf{m}}_t = -\bar{\mathbf{\Sigma}}^{-1}(\mathbf{m}_t - \bar{\mathbf{m}}), \qquad \dot{\mathbf{\Sigma}}_t = 2\mathrm{Id} - \bar{\mathbf{\Sigma}}^{-1}\mathbf{\Sigma}_t - \mathbf{\Sigma}_t\bar{\mathbf{\Sigma}}^{-1}.$$

- *Squared Wasserstein distance to a Gaussian.* For

  $$f(\alpha) = \frac{1}{2}\mathcal{W}_2^2(\alpha, \bar{\alpha}), \qquad \bar{\alpha} = \mathcal{N}(\bar{\mathbf{m}}, \bar{\mathbf{\Sigma}}),$$

the Gaussian Brenier map $T_{\alpha\to\bar\alpha}$ is affine,

$$T_{\alpha\to\bar\alpha}(x) = \bar{\mathbf{m}} + M(x-\mathbf{m}), \qquad M = \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{\Sigma}^{1/2}\bar{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{1/2})^{1/2}\boldsymbol{\Sigma}^{-1/2}.$$

Hence

$$\nabla_{\mathcal{W}} f(\alpha)(x) = x - T_{\alpha\to\bar\alpha}(x) = (\mathbf{m}-\bar{\mathbf{m}}) + (\mathrm{Id}-M)(x-\mathbf{m}),$$

and descent moves each Gaussian infinitesimally along the Bures–Wasserstein geodesic toward $\bar\alpha$.

- *Gaussian-only losses.* Sliced $\mathrm{SW}_2^2$ losses to a Gaussian, Gaussian Sinkhorn divergences, and any smooth closed formula depending only on $(\mathbf{m}, \boldsymbol{\Sigma})$ fit the same constrained-gradient template:

$$v_F(x) = \nabla_{\mathbf{m}} F + 2\nabla_{\boldsymbol{\Sigma}} F(x-\mathbf{m}).$$

For Gaussian Sinkhorn divergences this finite-dimensional flow is studied in [119].

Not every PDE preserves Gaussianity exactly. For instance, Wasserstein flows of relative Fisher information, related to quantum-drift or higher-order diffusion equations, typically require a Gaussian projection to close on $(\mathbf{m}, \boldsymbol{\Sigma})$. Such projected closures are still useful: they expose the finite-dimensional dynamics predicted by a variational model and make it easy to compare variational flows with non-variational affine dynamics such as drifting fields or the Gaussian transformer closure below.

**Proposition 14.13** (Centered Gaussian covariance catalogue)**.** *Let $\gamma = \mathcal{N}(0, \mathrm{Id})$ and let $\mu_t = \mathcal{N}(0, C_t)$ with $C_t \succ 0$. For the normalizations displayed below, the Wasserstein descent constrained to the centered Gaussian manifold satisfies $\dot C_t = h(C_t)$, with*

$$\begin{aligned}
\mathrm{KL}(\mu|\gamma) &: h(C) = 2(\mathrm{Id} - C),\\
\tfrac{1}{2}\mathcal{I}(\mu|\gamma) &: h(C) = 2(C^{-1} - C),\\
\mathcal{W}_2^2(\mu,\gamma) &: h(C) = 4(C^{1/2} - C),\\
\mathrm{MMD}_k^2(\mu,\gamma),\quad k(x,y) = \langle x, y\rangle^2 &: h(C) = 8(C - C^2),\\
S_\varepsilon(\mu,\gamma) &: h(C) = 4\left(C + \tfrac{\varepsilon^2}{16}\mathrm{Id}\right)^{1/2} - 2\left(C^2 + \tfrac{\varepsilon^2}{16}\mathrm{Id}\right)^{1/2} - 2C - \tfrac{\varepsilon}{2}\mathrm{Id},\\
\mathrm{SW}_2^2(\mu,\gamma) &: h(C) = V(C)C + CV(C),
\end{aligned}$$

*where $S_\varepsilon$ is the debiased Sinkhorn divergence for the quadratic cost $\|x-y\|^2$ and KL regularization strength $\varepsilon$, and*

$$V(C) = 2\int_{\mathbb{S}^{d-1}} \left(\frac{1}{\sqrt{\theta^\top C\theta}} - 1\right)\theta\theta^\top \,\mathrm{d}\sigma(\theta)$$

*for the normalized spherical measure $\sigma$. Here*

$$\mathcal{I}(\mu|\gamma) = \int \left|\nabla \log\rho(x) + x\right|^2 \rho(x)\,\mathrm{d}x \qquad (\mu = \rho\,\mathrm{d}x).$$

*Thus the unhalved Fisher divergence has right-hand side $4(C^{-1} - C)$. Multiplying any of these energies by a constant simply rescales the corresponding right-hand side.*

*Proof.* Each row is obtained by identifying the affine descent velocity $v(x) = M_C x$ generated by the corresponding Gaussian-constrained calculation and then applying Proposition 14.10, which gives $\dot C = M_C C + C M_C^\top$. For $\mathrm{KL}(\cdot|\gamma)$, the Fokker–Planck velocity is $(C^{-1} - \mathrm{Id})x$, hence $\dot C = 2(\mathrm{Id} - C)$. For the Fisher row, the restriction of $\frac{1}{2}\mathcal{I}$ to centered Gaussians is

$$\frac{1}{2}\left(\mathrm{tr}(C) + \mathrm{tr}(C^{-1}) - 2d\right).$$

Using Proposition 14.11 gives the descent velocity $(C^{-2} - \mathrm{Id})x$, hence $\dot C = 2(C^{-1} - C)$. This row should be read as a Gaussian projected closure of the fourth-order Fisher flow.

For $\mathcal{W}_2^2(\cdot,\gamma)$, the Brenier map from $\mathcal{N}(0, C)$ to $\gamma$ is $C^{-1/2}x$, so the descent velocity for the unhalved squared distance is $2(C^{-1/2} - \mathrm{Id})x$, giving $4(C^{1/2} - C)$. For the polynomial MMD row, centered Gaussians satisfy $\mathrm{MMD}_k^2(\mu,\gamma) = \|C - \mathrm{Id}\|_{\mathrm{F}}^2$; the first variation is quadratic and its descent velocity is $4(\mathrm{Id} - C)x$, giving $8(C - C^2)$.

Gaussian Sinkhorn dual potentials are quadratic, so the velocity is again linear; differentiating the closed Gaussian formula yields the displayed spectral expression. The square roots are spectral functions of $C$, hence commute with $C$, which is why the covariance ODE closes as a matrix function of $C$ alone. For sliced Wasserstein, each one-dimensional projection is a Gaussian transport with velocity $2((\theta^\top C\theta)^{-1/2} - 1)\langle\theta,\, x\rangle\theta$; averaging these velocities over $\mathbb{S}^{d-1}$ gives $v(x) = V(C)x$ and thus $\dot{C} = V(C)C + CV(C)$. □

**Example 14.14 (Linear mean-field networks as cross-moment flows).** In the two-layer model above, take the linear activation $\sigma(s) = s$, so that

$$\psi((u,v),z) = v\,\langle u,\, z\rangle.$$

The predictor is the linear map

$$G_{\alpha_t}(z) = Q_t z, \qquad Q_t = \int vu^\top \mathrm{d}\alpha_t(u,v) \in \mathbb{R}^{d'\times d}.$$

Thus the energy depends on the neuron law only through the cross moment $Q_t$, a subblock of the raw second moment of $x = (u,v)$. For square loss, set

$$S = \int zz^\top \mathrm{d}\rho(z,y), \qquad R = \int yz^\top \mathrm{d}\rho(z,y), \qquad H_t = Q_t S - R.$$

The first variation is

$$\delta f(\alpha_t)(u,v) = \langle H_t,\, vu^\top\rangle = v^\top H_t u.$$

Hence the particle velocity in parameter space is linear:

$$-\nabla_{(u,v)}\delta f(\alpha_t)(u,v) = -\begin{pmatrix} 0 & H_t^\top \\ H_t & 0 \end{pmatrix}\begin{pmatrix} u \\ v \end{pmatrix}.$$

Therefore a Gaussian law of neurons remains Gaussian. Its mean and covariance follow Proposition 14.10, with a matrix depending only on the current cross moment $Q_t$, a raw second-moment subblock that becomes a covariance subblock when the neuron law is centered. This exact closure is special to the linear activation; for nonlinear activations, Gaussian closures are usually projections rather than invariant families.

**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 14.15 (Flow matching and diffusion paths between Gaussians).** Consider a prescribed Gaussian interpolation $\alpha_t = \mathcal{N}(\mathbf{m}_t, \mathbf{\Sigma}_t)$. Proposition 14.10 shows that an exact flow-matching velocity can be taken affine:

$$v_t(x) = \dot{\mathbf{m}}_t + A_t(x - \mathbf{m}_t), \qquad A_t\mathbf{\Sigma}_t + \mathbf{\Sigma}_t A_t = \dot{\mathbf{\Sigma}}_t.$$

In the isotropic case $\mathbf{\Sigma}_t = s_t^2\mathrm{Id}$, this reduces to the transparent formula

$$v_t(x) = \dot{\mathbf{m}}_t + \frac{\dot{s}_t}{s_t}(x - \mathbf{m}_t).$$

For instance, the diffusion noising path

$$X_t = a_t X_0 + \sigma_t Z, \qquad Z \sim \mathcal{N}(0, \mathrm{Id}),$$

has $\mathbf{m}_t = a_t\mathbf{m}_0$ and $\mathbf{\Sigma}_t = a_t^2\mathbf{\Sigma}_0 + \sigma_t^2\mathrm{Id}$. 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 14.16 (Gaussian kernel drifting).** Let the target be $\beta = \mathcal{N}(\bar{\mathbf{m}}, \bar{\mathbf{\Sigma}})$ and assume $\mu_t = \mathcal{N}(\mathbf{m}_t, \mathbf{\Sigma}_t)$. For the Gaussian kernel

$$K_\varepsilon(x,y) = \exp(-\|x-y\|^2/(2\varepsilon)),$$

the normalized field (14.16) satisfies

$$B_\varepsilon[\mu_t](x) = -\varepsilon(\mathbf{\Sigma}_t + \varepsilon\mathrm{Id})^{-1}(x - \mathbf{m}_t).$$

Thus the drifting velocity (14.18) is affine and preserves Gaussianity. With

$$A_t = (\mathbf{\Sigma}_t + \varepsilon \mathrm{Id})^{-1}, \qquad \bar{A} = (\bar{\mathbf{\Sigma}} + \varepsilon \mathrm{Id})^{-1},$$

the ODE is

$$\dot{\mathbf{m}}_t = \varepsilon \bar{A}(\bar{\mathbf{m}} - \mathbf{m}_t), \qquad \dot{\mathbf{\Sigma}}_t = \varepsilon\big((A_t - \bar{A})\mathbf{\Sigma}_t + \mathbf{\Sigma}_t(A_t - \bar{A})\big).$$

This finite-dimensional model explains the stabilizing role of the self-normalized repulsion term in drifting: without it, the covariance equation loses the $A_t\mathbf{\Sigma}_t + \mathbf{\Sigma}_t A_t$ contribution.

**Example 14.17 (Gaussian closure of attention dynamics).** For the transformer PDE, assume $\alpha = \mathcal{N}(\mathbf{m}, \mathbf{\Sigma})$. Since exponential tilting preserves Gaussianity,

$$\frac{\int e^{\langle Qx, Ky\rangle}\, y \,\mathrm{d}\alpha(y)}{\int e^{\langle Qx, Kz\rangle}\,\mathrm{d}\alpha(z)} = \mathbf{m} + \mathbf{\Sigma}K^\top Q x.$$

Therefore

$$\Gamma_\theta[\alpha](x) = V\mathbf{m} + V\mathbf{\Sigma}K^\top Q x$$

is affine. The Gaussian token law is preserved and satisfies

$$\dot{\mathbf{m}}_t = (V_t + V_t\mathbf{\Sigma}_t K_t^\top Q_t)\mathbf{m}_t, \qquad \dot{\mathbf{\Sigma}}_t = B_t\mathbf{\Sigma}_t + \mathbf{\Sigma}_t B_t^\top, \qquad B_t = V_t\mathbf{\Sigma}_t K_t^\top Q_t.$$

When $V_t = Q_t^\top K_t$, the matrix $B_t = Q_t^\top K_t \mathbf{\Sigma}_t K_t^\top Q_t$ 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 [102] 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.

**Theorem 14.18** (Gelbrich theorem)**.** *For $\mu \in \mathcal{P}_2(\mathbb{R}^d)$, let*

$$\mathcal{R}\mu := \mathcal{N}(\mathbf{m}_\mu, \mathbf{\Sigma}_\mu)$$

*be the Gaussian with the same mean and covariance as $\mu$. Then*

$$\mathcal{W}_2^2(\mathcal{R}\mu, \mathcal{R}\nu) = \|\mathbf{m}_\mu - \mathbf{m}_\nu\|^2 + \mathcal{B}^2(\mathbf{\Sigma}_\mu, \mathbf{\Sigma}_\nu) \leqslant \mathcal{W}_2^2(\mu, \nu).$$

*Proof.* Take any coupling $(X, Y)$ of $\mu$ and $\nu$, center the variables, and write $C = \mathbb{E}[(X - \mathbf{m}_\mu)(Y - \mathbf{m}_\nu)^\top]$. In the positive definite case, positivity of the block covariance matrix implies the factorization $C = \mathbf{\Sigma}_\mu^{1/2} K \mathbf{\Sigma}_\nu^{1/2}$ with $\|K\|_{\mathrm{op}} \leqslant 1$, and therefore, by operator/nuclear norm duality,

$$\operatorname{tr} C \leqslant \operatorname{tr}\left((\mathbf{\Sigma}_\mu^{1/2}\mathbf{\Sigma}_\nu\mathbf{\Sigma}_\mu^{1/2})^{1/2}\right).$$

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

$$\mathbb{E}\|X - Y\|^2 \geqslant \|\mathbf{m}_\mu - \mathbf{m}_\nu\|^2 + \mathcal{B}^2(\mathbf{\Sigma}_\mu, \mathbf{\Sigma}_\nu).$$

Taking the infimum over couplings proves the inequality, while equality for Gaussian laws is Proposition 2.39. □

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 14.19** (Hugo Lavenant Gaussian-preservation criterion)**.** *Let $F : \mathcal{P}_2(\mathbb{R}^d) \to (-\infty, +\infty]$ satisfy*

$$F(\mathcal{R}\mu) \leqslant F(\mu) \qquad \forall \mu \in \mathcal{P}_2(\mathbb{R}^d),$$

*with $\mathcal{R}$ defined in Theorem 14.18. If $\gamma$ is Gaussian and $\nu$ minimizes the JKO step*

$$\eta \mapsto F(\eta) + \frac{1}{2\tau}\mathcal{W}_2^2(\gamma, \eta),$$

*then $\mathcal{R}\nu$ is also a minimizer. If the JKO minimizer is unique, it is Gaussian. Thus any unique Wasserstein gradient flow obtained as the limit of this JKO scheme preserves Gaussian initial data.*

*Proof.* For the JKO claim, $\mathcal{R}\gamma = \gamma$ because $\gamma$ is Gaussian. Hence, for any competitor $\eta$,

$$F(\mathcal{R}\eta) + \frac{1}{2\tau}\mathcal{W}_2^2(\gamma, \mathcal{R}\eta) \leqslant F(\eta) + \frac{1}{2\tau}\mathcal{W}_2^2(\gamma, \eta).$$

Applying this to a minimizer $\eta = \nu$ shows that $\mathcal{R}\nu$ is again a minimizer. Uniqueness forces $\nu = \mathcal{R}\nu$. □

> **Remark 14.20 (Gaussian barycenters from contraction).** The same projection argument also explains why quadratic Wasserstein barycenters of Gaussian measures are Gaussian. If $\beta_s$ are Gaussian and
>
> $$F(\mu) = \sum_s \lambda_s \mathcal{W}_2^2(\mu, \beta_s),$$
>
> then $\mathcal{R}\beta_s = \beta_s$, and Theorem 14.18 gives $F(\mathcal{R}\mu) \leqslant F(\mu)$. Thus the moment-matched Gaussian projection of any barycenter is again a barycenter; when the barycenter is unique, it must itself be Gaussian. This is the contraction viewpoint behind Corollary 10.9.

# 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.

## Acknowledgements

This work was supported by the European Research Council (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).

# Bibliography


[1] Martial Agueh and Guillaume Carlier. Barycenters in the Wasserstein space. *SIAM Journal on Mathematical Analysis*, 43(2):904–924, 2011.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient Flows in Metric Spaces and in the Space of Probability Measures*. Springer, 2006.

[8] 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.

[9] Francisco Andrade, Gabriel Peyré, and Clarice Poon. Sparsistency for inverse optimal transport. In *International Conference on Learning Representations*, 2024.

[10] 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.

[11] 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.

[12] 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.

[13] Franz Aurenhammer. Power diagrams: properties, algorithms and applications. *SIAM Journal on Computing*, 16(1):78–96, 1987.

[14] Franz Aurenhammer, Friedrich Hoffmann, and Boris Aronov. Minkowski-type theorems and least-squares clustering. *Algorithmica*, 20(1):61–76, 1998.

[15] Julio Backhoff Veraguas, Mathias Beiglböck, and Gudmund Pammer. Existence, duality, and cyclical monotonicity for weak transport costs. *arXiv preprint arXiv:1809.05893*, 2018.

[16] Martin Beckmann. A continuous model of transportation. *Econometrica*, 20:643–660, 1952.

[17] Aurélien Bellet, Amaury Habrard, and Marc Sebban. Metric learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 9(1):1–151, 2015.

[18] 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.

[19] 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.

[20] 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.

[21] Jean-David Benamou, Brittany D Froese, and Adam M Oberman. Numerical solution of the optimal transportation problem using the Monge–Ampère equation. *Journal of Computational Physics*, 260(1):107–126, 2014.

[22] Christian Berg, Jens Peter Reus Christensen, and Paul Ressel. *Harmonic Analysis on Semigroups*. Number 100 in Graduate Texts in Mathematics. Springer Verlag, 1984.

[23] Alain Berlinet and Christine Thomas-Agnan. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Kluwer Academic Publishers, 2003.

[24] 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.

[25] Dimitri P Bertsekas. A new algorithm for the assignment problem. *Mathematical Programming*, 21(1):152–171, 1981.

[26] Dimitri P Bertsekas. Auction algorithms for network flow problems: a tutorial introduction. *Computational Optimization and Applications*, 1(1):7–66, 1992.

[27] Dimitri P Bertsekas and Jonathan Eckstein. Dual coordinate step methods for linear network flow problems. *Mathematical Programming*, 42(1):203–243, 1988.

[28] Dimitris Bertsimas and John N Tsitsiklis. *Introduction to Linear Optimization*. Athena Scientific, 1997.

[29] Rajendra Bhatia, Tanvi Jain, and Yongdo Lim. On the Bures–Wasserstein distance between positive definite matrices. *Expositiones Mathematicae*, 37(2):165–191, 2019.

[30] 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.

[31] Garrett Birkhoff. Tres observaciones sobre el algebra lineal. *Universidad Nacional de Tucumán Revista Series A*, 5:147–151, 1946.

[32] Garrett Birkhoff. Extensions of jentzsch's theorem. *Transactions of the American Mathematical Society*, 85(1):219–227, 1957.

[33] Jose Blanchet and Karthyek Murthy. Quantifying distributional model risk via optimal transport. *Mathematics of Operations Research*, 44(2):565–600, 2019.

[34] 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.

[35] Vladimir I. Bogachev. *Measure Theory*. Springer, Berlin, 2007.

[36] Léa Bohbot, Cyril Letrouit, Gabriel Peyré, and Franccois-Xavier Vialard. Token sample complexity of attention. *arXiv preprint arXiv:2512.10656*, 2025.

[37] Emmanuel Boissard, Thibaut Le Gouic, and Jean-Michel Loubes. Distribution's template estimate with Wasserstein metrics. *Bernoulli*, 21(2):740–759, 2015.

[38] 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.

[39] Nicolas Bonneel and Julie Digne. A survey of optimal transport for computer graphics and computer vision. *Computer Graphics Forum*, 42(2):439–460, 2023.

[40] 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.

[41] Nicolas Bonnotte. *Unidimensional and evolution methods for optimal transportation*. PhD thesis, Université Paris-Sud, 2013.

[42] 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.

[43] Yann Brenier. Décomposition polaire et réarrangement monotone des champs de vecteurs. *C. R. Acad. Sci. Paris Sér. I Math.*, 305(19):805–808, 1987.

[44] Yann Brenier. Polar factorization and monotone rearrangement of vector-valued functions. *Communications on Pure and Applied Mathematics*, 44(4):375–417, 1991.

[45] 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.

[46] 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.

[47] Luis Caffarelli. The Monge-Ampere equation and optimal transportation, an elementary review. *Lecture Notes in Mathematics, Springer-Verlag*, pages 1–10, 2003.

[48] 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.

[49] 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.

[50] Guillaume Carlier, Victor Chernozhukov, and Alfred Galichon. Vector quantile regression beyond the specified case. *Journal of Multivariate Analysis*, 161:96–102, 2017.

[51] Guillaume Carlier and Ivar Ekeland. Matching for teams. *Economic Theory*, 42(2):397–418, 2010.

[52] 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.

[53] 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.

[54] 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.

[55] Shouvanik Chakrabarti, Yiming Huang, Tongyang Li, Soheil Feizi, and Xiaodi Wu. Quantum Wasserstein generative adversarial networks. *Advances in Neural Information Processing Systems*, 32, 2019.

[56] Timothy M Chan. Optimal output-sensitive convex hull algorithms in two and three dimensions. *Discrete & Computational Geometry*, 16(4):361–368, 1996.

[57] Louis H. Y. Chen, Larry Goldstein, and Qi-Man Shao. *Normal Approximation by Stein's Method*. Springer, 2011.

[58] 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.

[59] 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.

[60] 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.

[61] Yongxin Chen, Tryphon T Georgiou, and Allen Tannenbaum. Matrix optimal mass transport: a quantum mechanical approach. *arXiv preprint arXiv:1610.03041*, 2016.

[62] Sinho Chewi, Jonathan Niles-Weed, and Philippe Rigollet. Statistical optimal transport. *arXiv preprint arXiv:2407.18163*, 2024.

[63] 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.

[64] Lénaïc Chizat, Alex Delalande, and Tomas Vaškevičius. Sharper exponential convergence rates for Sinkhorn's algorithm in continuous settings. *arXiv preprint arXiv:2407.01202*, 2024.

[65] 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.

[66] 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.

[67] Imre Ciszár. Information-type measures of difference of probability distributions and indirect observations. *Studia Scientiarum Mathematicarum Hungarica*, 2:299–318, 1967.

[68] 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.

[69] 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.

[70] Marco Cuturi. Sinkhorn distances: lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems 26*, pages 2292–2300, 2013.

[71] Marco Cuturi and David Avis. Ground metric learning. *Journal of Machine Learning Research*, 15:533–564, 2014.

[72] 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.

[73] Marco Cuturi and Gabriel Peyré. A smoothed dual approach for variational Wasserstein problems. *SIAM Journal on Imaging Sciences*, 9(1):320–343, 2016.

[74] 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.

[75] George B Dantzig. Application of the simplex method to a transportation problem. *Activity Analysis of Production and Allocation*, 13:359–373, 1951.

[76] 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.

[77] 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.

[78] Mingyang Deng, He Li, Tianhong Li, Yilun Du, and Kaiming He. Generative modeling via drifting. *arXiv preprint arXiv:2602.04770*, 2026.

[79] 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.

[80] 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.

[81] 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.

[82] Richard M. Dudley. The speed of mean Glivenko-Cantelli convergence. *Annals of Mathematical Statistics*, 40(1):40–50, 1969.

[83] 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.

[84] 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.

[85] 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.

[86] 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.

[87] 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.

[88] 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.

[89] 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.

[90] 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.

[91] Robert Fortet. Résolution d'un système d'équations de M. Schrödinger. *Journal de Mathématiques Pures et Appliquées*, 19(1–4):83–105, 1940.

[92] 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.

[93] Joel Franklin and Jens Lorenz. On the scaling of multidimensional matrices. *Linear Algebra and its Applications*, 114:717–735, 1989.

[94] 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.

[95] 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.

[96] Alfred Galichon. *Optimal Transport Methods in Economics*. Princeton University Press, 2016.

[97] Alfred Galichon and Bernard Salanié. Matching with trade-offs: revealed preferences over competing characteristics. Technical report, Preprint SSRN-1487307, 2009.

[98] 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.

[99] Wilfrid Gangbo and Robert J McCann. The geometry of optimal transportation. *Acta Mathematica*, 177(2):113–161, 1996.

[100] Rui Gao and Anton J. Kleywegt. Distributionally robust stochastic optimization with Wasserstein distance. *Mathematics of Operations Research*, 48(2):603–655, 2023.

[101] Ankit Garg and Rafael Oliveira. Recent progress on scaling algorithms and applications. *arXiv preprint arXiv:1808.09669*, 2018.

[102] 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.

[103] 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.

[104] 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.

[105] Tryphon T Georgiou and Michele Pavon. Positive contraction mappings for classical and quantum Schrödinger systems. *Journal of Mathematical Physics*, 56(3):033301, 2015.

[106] Borjan Geshkovski, Cyril Letrouit, Yury Polyanskiy, and Philippe Rigollet. A mathematical perspective on transformers. *arXiv preprint arXiv:2312.10794*, 2023.

[107] 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.

[108] Franccois Golse, Emanuele Caglioti, and Thierry Paul. Quantum optimal transport is cheaper. *arXiv preprint arXiv:1908.01829*, 2019.

[109] 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.

[110] 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.

[111] Siegfried Graf and Harald Luschgy. *Foundations of Quantization for Probability Distributions*, volume 1730 of *Lecture Notes in Mathematics*. Springer, 2000.

[112] 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.

[113] 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.

[114] Mikhail Gromov. *Metric Structures for Riemannian and Non-Riemannian Spaces*. Progress in Mathematics. Birkhäuser, 2001.

[115] 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.

[116] Leonid Gurvits. Classical complexity and quantum entanglement. *Journal of Computer and System Sciences*, 69(3):448–484, 2004.

[117] 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.

[118] 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.

[119] Mathis Hardion and Théo Lacombe. The wasserstein gradient flow of the sinkhorn divergence between gaussian distributions. *arXiv preprint arXiv:2602.10726*, 2026.

[120] Ping He, Om Khangaonkar, Hamed Pirsiavash, Yikun Bai, and Soheil Kolouri. Sinkhorn-drifting generative models. *arXiv preprint arXiv:2603.12366*, 2026.

[121] 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.

[122] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851, 2020.

[123] Thomas Hofmann, Bernhard Schölkopf, and Alexander J Smola. Kernel methods in machine learning. *Annals of Statistics*, 36(3):1171–1220, 2008.

[124] Aapo Hyvärinen. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6:695–709, 2005.

[125] Martin Idel. A review of matrix scaling and Sinkhorn's normal form for matrices and positive maps. *arXiv preprint arXiv:1609.06349*, 2016.

[126] 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*, 2020.

[127] 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.

[128] Leonid Kantorovich. On the transfer of masses (in russian). *Doklady Akademii Nauk*, 37(2):227–229, 1942.

[129] LV Kantorovich and G.S. Rubinstein. On a space of totally additive functions. *Vestn Leningrad Universitet*, 13:52–59, 1958.

[130] Johan Karlsson and Axel Ringh. Generalized Sinkhorn iterations for regularizing inverse problems using optimal mass transport. *arXiv preprint arXiv:1612.02273*, 2016.

[131] 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.

[132] Philip A Knight. The Sinkhorn–Knopp algorithm: convergence and applications. *SIAM Journal on Matrix Analysis and Applications*, 30(1):261–275, 2008.

[133] Herbert Knothe. Contributions to the theory of convex bodies. *Michigan Mathematical Journal*, 4(1):39–52, 1957.

[134] Soheil Kolouri, Yang Zou, and Gustavo K Rohde. Sliced Wasserstein kernels for probability distributions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5258–5267, 2016.

[135] J. Kruithof. Telefoonverkeersrekening. *De Ingenieur*, 52:E15–E25, 1937.

[136] Harold W. Kuhn. The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1–2):83–97, 1955.

[137] Brian Kulis. Metric learning: a survey. *Foundations and Trends in Machine Learning*, 5(4):287–364, 2012.

[138] Hugo Lavenant and Filippo Santambrogio. The flow map of the Fokker–Planck equation does not provide optimal transport. *Applied Mathematics Letters*, 133:108225, 2022.

[139] Thibaut Le Gouic and Jean-Michel Loubes. Existence and consistency of Wasserstein barycenters. *Probability Theory and Related Fields*, 168:901–917, 2016.

[140] 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.

[141] Bas Lemmens and Roger Nussbaum. *Nonlinear Perron-Frobenius Theory*, volume 189. Cambridge University Press, 2012.

[142] Christian Léonard. From the Schrödinger problem to the Monge–Kantorovich problem. *Journal of Functional Analysis*, 262(4):1879–1920, 2012.

[143] 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.

[144] Christian Léonard. Revisiting Fortet's proof of existence of a solution to the Schrödinger system. *arXiv preprint arXiv:1904.13211*, 2019.

[145] 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.

[146] 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.

[147] 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.

[148] 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.

[149] Stuart Lloyd. Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28(2):129–137, 1982.

[150] 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.

[151] John Lott and Cédric Villani. Ricci curvature for metric-measure spaces via optimal transport. *Annals of Mathematics*, 169(3):903–991, 2009.

[152] 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.

[153] Shaojun Ma, Haodong Sun, Xiaojing Ye, Hongyuan Zha, and Haomin Zhou. Learning cost functions for optimal transport. *arXiv preprint arXiv:2002.09650*, 2020.

[154] Jan Maas. Gradient flows of the entropy for finite Markov chains. *Journal of Functional Analysis*, 261(8):2250–2292, 2011.

[155] 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.

[156] Jan Maas, Martin Rumpf, and Stefan Simon. Generalized optimal transport with singular sources. *arXiv preprint arXiv:1607.01186*, 2016.

[157] Robert J McCann. A convexity principle for interacting gases. *Advances in Mathematics*, 128(1):153–179, 1997.

[158] 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.

[159] Facundo Mémoli. On the use of Gromov–Hausdorff distances for shape comparison. In *Symposium on Point Based Graphics*, pages 81–90. Eurographics Association, 2007.

[160] Facundo Mémoli. Gromov–Wasserstein distances and the metric approach to object matching. *Foundations of Computational Mathematics*, 11(4):417–487, 2011.

[161] Quentin Mérigot. A multiscale approach to optimal transport. *Computer Graphics Forum*, 30(5):1583–1592, 2011.

[162] Quentin Mérigot. A comparison of two dual methods for discrete optimal transport. In *Geometric science of information*, pages 389–396. Springer, 2013.

[163] 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.

[164] Quentin Mérigot and Boris Thibert. Optimal transport: discretization and algorithms. *arXiv preprint arXiv:2003.00855*, 2020.

[165] 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.

[166] 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.

[167] Gaspard Monge. Mémoire sur la théorie des déblais et des remblais. *Histoire de l'Académie Royale des Sciences*, pages 666–704, 1781.

[168] Eduardo Fernandes Montesuma, Fred Ngolè Mboula, and Antoine Souloumiac. Recent advances in optimal transport for machine learning. *arXiv preprint arXiv:2306.16156*, 2023.

[169] 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.

[170] 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.

[171] Arkadi Nemirovski and Uriel Rothblum. On complexity of matrix scaling. *Linear Algebra and its Applications*, 302:435–460, 1999.

[172] Yurii Nesterov and Arkadii Nemirovskii. *Interior-point polynomial algorithms in convex programming*, volume 13. SIAM, 1994.

[173] 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.

[174] 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.

[175] James B. Orlin. A polynomial time primal network simplex algorithm for minimum cost flows. *Mathematical Programming*, 78(2):109–129, 1997.

[176] Felix Otto. The geometry of dissipative evolution equations: the porous medium equation. *Communications in Partial Differential Equations*, 26(1–2):101–174, 2001.

[177] Andrea Ottolini and Stefan Steinerberger. Greedy matching in optimal transport with concave cost. *arXiv preprint arXiv:2307.03140*, 2023.

[178] Nicolas Papadakis, Gabriel Peyré, and Edouard Oudet. Optimal transport with proximal splitting. *SIAM Journal on Imaging Sciences*, 7(1):212–238, 2014.

[179] 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.

[180] Gabriel Peyré. Entropic approximation of Wasserstein gradient flows. *SIAM Journal on Imaging Sciences*, 8(4):2323–2351, 2015.

[181] Gabriel Peyré. Optimal and diffusion transports in machine learning. *arXiv preprint arXiv:2512.06797*, 2025. Proc. 2026 International Congress of Mathematicians.

[182] Gabriel Peyré. Muon dynamics as a spectral Wasserstein flow. *arXiv preprint arXiv:2604.04891*, 2026.

[183] Gabriel Peyré. Robust sublinear convergence rates for iterative Bregman projections. *arXiv preprint arXiv:2602.01372*, 2026.

[184] 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.

[185] Gabriel Peyré and Marco Cuturi. Computational optimal transport with applications to data sciences. *Foundations and Trends in Machine Learning*, 11(5-6):355–607, 2019.

[186] 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.

[187] 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.

[188] 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.

[189] 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.

[190] Svetlozar T Rachev and Ludger Rüschendorf. *Mass Transportation Problems: Volume I: Theory*. Springer, 1998.

[191] Svetlozar T Rachev and Ludger Rüschendorf. *Mass Transportation Problems: Volume II: Applications*. Springer, 1998.

[192] Aaditya Ramdas, Nicolás García Trillos, and Marco Cuturi. On Wasserstein two-sample testing and related families of nonparametric tests. *Entropy*, 19(2):47, 2017.

[193] Erik Reinhard, Michael Adhikhmin, Bruce Gooch, and Peter Shirley. Color transfer between images. *IEEE Computer Graphics and Applications*, 21(5):34–41, 2001.

[194] Emmanuel Rio. Asymptotic constants for minimal distance in the central limit theorem. *Electronic Communications in Probability*, 16:96–103, 2011.

[195] Ralph Tyrell Rockafellar. *Convex Analysis*. Princeton university press, 2015.

[196] Murray Rosenblatt. Remarks on a multivariate transformation. *The Annals of Mathematical Statistics*, 23(3):470–472, 1952.

[197] Walter Rudin. *Real and Complex Analysis*. McGraw–Hill, third edition, 1987.

[198] Ludger Ruschendorf. Convergence of the iterative proportional fitting procedure. *Annals of Statistics*, 23(4):1160–1174, 1995.

[199] 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.

[200] Hans Samelson. On the perron-frobenius theorem. *Michigan Mathematical Journal*, 4(1):57–59, 1957.

[201] 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.

[202] Filippo Santambrogio. *Optimal Transport for Applied Mathematicians: Calculus of Variations, PDEs, and Modeling*. Birkhäuser, 2015.

[203] 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.

[204] Isaac J Schoenberg. Metric spaces and positive definite functions. *Transactions of the American Mathematical Society*, 44(3):522–536, 1938.

[205] Bernhard Schölkopf and Alexander J Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, 2002.

[206] Erwin Schrödinger. Über die Umkehrung der Naturgesetze. *Sitzungsberichte Preuss. Akad. Wiss. Berlin. Phys. Math.*, 144:144–153, 1931.

[207] 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.

[208] Aman Sinha, Hongseok Namkoong, and John Duchi. Certifying some distributional robustness with principled adversarial training. In *International Conference on Learning Representations*, 2018.

[209] Richard Sinkhorn. A relationship between arbitrary positive matrices and doubly stochastic matrices. *The Annals of Mathematical Statistics*, 35(2):876–879, 1964.

[210] Richard Sinkhorn. Diagonal equivalence to matrices with prescribed row and column sums. *American Mathematical Monthly*, 74:402–405, 1967.

[211] Richard Sinkhorn and Paul Knopp. Concerning nonnegative matrices and doubly stochastic matrices. *Pacific Journal of Mathematics*, 21(2):343–348, 1967.

[212] 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.

[213] 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.

[214] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

[215] 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.

[216] Bharath K Sriperumbudur, Kenji Fukumizu, Arthur Gretton, Bernhard Schölkopf, and Gert RG Lanckriet. On integral probability metrics, $\varphi$-divergences and binary classification. *arXiv preprint arXiv:0901.2698*, 2009.

[217] 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.

[218] 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.

[219] Karl-Theodor Sturm. On the geometry of metric measure spaces. I. *Acta Mathematica*, 196(1):65–131, 2006.

[220] Karl-Theodor Sturm. On the geometry of metric measure spaces. II. *Acta Mathematica*, 196(1):133–177, 2006.

[221] Karl-Theodor Sturm. The space of spaces: curvature bounds and gradient flows on the space of metric measure spaces. Preprint 1208.0434, arXiv, 2012.

[222] Gábor J Székely and Maria L Rizzo. Testing for equal distributions in high dimension. *InterStat*, 5(16.10), 2004.

[223] 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.

[224] 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.

[225] Cédric Villani. *Topics in Optimal Transportation*, volume 58 of *Graduate Studies in Mathematics Series*. American Mathematical Society, 2003.

[226] Cédric Villani. *Optimal Transport: Old and New*, volume 338. Springer, 2009.

[227] Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural Computation*, 23(7):1661–1674, 2011.

[228] 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.

[229] 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.

[230] James Vuckovic, Aristide Baratin, and Rémi Tachet des Combes. A mathematical theory of attention. *arXiv preprint arXiv:2007.02876*, 2020.

[231] 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.

[232] 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.

[233] Holger Wendland. *Scattered Data Approximation*. Cambridge University Press, 2005.

[234] Huan Xu and Shie Mannor. Distributionally robust markov decision processes. *Mathematics of Operations Research*, 37(2):288–300, 2012.

[235] Insoon Yang. A convex optimization approach to distributionally robust markov decision processes with Wasserstein distance. *IEEE Control Systems Letters*, 1(1):164–169, 2017.

[236] G Udny Yule. On the methods of measuring association between two attributes. *Journal of the Royal Statistical Society*, 75(6):579–652, 1912.

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.

| 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 |
| $\Sigma_n$ | Probability simplex of histograms of length $n$. | Definition 2.1 |
| $\delta_x$ | Dirac mass at $x$. | Definition 2.2 |
| $\alpha, \beta, \gamma$ | Source, target and auxiliary probability measures. | Eq. (2.5) |
| $\rho_\alpha$ | Density of $\alpha$ with respect to a reference measure. | Definition 2.5 |
| $\mathrm{d}\alpha, \mathrm{d}x$ | Integration against $\alpha$ 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. (6.8) |
| $\mathcal{C}(\mathcal{X})$ | 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) |
| $P_\sigma, \mathcal{P}_n^{\mathrm{perm}}$ | Permutation matrix and the set of all such matrices. | Definition 3.7 |
| $\mathcal{B}_n$ | Birkhoff polytope of bistochastic matrices. | Definition 3.8 |
| $\mathrm{a}, \mathrm{b}$ | Discrete probability histograms. | Eq. (3.1) |
| $\mathrm{P}$ | 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) |
| $\mathrm{D}$ | Ground distance matrix for discrete Wasserstein distances. | Definition 3.30 |
| $\mathrm{W}_p(\mathrm{a}, \mathrm{b})$ | Discrete $p$-Wasserstein distance. | Definition 3.30 |
| **Monge maps, one-dimensional OT and Gaussians** | | |
| $T, T$ | Transport map. | Eq. (2.5) |
| $T_\sharp \alpha$ | Push-forward of $\alpha$ by $T$. | Definition 2.8 |
| $T^\sharp g$ | Pullback of a test function, $T^\sharp g = g \circ T$. | Remark 2.9 |
| $\mathrm{Id}$ | Identity map. | Definition 2.8 |
| $\tilde{\mathcal{W}}_p$ | Directed Monge transport distance. | Eq. (2.6) |
| $\nabla\varphi$ | Brenier map for quadratic cost. | Theorem 2.18 |
| $C_\alpha$ | Cumulative distribution function of a 1-D measure. | Eq. (2.8) |

| Notation | Meaning | First reference |
|---|---|---|
| $C_\alpha^{-1}$ | Quantile function of a 1-D measure. | Eq. (2.9) |
| $\mathcal{N}(\mathbf{m}, \mathbf{\Sigma})$ | Gaussian law with mean $\mathbf{m}$ and covariance $\mathbf{\Sigma}$. | Eq. (2.14) |
| $\mathbf{m}_\alpha, \mathbf{\Sigma}_\alpha$ | Mean and covariance of a Gaussian measure $\alpha$. | Eq. (2.18) |
| $\mathcal{B}(\mathbf{\Sigma}_\alpha, \mathbf{\Sigma}_\beta)$ | Bures covariance distance. | Definition 2.38 |
| $\mathrm{tr}(\mathbf{\Sigma})$ | Trace of a matrix. | Eq. (2.18) |
| **Continuous Kantorovich OT and Wasserstein distances** | | |
| $\pi$ | Coupling or transport plan. | Definition 3.17 |
| $\mathcal{U}(\alpha, \beta)$ | Set of couplings between $\alpha$ and $\beta$. | Eq. (3.4) |
| $\mathcal{L}_c(\alpha, \beta)$ | Kantorovich optimal value with ground cost $c$. | Eq. (3.5) |
| $d$ | Ground distance on the underlying metric space. | Eq. (3.8) |
| $\mathcal{W}_p(\alpha, \beta)$ | $p$-Wasserstein distance. | Definition 3.33 |
| $\mathcal{W}_\infty(\alpha, \beta)$ | Worst-displacement Wasserstein distance. | Eq. (3.13) |
| $\mathfrak{A}, \mathfrak{B}$ | Probability laws over probability measures. | Eq. (3.9) |
| $\bar{\alpha}_\mathfrak{A}$ | Collapsed mixture associated with a law over measures. | Definition 3.47 |
| $\mathbb{W}_2$ | Wasserstein distance on the Wasserstein space. | Eq. (3.11) |
| $\Gamma$ | $c$-cyclically monotone subset of $\mathcal{X} \times \mathcal{Y}$. | Definition 3.27 |
| $\rho$ | Glued or composed coupling. | Lemma 3.32 |
| $\rightharpoonup$ | Weak* convergence of measures. | Definition 3.38 |
| $\mathrm{TV}, \|\cdot\|_{\mathrm{TV}}$ | Total variation divergence/norm. | Section 2.1 |
| **Duality, transforms and weak norms** | | |
| $\mathrm{f}, \mathrm{g}$ | Discrete dual potentials. | Eq. (4.2) |
| $f, g$ | Continuous dual potentials. | Eq. (4.5) |
| $\mathrm{R}(\mathrm{a}, \mathrm{b})$ | Feasible set of discrete dual potentials. | Eq. (4.1) |
| $\mathcal{R}(c)$ | Feasible set of continuous dual potentials. | Eq. (4.6) |
| $f^c, g^c$ | $c$-transform of a potential. | Definition 4.8 |
| $\mathbb{L}_j(\mathrm{g})$ | Laguerre/power cell in semi-discrete OT. | Eq. (5.4) |
| $Q_m(\alpha)$ | Optimal $m$-point quantization error. | Eq. (5.9) |
| $\mathrm{Lip}(f)$ | Lipschitz constant of $f$. | Eq. (5.10) |
| $\mathcal{W}_1$ | Kantorovich–Rubinstein distance/norm. | Eq. (5.11) |
| $\mathcal{W}_{1,G}$ | Graph Wasserstein-1/transshipment distance. | Proposition 5.8 |
| $d_G, \nabla_G, \mathrm{div}_G$ | Graph geodesic distance, gradient and divergence. | Proposition 5.8 |
| $\|\cdot\|_B$ | Dual norm induced by a discriminator class $B$. | Eq. (6.1) |
| $\mathcal{H}, k$ | Reproducing kernel Hilbert space and its kernel. | Definition 6.9 |
| $\mathrm{MMD}_k$ | Maximum mean discrepancy/kernel norm for $k$. | Definition 6.9 |
| $\mathcal{D}_\varphi, \mathrm{D}_\varphi$ | Continuous and discrete $\varphi$-divergences. | Eq. (6.6) |
| $\varphi'_\infty$ | Recession slope of an entropy function. | Definition 6.13 |
| $\varphi^\star$ | Legendre transform of $\varphi$. | Eq. (6.13) |
| $\mathrm{KL}, \mathrm{KL}$ | Continuous and discrete Kullback–Leibler divergences. | Definitions 7.5, 7.9 |
| $\mathfrak{h}$ | Hellinger divergence/distance. | Section 6.3 |
| JS | Jensen–Shannon divergence. | Section 6.3 |
| **Entropic regularization and Sinkhorn algorithms** | | |
| $\varepsilon$ | Entropic regularization strength. | Eq. (7.1) |
| $\mathrm{H}(\mathrm{P})$ | Shannon–Boltzmann entropy of a matrix. | Definition 7.1 |
| $\mathrm{L}_\mathrm{C}^\varepsilon(\mathrm{a}, \mathrm{b})$ | Discrete entropic OT value. | Eq. (7.1) |
| $\mathcal{L}_c^\varepsilon(\alpha, \beta)$ | Continuous entropic OT value. | Eq. (7.11) |
| $\mathrm{K}$ | Gibbs kernel $e^{-\mathrm{C}/\varepsilon}$. | Eq. (7.2) |
| $\mathrm{u}, \mathrm{v}$ | Left and right Sinkhorn scalings. | Eq. (7.2) |
| $\mathrm{diag}(\mathrm{u})\mathrm{K}\,\mathrm{diag}(\mathrm{v})$ | Scaling form of the entropic coupling. | Eq. (8.1) |
| $\odot$ | Entrywise product of vectors. | Eq. (7.4) |
| $\mathrm{u}^{(\ell)}, \mathrm{u}^{(\ell+1)}$ | Current and next Sinkhorn iterates. | Eq. (7.5) |
| $d_\mathcal{H}$ | Hilbert projective metric on positive vectors. | Definition 8.13 |

| Notation | Meaning | First reference |
|---|---|---|
| $\mathrm{Proj}^{\mathrm{KL}}$ | KL/Bregman projection. | Eq. (8.1) |
| $\bar{\mathcal{L}}_c^{\varepsilon}(\alpha,\beta)$ | Debiased Sinkhorn divergence. | Eq. (7.31) |
| **Extensions of OT** | | |
| $\psi_1, \psi_2$ | Entropy functions penalizing marginal mismatch. | Eq. (9.1) |
| $\mathrm{UW}_c, \mathrm{UW}_{c,\tau}$ | Relaxed unbalanced OT value with marginal penalties. | Eq. (9.1) |
| $L_c$ | Reverse-formulation local unbalanced cost. | Eq. (9.2) |
| $H_c$ | Homogeneous perspective of the local cost $L_c$. | Eq. (9.3) |
| HW | Homogeneous unbalanced formulation. | Eq. (9.4) |
| $\mathfrak{C}[\mathcal{X}]$ | Cone over the metric space $\mathcal{X}$. | Section 9.1 |
| $\mathrm{CW}, \mathrm{CW}_\kappa$ | Cone formulation of unbalanced OT, with $\mathrm{CW}_\kappa$ using growth scale $\kappa$. | Theorem 9.4, Eq. (12.11) |
| $\mathcal{A}_\kappa$ | Dynamic unbalanced perspective action for transport and growth. | Eq. (12.11) |
| $\mathrm{WFR}_\kappa$ | Wasserstein–Fisher–Rao dynamic distance with growth scale $\kappa$. | Eq. (12.11) |
| $\beta_s, \lambda_s$ | Input measures and weights in barycenter problems. | Eq. (10.2) |
| $\alpha^\star$ | Optimal measure, often a barycenter. | Section 10.1 |
| $\mathrm{SW}_p$ | Sliced Wasserstein distance. | Definition 9.5 |
| $\mathbb{S}^{d-1}$ | Unit sphere of projection directions. | Definition 9.5 |
| $P_\theta$ | Projection on direction $\theta$. | Definition 9.5 |
| $\mathrm{MaxSW}_p$ | Max-sliced Wasserstein distance. | Definition 9.8 |
| $\mathrm{MSWGG}_2$ | Min-sliced lifted-plan upper bound on $\mathcal{W}_2$. | Section 9.2 |
| $\mathrm{SW}_{p,k}, \mathrm{MaxSW}_{p,k}$ | Average and max Wasserstein distances over $k$-dimensional projections. | Definition 9.9 |
| $\mathcal{W}_\gamma$ | Spectral Wasserstein distance associated with a matrix gauge $\gamma$. | Eq. (9.7) |
| $\mathcal{B}_\gamma$ | Polar set defining the robust projected form of $\mathcal{W}_\gamma$. | Eq. (9.9) |
| $\mathcal{W}_{2,A}$ | Quadratic Wasserstein pseudodistance after projection by $A^{1/2}$. | Eq. (9.8) |
| $\mathrm{SRW}_{2,k}$ | Paty–Cuturi subspace robust Wasserstein distance. | Section 9.4 |
| $\mathrm{LOT}_\rho$ | Linear OT distance around reference $\rho$. | Eq. (9.6) |
| $\bar{T}_\pi$ | Barycentric projection of a coupling $\pi$. | Eq. (10.12) |
| $\bar{\beta}_\pi$ | Pushforward of $\alpha$ by the barycentric projection. | Eq. (10.12) |
| $\mathrm{WOT}_C$ | Weak OT value with conditional-law cost $C$. | Eq. (10.13) |
| $g^C$ | Weak $C$-transform in weak OT duality. | Proposition 10.14 |
| $u_t, V_t$ | Positive vector-valued density and spatial flux. | Eqs. (11.1), (11.2) |
| $\mathcal{W}_\Phi$ | Dynamic vector-valued BB-type cost. | Eq. (11.1) |
| $\mathrm{D}, \mathrm{D}'$ | Intra-domain distance matrices in discrete GW. | Eq. (11.5) |
| $\Delta$ | Discrepancy between intra-domain distances. | Eq. (11.5) |
| GW | Discrete Gromov–Wasserstein cost. | Eq. (11.5) |
| $\mathbb{X}, \mathbb{Y}$ | Metric-measure spaces. | Definition 11.6 |
| $\mathcal{GW}$ | Continuous Gromov–Wasserstein distance. | Eq. (11.6) |
| $d_{\mathrm{H}}, d_{\mathrm{GH}}$ | Hausdorff and Gromov–Hausdorff distances. | Section 11.2 |
| $\mathrm{FGW}_{\lambda,p}$ | Fused Gromov–Wasserstein distance. | Section 11.2 |
| $\mathbb{S}^m, \mathbb{S}^m_+$ | Real symmetric matrices and their positive semidefinite cone. | Definition 11.4 |
| $A_t, P_t$ | Positive matrix-valued density and spatial matrix flux. | Eqs. (11.3), (11.4) |
| $\mathcal{W}_{\mathrm{mat}}$ | Conservative matrix-valued BB-type cost. | Eq. (11.3) |
| $\mathbb{H}_n, \mathbb{H}_n^+, \mathbb{H}_n^{+,1}$ | Hermitian matrices, positive semidefinite Hermitian matrices and density matrices. | Definition 11.12 |
| $\mathrm{Tr}_A, \mathrm{Tr}_B$ | Partial traces of a bipartite matrix. | Eq. (11.9) |

| Notation | Meaning | First reference |
|---|---|---|
| $\mathrm{QOT}_C(A, B)$ | Finite-dimensional quantum OT value with cost observable $C$. | Eq. (11.10) |
| $\mathrm{QOT}_C^{\varepsilon}(A, B)$ | Entropically regularized quantum OT value. | Eq. (11.12) |
| $T_e(F, G), T_s(F, G)$ | Exact Gibbs coupling and symmetric Gurvits-scaling surrogate. | Eqs. (11.14), (11.15) |
| **Dynamic OT and Wasserstein gradient flows** | | |
| $\alpha_t$ | Time-dependent curve of probability measures. | Eq. (12.2) |
| $v_t$ | Eulerian velocity field transporting $\alpha_t$. | Eq. (12.2) |
| $T_t$ | Lagrangian particle flow map. | Eq. (12.1) |
| $P_t$ | Interpolant map in flow matching; later also path evaluation. | Eq. (14.1) |
| $\mathcal{W}_2^2$ via action | Benamou–Brenier dynamic formulation. | Eq. (12.8) |
| $\nabla_{\mathcal{W}} f(\alpha)$ | Wasserstein gradient of a functional. | Proposition 13.2 |
| $\delta f(\alpha)$ | First variation of $f$ at $\alpha$. | Proposition 13.2 |
| $\partial_t \alpha + \mathrm{div}(\alpha v) = 0$ | Continuity equation. | Eq. (12.2) |
| $\alpha_{t+\tau}$ | One JKO/minimizing-movement step. | Eq. (13.1) |
| $\mathcal{S} = C([0, 1]; \mathbb{R}^d)$ | Path space in the superposition formulation. | Chapter 12 |

# Index