Optimal Transport for Machine Learners

Compact teaching notes

Gabriel Peyré

June 13, 2026

1 Optimal Matching between Point Clouds

1.1 Monge Problem for Discrete Points

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 σ within the set $\mathrm{Perm}(n)$ of permutations of n elements that solves

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

1-D Case.

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 σ 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 $s\in\mathbb{R}$ $D(s):=\frac{h(s)-h(s-d)}{d}$ and $\Delta:=h(x-y')+h(x'-y)-h(x-y)-h(x'-y')=d(D(x'-y)-D(x-y))$. Because h is strictly convex, D is strictly increasing. Since $x-y<x'-y$, monotonicity yields $D(x-y)<D(x'-y)$, that is $\Delta>0$. □

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

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

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. □

Rational weights.

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 (μ,ν) is equivalent to the uniform assignment problem obtained by replacing each x_i by k_i identical copies and each y_j by ℓ_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 ℓ_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 $\sum_j n_{ij}=k_i$ and $\sum_i n_{ij}=\ell_j$, and the associated coupling $P_{ij}=n_{ij}/N$ has marginals k_i/N and ℓ_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. □

2D case.

Proposition 1.4 (Non-crossing optimal matchings)**.** *In dimension 2, for $c(x,y)=\|x-y\|$, if σ 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. □

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

1.2 Matching Algorithms

Hungarian primal-dual method. $u_i + v_j \leqslant C_{i,j} \qquad \forall i, j.$

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

Proposition 1.5 (Correctness and complexity of the Hungarian primal-dual method)**.** *Assume the Hungarian method terminates with a perfect matching σ contained in the equality graph $E(u, v) = \{(i, j)\ ;\ u_i + v_j = C_{i,j}\}$, where (u, v) is dual feasible, i.e. $u_i + v_j \leqslant C_{i,j}$ for all (i, j). Then σ 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 τ, 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 σ 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 σ reaches the dual lower bound and is optimal.
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 δ to source labels and subtracting δ from target labels preserves tightness; for edges from S to T^c, the definition of δ 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. □

2 Monge Problem between Measures

2.1 Measures

Histograms.

Definition 2.1 (Probability simplex)**.** The probability simplex of length n is $\Sigma_n \coloneqq \big\{\mathrm{a} \in \mathbb{R}_+^n\ ;\ \sum_{i=1}^n \mathrm{a}_i = 1\big\}$. Its elements are also called probability vectors or histograms.

Discrete measure, empirical measure.

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

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

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

General measures. A Dirac measure δ_x is then defined as $\delta_x(A) = 1$ if $x \in A$ and 0 otherwise, and this extends by linearity for discrete measures of the form (2.1) as
$\alpha(A) = \sum_{x_i \in A} \mathrm{a}_i$ 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.

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.

Definition 2.4 (Support of a measure)**.** The support $\mathrm{supp}(\alpha)$ of a Borel measure α on a metric space $\mathcal{X}$ is the smallest closed set of full α-mass. Equivalently, $x \in \mathrm{supp}(\alpha)$ if every open ball centered at x has positive α-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
$\langle f, \alpha \rangle \coloneqq \int f(x) \mathrm{d}\alpha(x)$. $\int_{\mathcal{X}} f(x) \mathrm{d}\alpha(x) = \sum_{i=1}^n \mathrm{a}_i f(x_i)$.

Relative densities.

Definition 2.5 (Relative density)**.** If α is absolutely continuous with respect to a reference measure λ, 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 $h \in \mathcal{C}(\mathcal{X})$, $\int_{\mathcal{X}} h(x) \mathrm{d}\alpha(x) = \int_{\mathcal{X}} h(x) \rho_\alpha(x) \mathrm{d}\lambda(x)$.

Total variation norm.

Definition 2.6 (Total variation)**.** For a finite signed Radon measure α on a compact space $\mathcal{X}$, $\|\alpha\|_{\mathrm{TV}} \coloneqq \sup_{f \in \mathcal{C}(\mathcal{X})} \{\langle f, \alpha \rangle\ ;\ \|f\|_\infty \leqslant 1\}$.

$|\alpha|(A) \coloneqq \sup_{A = \cup_i B_i} \sum_i |\alpha(B_i)|$, 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 α on a compact space $\mathcal{X}$, $\|\alpha\|_{\mathrm{TV}} = |\alpha|(\mathcal{X})$.*

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 $(\Omega, \mathbb{P})$.
$\alpha(A) = \mathbb{P}(\{\omega \in \Omega \,;\, X(\omega) \in A\})$ for Borel sets $A \subset \mathcal{X}$. $\int_{\mathcal{X}} f(x) \,\mathrm{d}\alpha(x) = \mathbb{E}[f(X)]$.

2.2 Push Forward

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

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

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

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 $T_\sharp \alpha \in \mathcal{M}_+^1(\mathcal{Y})$.

Proposition 2.9 (Push-forward formula for densities)**.** *Let α and β have densities ρ_α and ρ_β 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 $h \in \mathcal{C}(\mathcal{Y})$

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

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

2.3 Monge's Formulation

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

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

Proposition 2.10 (Empirical Monge maps and matchings)**.** *Assume that the source atoms $x_1, \ldots, x_n$ are distinct and that $\alpha = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$, $\beta = \frac{1}{n}\sum_{j=1}^n \delta_{y_j}$. If $T_\sharp \alpha = \beta$, then for each distinct target value z in the support of β, exactly $n\beta(\{z\})$ source atoms are mapped to z. In particular, if the y_j are distinct, then there exists a permutation $\sigma \in \mathrm{Perm}(n)$ such that $T(x_i) = y_{\sigma(i)}$ for all i. Conversely, every such assignment of source atoms to target atoms with the correct masses defines a feasible Monge map on the support of α, and in the distinct-target case $\int_{\mathcal{X}} c(x, T(x)) \mathrm{d}\alpha(x) = \frac{1}{n}\sum_{i=1}^n c(x_i, y_{\sigma(i)})$. If source locations are repeated, they should first be merged into atoms with larger masses; such atoms cannot be split by a Monge map.*

Proof. Since $T_\sharp \alpha = \beta$, all images $T(x_i)$ must belong to the support of β; otherwise the push-forward would give positive mass to a point outside that support. For any target atom z, $\beta(\{z\}) = \alpha(T^{-1}(\{z\})) = \frac{1}{n}\#\{i \,;\, T(x_i) = z\}$. This proves the counting statement. If all target atoms have mass $1/n$, each target receives exactly one source atom, which is a permutation. The converse and the cost identity follow by direct substitution. □

Proposition 2.11 (Existence of transport maps from atomless sources)**.** *Let α and β be Borel probability measures on Polish spaces, and assume that α is atomless. Then there exists a measurable map T such that $T_\sharp\alpha = \beta$.*

Proof. A standard measure-isomorphism theorem identifies the atomless probability space generated by α with Lebesgue measure on $[0,1]$, modulo null sets. It is therefore enough to construct a map from $[0,1]$ to the target law β. Choose a Borel isomorphism i from the support of β onto a Borel subset of $[0,1]$, set $\nu = i_\sharp\beta$, and use the generalized inverse of the cumulative distribution function of ν. This map sends Lebesgue measure on $[0,1]$ to ν 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 α to β. □

Monge distance. When $\mathcal{X} = \mathcal{Y}$ and $c(x,y) = d(x,y)^p$ for a metric d, set
$\mathcal{E}_\alpha(T) \coloneqq \int_{\mathcal{X}} d(x,T(x))^p \mathrm{d}\alpha(x)$.

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

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

Proposition 2.12 (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, $\alpha = \beta$.
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 ε-minimizers $S_\sharp\alpha = \gamma$ and $T_\sharp\gamma = \beta$ such that $\mathcal{E}_\alpha(S)^{1/p} \leqslant \tilde{\mathcal{W}}_p(\alpha,\gamma) + \varepsilon$, $\quad \mathcal{E}_\gamma(T)^{1/p} \leqslant \tilde{\mathcal{W}}_p(\gamma,\beta) + \varepsilon$. The composed map is feasible from α to β, and Minkowski's inequality gives

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

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

2.4 Existence and Uniqueness of the Monge Map

Brenier's theorem.

Theorem 2.13 (Brenier)**.** *Let $\alpha,\beta \in \mathcal{M}_+^1(\mathbb{R}^d)$ have finite second moments, and assume that α is absolutely continuous with respect to Lebesgue measure. For the quadratic cost $c(x,y) = \|x-y\|^2$, there exists a convex function $\varphi : \mathbb{R}^d \to \mathbb{R}\cup\{+\infty\}$ such that $T = \nabla\varphi, \quad T_\sharp\alpha = \beta$, and T is the unique optimal Monge map α-almost everywhere. The optimal Kantorovich plan is $(\mathrm{Id},T)_\sharp\alpha$.*

Proof. 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 φ. Hence the support of every optimal plan lies in the graph of the subdifferential $\partial\varphi$. Since α has a density and convex functions are differentiable Lebesgue-almost everywhere, $\partial\varphi(x)$ is a singleton for α-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 α-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. □

Definition 2.14 (Measures not charging hypersurfaces)**.** A Borel measure α on $\mathbb{R}^d$ does not charge hypersurfaces if $\alpha(S) = 0$ for every countably $(d-1)$-rectifiable set S, i.e. every set that can be covered, up to an $\mathcal{H}^{d-1}$-null set, by countably many C^1 hypersurfaces.

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

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.

Proposition 2.15 (Polar factorization)**.** *Let $\Omega \subset \mathbb{R}^d$ be endowed with normalized Lebesgue measure λ, 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 φ such that $u = \nabla\varphi \circ s \qquad \lambda$-a.e. The Brenier factor $\nabla\varphi$ is unique λ-almost everywhere.*

Proof. By Brenier's theorem there is a unique gradient of a convex function $T = \nabla\varphi$ transporting λ to ν. 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 2.13. □

The name is meant to echo the usual polar decomposition of matrices. This analogy becomes literal for linear maps under the Gaussian reference measure.
$A = SO, \qquad O = S^\dagger A,$

Displacement interpolation.

Definition 2.16 (Monge and McCann displacement interpolation)**.** If $T_\sharp\alpha = \beta$, the Monge interpolation generated by T is the curve $T_t(x) \coloneqq (1-t)x + tT(x), \qquad \alpha_t \coloneqq (T_t)_\sharp\alpha, \qquad t \in [0,1]$. For the quadratic cost, when T is the optimal Brenier map, this curve is called McCann's displacement interpolation.

Proposition 2.17 (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 $\tilde{\mathcal{W}}_p(\alpha,\beta)$ in (2.6). Set $\alpha_t = (T_t)_\sharp\alpha$ with $T_t = (1-t)\mathrm{Id} + tT$. Assume that, for every $t < 1$, T_t is one-to-one on a full α-measure Borel set, so that T_t^{-1} is defined α_t-almost everywhere. Then, for $0 \leqslant s \leqslant t \leqslant 1$, $\tilde{\mathcal{W}}_p(\alpha_s,\alpha_t) = (t-s)\tilde{\mathcal{W}}_p(\alpha,\beta)$. Thus $t \mapsto \alpha_t$ is an oriented constant-speed geodesic for the directed Monge distance. In particular, for $p = 2$, this applies to the Brenier map under the hypotheses of Theorem 2.13.*

Proof. The case $s = t$ is trivial, so assume $s < t$. Since $s < 1$, the inverse T_s^{-1} is defined α_s-almost everywhere along the transported particles. Define $S_{s,t} \coloneqq T_t \circ T_s^{-1} \qquad \alpha_s$-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 α_t to β and gives $\tilde{\mathcal{W}}_p(\alpha_t,\beta) \leqslant (1-t)\tilde{\mathcal{W}}_p(\alpha,\beta)$; if $t = 1$, this latter distance is zero. Using the triangle inequality from Proposition 2.12,

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

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

Regularity and the Monge–Ampère equation.

Proposition 2.18 (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 Ω and $\beta = \eta(y)\mathrm{d}y$ be supported on Λ, 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 φ transporting α to β 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 φ; the boundary statement follows from the boundary regularity theory under uniform convexity and compatibility assumptions. □

For smooth densities, the change-of-variables formula (2.4) gives the Monge–Ampère equation

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

Proposition 2.19 (Linearization of the Monge–Ampère equation)**.** *Let $\rho_\varepsilon = \rho_0 + \varepsilon r + o(\varepsilon)$ be a smooth perturbation of a positive reference density ρ_0 on a smooth bounded domain, with $\int r = 0$. If $T_\varepsilon(x) = x + \varepsilon\nabla u(x) + o(\varepsilon)$ transports ρ_0 to ρ_ε, then, to first order, $-\nabla\cdot(\rho_0\nabla u) = r$. In particular, when ρ_0 is constant, the linearized equation is $-\Delta u = r/\rho_0$.*

Proof. The change-of-variables equation for T_ε is $\rho_0(x) = \rho_\varepsilon(T_\varepsilon(x))\det(\nabla T_\varepsilon(x))$. Expanding $\rho_\varepsilon(x + \varepsilon\nabla u) = \rho_0(x) + \varepsilon r(x) + \varepsilon\langle\nabla\rho_0,\, \nabla u\rangle + o(\varepsilon)$ and $\det(I + \varepsilon\nabla^2 u) = 1 + \varepsilon\Delta u + o(\varepsilon)$ 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

Definition 2.20 (Cumulative and quantile functions)**.** For $\alpha \in \mathcal{M}_+^1(\mathbb{R})$, its cumulative distribution function is

$$\mathcal{C}_\alpha(x) \coloneqq \alpha((-\infty, x]). \tag{2.8}$$

Its generalized inverse, or quantile function, is

$$\mathcal{C}_\alpha^{-1}(r) \coloneqq \inf\{x \in \mathbb{R}\ ;\ \mathcal{C}_\alpha(x) \geqslant r\}, \qquad r \in (0,1). \tag{2.9}$$

Proposition 2.21 (Quantile push-forward)**.** *One has $(\mathcal{C}_\alpha^{-1})_\sharp\mathrm{Leb}_{[0,1]} = \alpha$. If α has no atoms, then $(\mathcal{C}_\alpha)_\sharp\alpha = \mathrm{Leb}_{[0,1]}$.*

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

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

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

If α has no atoms, the map

$$T = \mathcal{C}_\beta^{-1} \circ \mathcal{C}_\alpha \tag{2.10}$$

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

Proposition 2.22 (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 $\pi^\star = (\mathcal{C}_\alpha^{-1}, \mathcal{C}_\beta^{-1})_\sharp \mathrm{Leb}_{[0,1]}$ minimizes $\int |x-y|^p \mathrm{d}\pi(x,y)$ among couplings. If α has no atoms, it is induced by the monotone Monge map (2.10).*

Proof. The displayed measure is a coupling by Proposition 2.21. Its support is monotone: for $s < t$, both quantile functions satisfy $\mathcal{C}_\alpha^{-1}(s) \leqslant \mathcal{C}_\alpha^{-1}(t)$ and $\mathcal{C}_\beta^{-1}(s) \leqslant \mathcal{C}_\beta^{-1}(t)$. 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 α has no atoms, $(\mathcal{C}_\alpha)_\sharp \alpha = \mathrm{Leb}_{[0,1]}$, so the coupling is generated by (2.10). □

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.

Proposition 2.23 (One-dimensional Wasserstein formulas)**.** *Let $\alpha, \beta \in \mathcal{M}_+^1(\mathbb{R})$ have finite p-th moments. For every $p \geqslant 1$,*

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

For $p = 1$, this is equivalently

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

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

Proof. The first formula follows from Proposition 2.22: the optimal coupling is obtained by taking the same quantile level r for both measures. For $p = 1$, use the layer-cake identity. If q_α and q_β are the two quantile functions, then $\int_0^1 |q_\alpha(r) - q_\beta(r)|\mathrm{d}r = \int_{\mathbb{R}} \lambda\big(\{r : q_\alpha(r) \leqslant x < q_\beta(r)\} \cup \{r : q_\beta(r) \leqslant x < q_\alpha(r)\}\big)\mathrm{d}x$. The measure of the displayed set is exactly $|\mathcal{C}_\alpha(x) - \mathcal{C}_\beta(x)|$ for almost every x. □

Formula (2.11) means that the map $\alpha \mapsto \mathcal{C}_\alpha^{-1}$ 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.

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

Triangular rearrangements.

Proposition 2.24 (Knothe–Rosenblatt triangular rearrangement)**.** *Let $\alpha, \beta \in \mathcal{M}_+^1(\mathbb{R}^d)$. Assume, for simplicity, that the first marginal of α and the one-dimensional conditional laws of α 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 α-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 α and the first marginal of β. Suppose that $T_1,\dots,T_{k-1}$ have already been constructed. Write $x_{<k} = (x_1,\dots,x_{k-1})$ and $T_{<k} = (T_1,\dots,T_{k-1})$. By induction, $(T_{<k})_\sharp \alpha_{<k} = \beta_{<k}$, where $\alpha_{<k}$ and $\beta_{<k}$ are the first $(k-1)$-coordinate marginals. Let $\alpha^k_{x_{<k}}$ and $\beta^k_{y_{<k}}$ be regular conditional laws of the k-th coordinate given the previous coordinates. Define $T_k(x_{<k},\cdot)$ as the one-dimensional monotone rearrangement from $\alpha^k_{x_{<k}}$ to $\beta^k_{T_{<k}(x_{<k})}$.

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

Proposition 2.25 (Anisotropic Brenier maps converge to Knothe–Rosenblatt)**.** *Let α, β be compactly supported probability measures on $\mathbb{R}^d$ with densities bounded above and below on their rectangular supports, and assume that the conditional laws entering Proposition 2.24 are atomless. For $\varepsilon > 0$, set $c_\varepsilon(x,y) := \sum_{k=1}^d \varepsilon^{k-1} |x_k - y_k|^2$. Let T_ε be the Monge map from α to β for the cost c_ε, and let T_{KR} be the triangular Knothe–Rosenblatt rearrangement with the coordinate order used above. Then*

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

Proof. Let $\pi_\varepsilon = (\mathrm{Id}, T_\varepsilon)_\sharp \alpha$. Since the supports are compact, a subsequence converges weakly to some coupling π between α and β. The optimality of π_ε 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 π minimizes the one-dimensional quadratic cost in the first coordinate among all couplings. Since the first marginal of α is atomless, the one-dimensional minimizer is the monotone rearrangement, so $y_1 = T_1(x_1)$ under π.

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 ε, 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 π.

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_{KR}. This graph coupling is unique, so the whole family π_ε 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_{KR}

is continuous. On K, the set $\{(x,y)\ ;\ x\in K,\ \|y-T_{\mathrm{KR}}(x)\|\geqslant\delta\}$ is closed and has zero mass under the limiting graph coupling. Portmanteau's theorem gives

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

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

2.6 Gaussian Measures and the Bures Metric

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

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

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 $(m,\sigma)\in\mathbb{R}\times\mathbb{R}_+$.

Multivariate Gaussians.

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

Proposition 2.26 (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 β. □

Definition 2.27 (Bures metric)**.** For positive semidefinite covariance matrices Σ and Λ, the Bures metric is

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

Proposition 2.28 (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 2.27.

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

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

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

Proposition 2.29 (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 $\|\Sigma^{1/2}-\Gamma^{1/2}Q_2Q_1\|_F\leqslant\|\Sigma^{1/2}-\Lambda^{1/2}Q_1\|_F+\|\Lambda^{1/2}-\Gamma^{1/2}Q_2\|_F$. 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 $U_t=[\sqrt{1-t}\,U_0,\sqrt{t}\,U_1]$, $V_t=[\sqrt{1-t}\,V_0,\sqrt{t}\,V_1]$. Then $U_tU_t^\top=(1-t)\Sigma_0+t\Sigma_1$ and $V_tV_t^\top=(1-t)\Lambda_0+t\Lambda_1$, 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. □

3 Kantorovich Relaxation

3.1 Discrete Relaxation

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

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

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

Equivalently,

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

Definition 3.2 (Discrete product coupling)**.** Given weights $\mathrm{a} \in \Sigma_n$ and $\mathrm{b} \in \Sigma_m$, the discrete product, or trivial, coupling is $(\mathrm{a} \otimes \mathrm{b})_{i,j} \coloneqq \mathrm{a}_i \mathrm{b}_j$. It belongs to $\mathrm{U}(\mathrm{a}, \mathrm{b})$ and corresponds to choosing the source and target labels independently.

Proposition 3.3 (Discrete product optimality is degenerate)**.** *Assume that the zero-mass rows and columns have been removed, so that* $\mathrm{a}_i > 0$ *and* $\mathrm{b}_j > 0$*, and let* C *be a finite cost matrix. The product plan* $\mathrm{a} \otimes \mathrm{b}$ *minimizes* $\mathrm{P} \mapsto \langle \mathrm{C}, \mathrm{P} \rangle$ *over* $\mathrm{U}(\mathrm{a}, \mathrm{b})$ *if and only if every coupling* $\mathrm{P} \in \mathrm{U}(\mathrm{a}, \mathrm{b})$ *minimizes it.*

Proof. The reverse implication is immediate. Assume conversely that $\mathrm{a} \otimes \mathrm{b}$ is optimal and let $\mathrm{Q} \in \mathrm{U}(\mathrm{a}, \mathrm{b})$ be arbitrary. Since all entries of $\mathrm{a} \otimes \mathrm{b}$ are positive, there exists $t > 0$ small enough that $\mathrm{R} \coloneqq (1+t)(\mathrm{a} \otimes \mathrm{b}) - t\mathrm{Q}$ has nonnegative entries. Its row and column sums are $\mathrm{R}\mathbb{1}_m = (1+t)\mathrm{a} - t\mathrm{a} = \mathrm{a}, \qquad \mathrm{R}^\top \mathbb{1}_n = (1+t)\mathrm{b} - t\mathrm{b} = \mathrm{b}$, so $\mathrm{R} \in \mathrm{U}(\mathrm{a}, \mathrm{b})$. Moreover, $\mathrm{a} \otimes \mathrm{b} = \frac{1}{1+t}\mathrm{R} + \frac{t}{1+t}\mathrm{Q}$. By optimality of $\mathrm{a} \otimes \mathrm{b}$, both $\langle \mathrm{C}, \mathrm{R} \rangle$ and $\langle \mathrm{C}, \mathrm{Q} \rangle$ are at least $\langle \mathrm{C}, \mathrm{a} \otimes \mathrm{b} \rangle$. Taking the scalar product of the convex-combination identity with C gives $\langle \mathrm{C}, \mathrm{a} \otimes \mathrm{b} \rangle = \frac{1}{1+t}\langle \mathrm{C}, \mathrm{R} \rangle + \frac{t}{1+t}\langle \mathrm{C}, \mathrm{Q} \rangle$. A convex average of two numbers not smaller than $\langle \mathrm{C}, \mathrm{a} \otimes \mathrm{b} \rangle$ can equal $\langle \mathrm{C}, \mathrm{a} \otimes \mathrm{b} \rangle$ only if both numbers are equal to it. Hence $\langle \mathrm{C}, \mathrm{Q} \rangle = \langle \mathrm{C}, \mathrm{a} \otimes \mathrm{b} \rangle$, 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.

$$\mathrm{L}_\mathrm{C}(\mathrm{a}, \mathrm{b}) \coloneqq \min_{\mathrm{P} \in \mathrm{U}(\mathrm{a},\mathrm{b})} \langle \mathrm{C}, \mathrm{P} \rangle \coloneqq \min_{\mathrm{P} \in \mathrm{U}(\mathrm{a},\mathrm{b})} \sum_{i,j} \mathrm{C}_{i,j} \mathrm{P}_{i,j}. \tag{3.2}$$

Proposition 3.4 (Sparse optimal plans)**.** *Assume* $\mathrm{a}_i > 0$, $\mathrm{b}_j > 0$ *and* $\sum_i \mathrm{a}_i = \sum_j \mathrm{b}_j = 1$*. 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 $E = \{(i,j) : \mathrm{P}_{ij} > 0\}$ 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 $\mathrm{P} + tH$ and $\mathrm{P} - tH$ 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* $\mathrm{a} \in \mathbb{R}_+^n$ *and* $\mathrm{b} \in \mathbb{R}_+^m$ *have the same positive total mass. 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. □

One-dimensional cases.

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)*.*

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 η 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. □

Permutation matrices as couplings.

Definition 3.7 (Permutation matrices)**.** For a permutation $\sigma \in \mathrm{Perm}(n)$, its permutation matrix P_σ is

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

The set of all permutation matrices is $\mathcal{P}_n^{\mathrm{perm}} \coloneqq \{P_\sigma \ ; \ \sigma \in \mathrm{Perm}(n)\}$.

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

Definition 3.8 (Birkhoff polytope)**.** The Birkhoff polytope is the convex set of bistochastic matrices $\mathcal{B}_n \coloneqq \left\{P \in \mathbb{R}_+^{n \times n} \ ; \ P\mathbb{1}_n = \mathbb{1}_n \quad \text{and} \quad P^\top \mathbb{1}_n = \mathbb{1}_n\right\}$.

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, $\mathcal{P}_n^{\mathrm{perm}} = \mathcal{B}_n \cap \{0,1\}^{n \times n} \subset \mathcal{B}_n$, $\min_{P \in \mathcal{B}_n} \langle \mathrm{C}, P \rangle \leqslant \min_{P \in \mathcal{P}_n^{\mathrm{perm}}} \langle \mathrm{C}, P \rangle$.

Definition 3.9 (Extreme points)**.** For a compact convex set $\mathcal{C}$ in a finite-dimensional vector space, $\mathrm{Extr}(\mathcal{C}) \coloneqq \{x \in \mathcal{C} \ ; \ x = (y+z)/2, \ y, z \in \mathcal{C} \Rightarrow y = z = x\}$.

Proposition 3.10 (Existence of extreme points)**.** *If $\mathcal{C}$ is a non-empty compact convex subset of a finite-dimensional vector space, then $\mathrm{Extr}(\mathcal{C})$ is non-empty.*

Proof. Among all non-empty faces of $\mathcal{C}$, 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. □

Proposition 3.11 (Linear programs have extreme minimizers)**.** *Let $\mathcal{C}$ be non-empty and compact. For every linear form ℓ, $\mathrm{Extr}(\mathcal{C}) \cap \mathrm{argmin}_{x \in \mathcal{C}} \ell(x) \neq \emptyset$.*

Proof. The set $S = \mathrm{argmin}_{x \in \mathcal{C}} \ell(x)$ is non-empty, compact and convex. By Proposition 3.10, it has an extreme point x. If $x = (y+z)/2$ with $y, z \in \mathcal{C}$, then by linearity and optimality of x, both y and z also minimize ℓ on $\mathcal{C}$, hence $y, z \in S$. Since x is extreme in S, $y = z = x$. Thus x is extreme in $\mathcal{C}$. □

Theorem 3.12 (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 $P_\sigma = \frac{Q+R}{2}$ with $Q, R \in \mathcal{B}_n$. Every bistochastic matrix has entries in $[0,1]$. Since the only extreme points of $[0,1]$ are 0 and 1, each entry of P_σ 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_σ is extreme.
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 $(i_1, j_1, i_2, j_2, \ldots, i_p, j_p)$, $i_{p+1} = i_1$, 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, $0 < P_{i_s,j_s} < 1$ and $0 < P_{i_{s+1},j_s} < 1$. 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 $A \coloneqq \{(i_s, j_s)\}_{s=1}^p$, $B \coloneqq \{(i_{s+1}, j_s)\}_{s=1}^p$.

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

By the definition of ε, all modified entries stay in $[0,1]$, so Q and R are nonnegative. Each row vertex i_s of the cycle is incident to exactly one edge of A and one edge of B; the $+\varepsilon/2$ and $-\varepsilon/2$ perturbations therefore cancel in that row. The same cancellation holds in each column vertex j_s, and all other rows and columns are unchanged. $Q\mathbb{1}_n = R\mathbb{1}_n = \mathbb{1}_n$, $Q^\top \mathbb{1}_n = R^\top \mathbb{1}_n = \mathbb{1}_n$, 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. □

Corollary 3.13 (Kantorovich for matching)**.** *If $m = n$ and $\mathrm{a} = \mathrm{b} = \mathbb{1}_n/n$, then the discrete Kantorovich problem (3.2) admits an optimal solution of the form P_σ/n. The associated permutation σ solves the assignment problem of Section 1.1.*

Proof. The feasible set is $\mathcal{B}_n/n$. By Proposition 3.11, the linear objective has an optimal extreme point. Since scaling preserves extreme points and Theorem 3.12 identifies the extreme points of $\mathcal{B}_n$, this optimizer is P_σ/n for some permutation σ. Its cost is exactly $n^{-1} \sum_i C_{i,\sigma(i)}$, so σ is an optimal assignment. □

3.2 Linear-Programming Algorithms

Transportation simplex and network simplex.

Interior-point methods.

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

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.

3.3 Relaxation for Arbitrary Measures

Continuous couplings.

Definition 3.14 (Marginals of a joint measure)**.** Let $\pi \in \mathcal{M}_+^1(\mathcal{X}\times\mathcal{Y})$ and let $P_{\mathcal{X}}(x,y) = x$, $P_{\mathcal{Y}}(x,y) = y$ be the coordinate projections. The marginals of π are $\pi_1 \coloneqq (P_{\mathcal{X}})_\sharp\pi \in \mathcal{M}_+^1(\mathcal{X})$, $\pi_2 \coloneqq (P_{\mathcal{Y}})_\sharp\pi \in \mathcal{M}_+^1(\mathcal{Y})$. Equivalently, for all bounded continuous test functions f on $\mathcal{X}$ and g on $\mathcal{Y}$, $\int_{\mathcal{X}\times\mathcal{Y}} f(x)\mathrm{d}\pi(x,y) = \int_{\mathcal{X}} f\mathrm{d}\pi_1$, $\int_{\mathcal{X}\times\mathcal{Y}} g(y)\mathrm{d}\pi(x,y) = \int_{\mathcal{Y}} g\mathrm{d}\pi_2$.

$\int_{\mathcal{Y}} \mathrm{d}\pi(x,y) = \mathrm{d}\alpha(x)$, $\int_{\mathcal{X}} \mathrm{d}\pi(x,y) = \mathrm{d}\beta(y)$, which is made rigorous by Definition 3.14, 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 3.15 (Couplings)**.** Given $\alpha \in \mathcal{M}_+^1(\mathcal{X})$ and $\beta \in \mathcal{M}_+^1(\mathcal{Y})$, the set of couplings between α and β is

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

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

Definition 3.16 (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.

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

Proposition 3.17 (Product optimality is degenerate)**.** *Assume that $\mathcal{X}$ and $\mathcal{Y}$ are compact metric spaces and that $c\in\mathcal{C}(\mathcal{X}\times\mathcal{Y})$. $\alpha\otimes\beta \in \operatorname*{argmin}_{\pi\in\mathcal{U}(\alpha,\beta)} \int c\,\mathrm{d}\pi$, every coupling in $\mathcal{U}(\alpha,\beta)$ 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 α_i and β_i the normalized restrictions of α to U_i and of β to V_i, and choose $0 < \lambda \leqslant \min\{\alpha(U_0)\beta(V_0), \alpha(U_1)\beta(V_1)\}$. The exchanged measure $\tilde\pi = \alpha\otimes\beta - \lambda\,\alpha_0\otimes\beta_0 - \lambda\,\alpha_1\otimes\beta_1 + \lambda\,\alpha_0\otimes\beta_1 + \lambda\,\alpha_1\otimes\beta_0$ is nonnegative and has the same two marginals as $\alpha\otimes\beta$. The uniform strict inequality on the neighborhoods implies that $\int c\,\mathrm{d}\tilde\pi < \int c\,\mathrm{d}(\alpha\otimes\beta)$, contradicting optimality.
Fixing any $x_\star \in \mathrm{supp}(\alpha)$ and $y_\star \in \mathrm{supp}(\beta)$, the equality of cross differences gives, for all $(x,y) \in \mathrm{supp}(\alpha)\times\mathrm{supp}(\beta)$, $c(x,y) = c(x,y_\star) + c(x_\star,y) - c(x_\star,y_\star)$. Thus $c = u + v$ on the product support. 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.
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.

Continuous Kantorovich problem.

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

Proposition 3.18 (Existence on compact spaces)**.** *Assume that $\mathcal{X}$ and $\mathcal{Y}$ are compact metric spaces and that $c\in\mathcal{C}(\mathcal{X}\times\mathcal{Y})$. 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. □

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

Monge–Kantorovich equivalence.

Corollary 3.19 (Monge–Kantorovich equivalence under Brenier)**.** *Assume that α 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 φ. When α has a density, φ is differentiable α-almost everywhere, so $\partial\varphi(x) = \{\nabla\varphi(x)\}$ for α-almost every x. Thus every optimal coupling is concentrated on the graph of $T = \nabla\varphi$ and must equal $(\mathrm{Id}, T)_\sharp\alpha$. The graph coupling is feasible and optimal, and the two formulations have the same value. □

If α does not have a density, then φ may be non-smooth on a set charged by α, and non-smooth points can lead to mass splitting. This is the continuous counterpart of the gap between the uniform matching case of Corollary 3.13 and the general splitting case.

3.4 c-Cyclical Monotonicity

Support and c-cyclical monotonicity.

Definition 3.20 (Support)**.** For a Radon measure π on $\mathcal{X}\times\mathcal{Y}$, $\mathrm{supp}(\pi) \coloneqq \{(x,y)\ ;\ \pi(U\times V) > 0 \text{ for every open } U \ni x,\ V \ni y\}$.

Definition 3.21 (c-cyclical monotonicity)**.** A set $\Gamma \subset \mathcal{X}\times\mathcal{Y}$ is c-cyclically monotone if, for every $k \geqslant 2$, every finite family $(x_i, y_i)_{i=1}^k \subset \Gamma$ and every permutation σ of $\{1,\dots,k\}$, $\sum_{i=1}^k c(x_i,y_i) \leqslant \sum_{i=1}^k c(x_i, y_{\sigma(i)})$.

$\sum_{i=1}^k c(x_i,y_i) \leqslant \sum_{i=1}^k c(x_i,y_{i+1}), \qquad y_{k+1} = y_1.$

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 3.13, 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.

Theorem 3.22 (Optimal plans are c-cyclically monotone)**.** *Assume c is continuous. For any optimal plan π solving the Kantorovich problem (3.5), $\mathrm{supp}(\pi)$ is c-cyclically monotone.*

Proof. Suppose that $\mathrm{supp}(\pi)$ is not c-cyclically monotone. Then there exist points $(x_i,y_i)_{i=1}^k$ in the support and a permutation σ such that $\sum_i c(x_i,y_i) > \sum_i c(x_i,y_{\sigma(i)})$. By continuity of c, after shrinking neighborhoods $U_i \ni x_i$ and $V_i \ni y_i$, the same strict inequality holds uniformly for every choice of points in these neighborhoods: $\sum_i c(u_i,v_i) > \sum_i c(u_i,\tilde v_{\sigma(i)}) \qquad (u_i \in U_i,\ v_i \in V_i,\ \tilde v_{\sigma(i)} \in V_{\sigma(i)})$. Choose the sets so that $\pi(U_i\times V_i) > 0$. 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 λ 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 σ is a permutation. Hence $\tilde\pi \in \mathcal{U}(\alpha,\beta)$. Integrating the uniform strict inequality against the product probability $\otimes_i(\pi_i/\lambda)$ shows that the reinserted crossed terms have strictly smaller cost than the removed diagonal terms. This contradicts the optimality of π. □

Monotonicity. Assume the optimal plan is induced by a measurable map $T : \mathcal{X} \to \mathcal{Y}$, 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,
$\langle T(x) - T(y),\ x - y\rangle \geqslant 0,$

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

OT defines a distance.

Lemma 3.23 (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 $\sum_k \mathrm{S}_{i,j,k} = \mathrm{P}_{i,j}$ and $\sum_i \mathrm{S}_{i,j,k} = \mathrm{Q}_{j,k}$. $\mathrm{R}_{i,k} \coloneqq \sum_j \mathrm{S}_{i,j,k}$, belongs to $\mathrm{U}(\mathrm{a},\mathrm{c})$. For the canonical construction below, this glued coupling is the twisted matrix product $\mathrm{R} = \mathrm{P}\,\mathrm{diag}(1/\mathrm{b})\mathrm{Q}$, $\mathrm{R}_{i,k} = \sum_{j:\mathrm{b}_j>0} \frac{\mathrm{P}_{i,j}\mathrm{Q}_{j,k}}{\mathrm{b}_j}$. In the matrix notation, $1/\mathrm{b}_j$ is understood as 0 when $\mathrm{b}_j = 0$.*

Proof. One verifies that

$$\mathrm{S}_{i,j,k} = \begin{cases} \frac{\mathrm{P}_{i,j}\mathrm{Q}_{j,k}}{\mathrm{b}_j} & \text{if} \quad \mathrm{b}_j \neq 0 \\ 0 \text{ otherwise} \end{cases} \tag{3.6}$$

is acceptable. Indeed, if $\mathrm{b}_j \neq 0$ $\sum_k \mathrm{S}_{i,j,k} = \sum_k \frac{\mathrm{P}_{i,j}\mathrm{Q}_{j,k}}{\mathrm{b}_j} = \frac{\mathrm{P}_{i,j}}{\mathrm{b}_j}(\mathrm{Q}\mathbb{1}_m)_j = \frac{\mathrm{P}_{i,j}}{\mathrm{b}_j}\mathrm{b}_j$. If $\mathrm{b}_j = 0$, then necessarily $\mathrm{P}_{i,j} = 0$ and $\sum_k \mathrm{S}_{i,j,k} = 0 = \mathrm{P}_{i,j}$. The same computation gives the other prescribed marginal:

$$\sum_i \mathrm{S}_{i,j,k} = \begin{cases} \frac{\mathrm{Q}_{j,k}}{\mathrm{b}_j}\sum_i \mathrm{P}_{i,j} = \mathrm{Q}_{j,k} & \text{if} \quad \mathrm{b}_j > 0 \\ 0 = \mathrm{Q}_{j,k} & \text{if} \quad \mathrm{b}_j = 0. \end{cases}$$

Summing over j then gives the displayed formula for R. Its row and column sums are $\sum_k \mathrm{R}_{i,k} = \sum_j \mathrm{P}_{i,j} = \mathrm{a}_i, \qquad \sum_i \mathrm{R}_{i,k} = \sum_j \mathrm{Q}_{j,k} = \mathrm{c}_k$, so $\mathrm{R} \in \mathrm{U}(\mathrm{a},\mathrm{c})$. □

When the cost matrix is the pth power of a distance matrix, the discrete Kantorovich value becomes a metric on histograms.

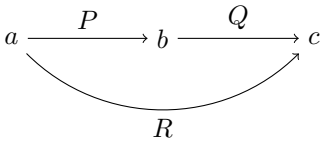
Definition 3.24 (Discrete Wasserstein distance)**.** Let $\mathrm{D} \in \mathbb{R}_+^{n\times n}$ be a distance matrix on $[\![n]\!]$ and let $p \geqslant 1$. The discrete p-Wasserstein distance between histograms $\mathrm{a},\mathrm{b} \in \Sigma_n$ is

$$\mathrm{W}_p(\mathrm{a},\mathrm{b}) \coloneqq \mathrm{L}_{\mathrm{D}^p}(\mathrm{a},\mathrm{b})^{1/p}. \tag{3.7}$$

It depends on the chosen ground distance D.

Proposition 3.25 (Metric property of discrete Wasserstein distance)**.** *For every distance matrix* D *on* $[\![n]\!]$*, Definition 3.24 defines a distance on* Σ_n*:* W_p *is symmetric, positive,* $\mathrm{W}_p(\mathrm{a},\mathrm{b})=0$ *if and only if* $\mathrm{a}=\mathrm{b}$*, and it satisfies the triangle inequality* $\forall\,\mathrm{a},\mathrm{b},\mathrm{c}\in\Sigma_n,\quad \mathrm{W}_p(\mathrm{a},\mathrm{c})\leqslant \mathrm{W}_p(\mathrm{a},\mathrm{b})+\mathrm{W}_p(\mathrm{b},\mathrm{c})$.

Proof. For symmetry, since D^p is symmetric, we use the fact that if $\mathrm{P}\in\mathrm{U}(\mathrm{a},\mathrm{b})$ is optimal for $\mathrm{W}_p(\mathrm{a},\mathrm{b})$, then $\mathrm{P}^\top\in\mathrm{U}(\mathrm{b},\mathrm{a})$ is optimal for $\mathrm{W}_p(\mathrm{b},\mathrm{a})$. For definiteness, since $\mathrm{C}=\mathrm{D}^p$ has a null diagonal, $\mathrm{W}_p(\mathrm{a},\mathrm{b})=0$ is achieved by the diagonal coupling $\mathrm{P}^\star=\mathrm{diag}(\mathrm{a})=\mathrm{diag}(\mathrm{b})$ when $\mathrm{a}=\mathrm{b}$; by positivity of all off-diagonal elements of D^p, $\mathrm{W}_p(\mathrm{a},\mathrm{b})>0$ whenever $\mathrm{a}\neq\mathrm{b}$ because any admissible coupling then has a nonzero element outside the diagonal.
To prove the triangle inequality in this discrete setting, we consider $\mathrm{a},\mathrm{b},\mathrm{c}\in\Sigma_n$, 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 3.23 which defines $\mathrm{S}\in\mathbb{R}_+^{n^3}$ with marginals $\sum_k \mathrm{S}_{\cdot,\cdot,k}=\mathrm{P}$ and $\sum_i \mathrm{S}_{i,\cdot,\cdot}=\mathrm{Q}$. We define $\mathrm{R}=\sum_j \mathrm{S}_{\cdot,j,\cdot}$, which is an element of $\mathrm{U}(\mathrm{a},\mathrm{c})$.



Note that if one assumes $\mathrm{b}>0$ then $\mathrm{R}=\mathrm{P}\,\mathrm{diag}(1/\mathrm{b})\mathrm{Q}$.
The triangle inequality follows from

$$\begin{aligned}
\mathrm{W}_p(\mathrm{a},\mathrm{c}) &= \Big(\min_{\tilde{\mathrm{R}}\in\mathrm{U}(\mathrm{a},\mathrm{c})}\langle\tilde{\mathrm{R}},\,\mathrm{D}^p\rangle\Big)^{1/p}\leqslant\langle\mathrm{R},\,\mathrm{D}^p\rangle^{1/p}\\
&=\Big(\sum_{i,k}\mathrm{D}_{ik}^p\sum_j\mathrm{S}_{i,j,k}\Big)^{1/p}\leqslant\Big(\sum_{i,j,k}\Big(\mathrm{D}_{ij}+\mathrm{D}_{j,k}\Big)^p\mathrm{S}_{i,j,k}\Big)^{1/p}\\
&\leqslant\Big(\sum_{i,j,k}\mathrm{D}_{ij}^p\mathrm{S}_{i,j,k}\Big)^{1/p}+\Big(\sum_{i,j,k}\mathrm{D}_{j,k}^p\mathrm{S}_{i,j,k}\Big)^{1/p}\\
&=\Big(\sum_{i,j}\mathrm{D}_{i,j}^p\sum_k\mathrm{S}_{i,j,k}\Big)^{1/p}+\Big(\sum_{j,k}\mathrm{D}_{j,k}^p\sum_i\mathrm{S}_{i,j,k}\Big)^{1/p}\\
&=\Big(\sum_{i,j}\mathrm{D}_{i,j}^p\mathrm{P}_{i,j}\Big)^{1/p}+\Big(\sum_{j,k}\mathrm{D}_{j,k}^p\mathrm{Q}_{j,k}\Big)^{1/p}=\mathrm{W}_p(\mathrm{a},\mathrm{b})+\mathrm{W}_p(\mathrm{b},\mathrm{c}).
\end{aligned}$$

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.
Lemma 3.26 (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* $(P_{\mathcal{X},\mathcal{Y}})_\sharp\sigma=\pi\quad$ *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 π and ξ against β 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 $\forall\,h\in\mathcal{C}(\mathcal{X}\times\mathcal{Y}),\quad\int_{\mathcal{Y}}\Big(\int_{\mathcal{X}}h(x,y)\mathrm{d}\pi_y(x)\Big)\mathrm{d}\beta(y)=\int h(x,y)\mathrm{d}\pi(x,y)$. and similarly for ξ.
When $\beta=\sum_j \mathrm{b}_j\delta_{y_j}$ and $\pi=\sum_{i,j}\mathrm{P}_{i,j}\delta_{(x_i,y_j)}$, then this conditional distribution is defined on the support of β as $\pi_{y_j}=\sum_i\frac{\mathrm{P}_{i,j}}{\mathrm{b}_j}\delta_{x_i}$ (and similarly for ξ).
The glued measure is then defined by the conditional-product formula $\forall\,g\in\mathcal{C}(\mathcal{X}\times\mathcal{Y}\times\mathcal{Z}),\ \int g(x,y,z)\mathrm{d}\sigma(x,y,z)=\int g(x,y,z)\mathrm{d}\pi_y(x)\mathrm{d}\xi_y(z)\mathrm{d}\beta(y)$. For discrete measures, this matches the definition (3.6), since $\sigma=\sum_{i,j,k}\mathrm{S}_{i,j,k}\delta_{x_i,y_j,z_k}$ where $\mathrm{S}_{i,j,k}=\frac{\mathrm{P}_{i,j}}{\mathrm{b}_j}\frac{\mathrm{Q}_{j,k}}{\mathrm{b}_j}\mathrm{b}_j$. □

Definition 3.27 (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

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

It depends on the ground distance d.

Proposition 3.28 (Metric property of the Wasserstein distance)**.** *Definition 3.27 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 π 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 3.26.
We define the composition of the two couplings (π,ξ) as $\rho\coloneqq(P_{\mathcal{X},\mathcal{Z}})_\sharp\sigma$.

Note that if π and ξ are couplings induced by two Monge maps $T_{\mathcal{X}}(x)$ and $T_{\mathcal{Y}}(y)$, then ρ 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

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

□

Interpolation induced by an optimal plan.

Definition 3.29 ($\mathcal{W}_2$ 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 $e_t(x,y) \coloneqq (1-t)x + ty$, $\quad \alpha_t \coloneqq (e_t)_\sharp \pi^\star$. The curve $(\alpha_t)_{t\in[0,1]}$ is the displacement, or McCann, $\mathcal{W}_2$ geodesic induced by $\pi^\star$.

Proposition 3.30 (Optimal-plan interpolation is a $\mathcal{W}_2$ geodesic)**.** *Let $(\alpha_t)_{t\in[0,1]}$ be defined by Definition 3.29. Then, for every $0 \leqslant s \leqslant t \leqslant 1$, $\mathcal{W}_2(\alpha_s,\alpha_t) = (t-s)\,\mathcal{W}_2(\alpha_0,\alpha_1)$. Thus $t \mapsto \alpha_t$ is a constant-speed geodesic for the metric $\mathcal{W}_2$.*

Proof. Push the optimal plan $\pi^\star$ forward by (e_s, e_t). $\int \|z-z'\|^2 \mathrm{d}\gamma_{s,t}(z,z') = \int \|e_t(x,y) - e_s(x,y)\|^2 \mathrm{d}\pi^\star(x,y) = (t-s)^2\,\mathcal{W}_2^2(\alpha_0,\alpha_1)$. 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 3.28, 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. □

General geodesic spaces. For Dirac masses in Euclidean space, the $\mathcal{W}_2$ geodesic from δ_x to δ_y is $t \mapsto \delta_{(1-t)x+ty}$. The same idea extends to any geodesic metric space $(\mathcal{X}, d)$, meaning that each pair of points can be joined by a constant-speed metric geodesic. For each pair (x,y), one replaces the Euclidean segment by a curve $\gamma^{x,y} : [0,1] \to \mathcal{X}$ such that $\gamma_0^{x,y} = x$, $\gamma_1^{x,y} = y$, and $d(\gamma_s^{x,y}, \gamma_t^{x,y}) = |t-s| d(x,y)$.
Comparison with Monge.
3.6 Metric Properties: Topology and Applications
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 3.31 (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 π and to the convex function $\varphi(r) = r^{q/p}$ with $c(x,y) = d(x,y)^p$, so that one gets $\big(\int d(x,y)^p \mathrm{d}\pi(x,y)\big)^{\frac{q}{p}} \leqslant \int d(x,y)^q \mathrm{d}\pi(x,y)$. The right inequality follows from $d(x,y)^q \leqslant \mathrm{diam}(\mathcal{X})^{q-p} d(x,y)^p$. □

Definition 3.32 (Weak* topology)**.** $(\alpha_k)_k$ converges weakly* to α in $\mathcal{M}_+^1(\mathcal{X})$ (denoted $\alpha_k \rightharpoonup \alpha$) if and only if for any bounded continuous function $f \in \mathcal{C}_b(\mathcal{X})$, $\int_{\mathcal{X}} f \mathrm{d}\alpha_k \to \int_{\mathcal{X}} f \mathrm{d}\alpha$. On compact spaces, $\mathcal{C}_b(\mathcal{X}) = \mathcal{C}(\mathcal{X})$, which is why the boundedness is often left implicit there.

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.
$\|\alpha - \beta\|_{\mathrm{TV}} = |\alpha - \beta|(\mathcal{X})$,
Proposition 3.33 (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 (a, b) and without loss of generality assume they have the same support $(x_i)_i$ and we denote $\mathrm{D} \coloneqq (d(x_i,x_j))_{i,j}$ 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 $\bar{\mathrm{a}} \coloneqq \mathrm{a} - \mathrm{c} = (\mathrm{a}-\mathrm{b})_+ \geqslant 0$ and $\bar{\mathrm{b}} \coloneqq \mathrm{b} - \mathrm{c} = (\mathrm{b}-\mathrm{a})_+ \geqslant 0$ 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 $\hat{\mathrm{P}} \coloneqq \mathrm{diag}(\mathrm{c}) + \frac{\bar{\mathrm{a}}\otimes\bar{\mathrm{b}}}{\langle \bar{\mathrm{a}}, \mathbb{1}\rangle} \geqslant 0$ satisfies $\hat{\mathrm{P}}\mathbb{1} = \mathrm{c} + \bar{\mathrm{a}} = \mathrm{a}$ and $\hat{\mathrm{P}}^\top \mathbb{1} = \mathrm{c} + \bar{\mathrm{b}} = \mathrm{b}$ 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

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

□

$\|\delta_{x_n} - \delta_x\|_{\mathrm{TV}} = 2$ and $\mathcal{W}_p(\delta_{x_n}, \delta_x) = d(x_n, x)$.
Proposition 3.34 (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 3.31 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 $p \geqslant 1$. □

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

Proposition 3.35 (Comparison with total variation on discrete spaces)**.** *One has*

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

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 π gives the result using Proposition 3.33. □

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

3.7 Wasserstein over Wasserstein

Proposition 3.36 (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. 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 3.34 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.

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

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 3.37 (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.

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

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
$\mathfrak{A} = \sum_i a_i \delta_{\mathcal{N}(m_i, \Sigma_i)}$

Proposition 3.38 (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 (α, β) 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 Π gives a coupling $\bar{\pi}$ between $\bar{\alpha}_{\mathfrak{A}}$ and $\bar{\beta}_{\mathfrak{B}}$. Its cost satisfies $\int_{\mathcal{X}\times\mathcal{X}} d(x, y)^2 \mathrm{d}\bar{\pi}(x, y) \leqslant \int_{\mathcal{P}_2(\mathcal{X})^2} \mathcal{W}_2^2(\alpha, \beta) \mathrm{d}\Pi(\alpha, \beta)$ up to the arbitrary selection error. Taking first the infimum over $\bar{\pi}$ and then over Π proves the claim. □

$\alpha_x = (d_{\mathcal{X}}(x, \cdot))_\sharp \mu_{\mathcal{X}} \in \mathcal{P}(\mathbb{R}_+), \qquad \mathfrak{D}_{\mathcal{X}} = (x \mapsto \alpha_x)_\sharp \mu_{\mathcal{X}} \in \mathcal{P}(\mathcal{P}(\mathbb{R}_+))$.

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

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

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_θ-Lipschitz loss, the Kantorovich–Rubinstein dual gives the transparent upper bound
$\sup_{\beta:\, \mathcal{W}_1(\beta, \hat{\alpha}_n) \leqslant \rho} \int \ell_\theta \mathrm{d}\beta \leqslant \frac{1}{n}\sum_i \ell_\theta(z_i) + \rho L_\theta$,

Proposition 3.39 (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 η-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, $F(t) \coloneqq \mathcal{W}_p((1-t)\delta_0 + t\delta_1, \delta_0) = t^{1/p}$ satisfies $F(1/2) > (F(0) + F(1))/2$. □

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

$\mathcal{W}_\infty$ robustness.

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

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.

Proposition 3.40 ($\mathcal{W}_\infty$ 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 ℓ, $\sup_{\beta:\, \mathcal{W}_\infty(\beta,\hat{\alpha}) \leqslant \rho} \int \ell(z) \mathrm{d}\beta(z) = \sum_{i=1}^n a_i \sup_{z \in \overline{B}(z_i,\rho)} \ell(z)$.*

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 π with respect to the first marginal gives $\pi = \sum_i a_i \delta_{z_i} \otimes \nu_i$, where each ν_i is supported in the closed ball $\overline{B}(z_i, \rho)$ and $\beta = \sum_i a_i \nu_i$. Hence $\int \ell \,\mathrm{d}\beta = \sum_i a_i \int \ell \,\mathrm{d}\nu_i \leqslant \sum_i a_i \sup_{\overline{B}(z_i,\rho)} \ell$. The reverse inequality follows by choosing, for each i, a maximizer $z_i^\star \in \overline{B}(z_i, \rho)$ and setting $\beta = \sum_i a_i \delta_{z_i^\star}$. The coupling $\sum_i a_i \delta_{(z_i, z_i^\star)}$ has essential displacement at most ρ, so this β is feasible and attains the displayed value. □

3.9 Quantitative Central Limit Theorems

Proposition 3.41 (Berry–Esseen bound in $\mathcal{W}_1$)**.** *Let $(X_i)_{i=1}^n$ be i.i.d. real random variables such that $\mathbb{E}X_i = 0$, $\mathbb{E}X_i^2 = 1$ and $\mathbb{E}|X_i|^3 < +\infty$. If α_n is the law of $n^{-1/2}\sum_i X_i$ and γ is the standard Gaussian law, then $\mathcal{W}_1(\alpha_n, \gamma) \leqslant \frac{C\,\mathbb{E}|X_1|^3}{\sqrt{n}}$, where C is a universal constant.*

Proof. By Kantorovich–Rubinstein duality,

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

For each such h, solve Stein's equation $f_h'(x) - x f_h(x) = h(x) - \mathbb{E}h(G)$. The solution satisfies 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. □

4 Dual Problem

4.1 Discrete dual

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 $\mathrm{C} \in \mathbb{R}^{n\times m}$, a pair $(\mathrm{f}, \mathrm{g}) \in \mathbb{R}^n \times \mathbb{R}^m$ is admissible if it lies below the cost:

$$\mathrm{R}(\mathrm{a}, \mathrm{b}) \coloneqq \{(\mathrm{f}, \mathrm{g}) \in \mathbb{R}^n \times \mathbb{R}^m \,;\, \forall\, (i, j) \in [\![n]\!] \times [\![m]\!], \mathrm{f}_i + \mathrm{g}_j \leqslant \mathrm{C}_{i,j}\}\,. \tag{4.1}$$

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

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

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

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

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

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

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

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

$$\mathrm{Supp}(\mathrm{P}) \subset \{(i, j) \in [\![n]\!] \times [\![m]\!] \,;\, \mathrm{f}_i + \mathrm{g}_j = \mathrm{C}_{i,j}\}\,. \tag{4.4}$$

4.2 Auction Algorithm and Dual Prices

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

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

For fixed prices p, eliminating the bidder utilities u_i in the dual minimization gives the convex objective
$D(p) = \sum_j p_j + \sum_i \max_j(a_{i,j} - p_j)$,

Definition 4.3 (ε-complementary slackness)**.** An assignment σ and prices p satisfy ε-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 σ satisfies ε-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 σ is optimal.*

Proof. Let τ be any assignment. By ε-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 σ and τ 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 σ. Since $a = -C$, the same statement says that the cost of σ 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. □

4.3 General formulation

Proposition 4.5 (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

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

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. □

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

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 $g = f^c$.

4.4 c-transforms

Best-response potentials and the c-transform. $\sup_{g\in\mathcal{C}(\mathcal{Y})} \left\{\int g\mathrm{d}\beta \; ; \; \forall\,(x,y), g(y) \leqslant c(x,y) - f(x)\right\}.$ $\forall\, y \in \mathcal{Y}, \quad g(y) \leqslant f^c(y)$

Definition 4.6 (c-transform)**.** For a function $f : \mathcal{X} \to \mathbb{R} \cup \{-\infty\}$, its c-transform is

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

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

Proposition 4.7 (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 $g = f^c$ β-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 $f = g^{\bar{c}}$ α-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 β is nonnegative, the largest possible value of $\int g\mathrm{d}\beta$ is obtained by saturating this pointwise upper bound on the support of β. The proof for $f = g^{\bar{c}}$ is identical after exchanging the two marginals.
For the primal formula, disintegrate any feasible π 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 π_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.

Proposition 4.8 (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,

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

□

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

The failure of alternate optimization.

Proposition 4.9 (Algebra of c-transforms)**.**

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

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 $f^c \leqslant f^{c\bar{c}c}$.

□

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

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.

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

5.1 Semi-dual

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

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

5.2 Semi-discrete

Discrete target and Laguerre cells.

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

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

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

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

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.

Mass balance.

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

Proposition 5.2 (Gradient of the semi-discrete dual)**.** *If α 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 α-almost every x, the minimizing index in $\min_j c(x, y_j) - \mathrm{g}_j$ is unique. If this index is $j(x)$, then the directional derivative in a direction $h \in \mathbb{R}^m$ is

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

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. □

Stochastic optimization.

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

$$\nabla_{\mathrm{g}} E(\mathrm{g}, x) = \left(\mathrm{b}_j - \mathbb{1}_{\mathbb{L}_j(\mathrm{g})}(x)\right)_{j=1}^{m},$$

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

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

where ℓ_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.

5.3 Optimal Quantization

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

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 Ω, 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 Ω 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 $\int d_Y(x)^p \mathrm{d}\alpha(x) = \int_0^{+\infty} pt^{p-1}\alpha(\{d_Y > t\})\mathrm{d}t \geqslant ct_0^p \simeq cm^{-p/d}$. Taking the p-th root and minimizing over Y proves the lower bound. □

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 $\mathbb{V}_j(Y) \coloneqq \{x\ ;\ \forall\, j',\ c(x, y_j) \leqslant c(x, y_{j'})\}$, up to arbitrary tie-breaking on common boundaries.

Proof. For any coupling between α and a measure supported on Y, the conditional destination of a point x belongs to Y, hence its conditional cost is at least $\min_j c(x, y_j)$. Integrating gives the lower bound. Conversely, choose a measurable nearest-codepoint map $T_Y(x) \in \operatorname{argmin}_j c(x, y_j)$, breaking ties measurably, and set $\mathrm{b}_j = \alpha(T_Y^{-1}(y_j))$. Then $(T_Y)_\sharp\alpha = \sum_j \mathrm{b}_j\delta_{y_j}$ and the induced transport reaches the displayed lower bound. □

$\mathcal{Q}_m(\alpha)^p = \min_Y \mathcal{F}(Y)$, $\qquad \mathcal{F}(Y) \coloneqq \int_{\mathcal{X}} \min_{1\leqslant j\leqslant m} c(x, y_j)\mathrm{d}\alpha(x)$. $y_j \in \operatorname*{argmin}_y \int_{\mathbb{V}_j(Y)} c(x, y)\mathrm{d}\alpha(x)$. $y_j = \frac{\int_{\mathbb{V}_j(Y)} x\mathrm{d}\alpha(x)}{\int_{\mathbb{V}_j(Y)} \mathrm{d}\alpha}$.

5.4 Wasserstein-1 norm

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. □

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

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

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

$\mathcal{W}_1(\alpha,\beta) = \max_{(f_k)_k} \left\{ \sum_k f_k r_k \;;\; \forall\, k,\ |f_{k+1} - f_k| \leqslant z_{k+1} - z_k \right\}.$

$\mathcal{W}_1$ on Euclidean spaces.

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

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

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

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

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 γ joining i to j.

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 a, b on V, set $r = \mathrm{a} - \mathrm{b}$ and orient each edge $e = (i,j)$. If $(\nabla_G f)_e = f_j - f_i$, $\mathrm{div}_G = -\nabla_G^*$ 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 ℓ_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. □

6 Divergences and Dual Norms

6.1 Dual norms (Integral Probability Metrics)

Integral probability metrics.

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

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

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

Proposition 6.2 (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 $\mathcal{C}(\mathcal{X})$ can be uniformly approximated by elements of $\mathrm{Span}(B)$*, then $\|\alpha_n - \alpha\|_B \to 0$ implies $\alpha_n \rightharpoonup \alpha$.*
2. If $B \subset \mathcal{C}(\mathcal{X})$ 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 $|\langle f,\, \alpha_n - \alpha\rangle| \leqslant \|\alpha_n - \alpha\|_B$ for $f \in B$. By linearity, integrals converge for every $h \in \mathrm{Span}(B)$. Let $u \in \mathcal{C}(\mathcal{X})$ and choose $h \in \mathrm{Span}(B)$ with $\|u - h\|_\infty \leqslant \eta$. Since α_n and α are probabilities, $|\langle u,\, \alpha_n - \alpha\rangle| \leqslant |\langle h,\, \alpha_n - \alpha\rangle| + 2\eta$. Taking the limsup as $n \to \infty$ and then letting $\eta \to 0$ gives $\langle u,\, \alpha_n\rangle \to \langle u,\, \alpha\rangle$ for all $u \in \mathcal{C}(\mathcal{X})$, 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 6.3 (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 $\mathcal{C}(\mathcal{X})$ on compact metric spaces.
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 6.2 gives $\mathcal{W}_1(\alpha_n, \alpha) \to 0$. Proposition 3.31 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

Definition 6.4 (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 $r \in \mathbb{R}^n$,

$$\sum_{i,j=1}^{n} r_i r_j k(x_i, x_j) \geqslant 0. \tag{6.2}$$

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

Definition 6.5 (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 ξ with finite kernel energy, the associated norm is

$$\|\xi\|_k^2 \coloneqq \iint_{\mathcal{X}\times\mathcal{X}} k(x,y)\mathrm{d}\xi(x)\mathrm{d}\xi(y). \tag{6.3}$$

For two probability measures, the maximum mean discrepancy associated with k is $\mathrm{MMD}_k(\alpha, \beta) \coloneqq \|\alpha - \beta\|_k$.

Proposition 6.6 (Kernel norm as an RKHS dual norm)**.** *Let $\mathcal{H}$ be the RKHS with reproducing kernel k, and assume that the kernel mean embedding $m_\xi \coloneqq \int k(x,\cdot)\mathrm{d}\xi(x)$ is well-defined for the signed measure ξ. Then $\|\xi\|_k = \sup_{\|h\|_{\mathcal{H}}\leqslant 1} \int h(x)\mathrm{d}\xi(x)$, so $\|\cdot\|_k$ is the dual 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 (6.3). □

Proposition 6.7 (Universal kernels metrize weak convergence)**.** *Assume that $\mathcal{X}$ is compact and that the RKHS generated by the continuous kernel k is dense in $\mathcal{C}(\mathcal{X})$ 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 $h \in \mathcal{C}(\mathcal{X})$ and any $\eta > 0$, choose $g \in \mathcal{H}$ with $\|h - g\|_\infty \leqslant \eta$. Since α_n and α are probabilities,

$$\left|\int h\,\mathrm{d}(\alpha_n - \alpha)\right| \leqslant 2\eta + \left|\int g\,\mathrm{d}(\alpha_n - \alpha)\right|,$$

and the last term tends to zero. This proves weak convergence. Conversely, if $\alpha_n \rightharpoonup \alpha$, then $\alpha_n \otimes \alpha_n$, $\alpha_n \otimes \alpha$ and $\alpha \otimes \alpha$ converge weakly on the compact product space. Applying this to the continuous bounded function k in the identity $\mathrm{MMD}_k(\alpha_n, \alpha)^2 = \iint k\,\mathrm{d}\alpha_n\mathrm{d}\alpha_n - 2\iint k\,\mathrm{d}\alpha_n\mathrm{d}\alpha + \iint k\,\mathrm{d}\alpha\mathrm{d}\alpha$ gives convergence to zero. □

In the special case where α is a discrete measure, one thus has the simple expression

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

$$\|\alpha - \beta\|_k^2 = \sum_{i,i'} \mathrm{a}_i \mathrm{a}_{i'} k(x_i, x_{i'}) + \sum_{j,j'} \mathrm{b}_j \mathrm{b}_{j'} k(y_j, y_{j'}) - 2\sum_{i,j} \mathrm{a}_i \mathrm{b}_j k(x_i, y_j). \tag{6.4}$$

6.3 φ-divergences

Definition by density ratios.

Definition 6.8 (Entropy function)**.** A function $\varphi : \mathbb{R} \to \mathbb{R} \cup \{\infty\}$ is an entropy function if it is lower semicontinuous, convex, $\mathrm{dom}\,\varphi \subset [0, \infty[$, and satisfies the feasibility condition $\mathrm{dom}\,\varphi \cap (0, +\infty) \neq \emptyset$. The speed of growth of φ at ∞ is described by

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

If $\varphi'_\infty = \infty$, then φ grows faster than any linear function and φ is said to be *superlinear*.

Definition 6.9 (φ-Divergences)**.** Let φ 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 α with respect to β: this means that α 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

$$\mathcal{D}_\varphi(\alpha|\beta) \coloneqq \int_{\mathcal{X}} \varphi\left(\frac{\mathrm{d}\alpha}{\mathrm{d}\beta}\right)\mathrm{d}\beta + \varphi'_\infty \alpha^\perp(\mathcal{X}) \tag{6.5}$$

if α, β are nonnegative and ∞ otherwise.

$$\alpha = \sum_i \mathrm{a}_i \delta_{x_i} \quad \text{and} \quad \beta = \sum_i \mathrm{b}_i \delta_{x_i} \tag{6.6}$$

$$\mathrm{D}_\varphi(\mathrm{a}|\mathrm{b}) = \sum_{i \in \mathrm{Supp}(\mathrm{b})} \varphi\left(\frac{\mathrm{a}_i}{\mathrm{b}_i}\right) \mathrm{b}_i + \varphi'_\infty \sum_{i \notin \mathrm{Supp}(\mathrm{b})} \mathrm{a}_i, \tag{6.7}$$

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

Proposition 6.10 (Basic properties of φ-divergences)**.** *If φ is an entropy function, then $\mathcal{D}_\varphi$ is jointly 1-homogeneous, convex and weak-* lower semicontinuous in (α, β).*

Proof. 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. $\mathrm{D}_\varphi(\mathrm{a}|\mathrm{b}) = \sum_i \psi(\mathrm{a}_i, \mathrm{b}_i)$, and it is enough to show that ψ 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 $\theta_1 = \frac{\tau v_1}{\tau v_1 + \lambda v_2}$, $\quad \theta_2 = \frac{\lambda v_2}{\tau v_1 + \lambda v_2}$. 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 φ 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 ψ.

□

Proposition 6.11 (Non-negativity of φ-divergences)**.** *Assume that φ is normalized by $\varphi(1) = 0$. For probability distributions $(\alpha, \beta) \in \mathcal{M}_+^1(\mathcal{X})$, one has $\mathcal{D}_\varphi(\alpha|\beta) \geqslant 0$. If φ is strictly convex, then one has $\mathcal{D}_\varphi(\alpha|\beta) = 0$ if and only if $\alpha = \beta$.*

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 ψ from the previous proof, $\mathcal{D}_\varphi(\alpha|\beta) = \int \psi(a,b)\mathrm{d}m$. Since α and β 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 φ 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.

Main families of φ-divergences. $\varphi_\gamma(s) = \frac{s^\gamma - \gamma s + \gamma - 1}{\gamma(\gamma - 1)} \qquad (\gamma \neq 0, 1)$

$$\mathrm{JS}(\alpha,\beta)^2 = \frac{1}{2}\,\mathrm{KL}\left(\alpha \middle| \frac{\alpha+\beta}{2}\right) + \frac{1}{2}\,\mathrm{KL}\left(\beta \middle| \frac{\alpha+\beta}{2}\right),$$

Variational dual formula.

Proposition 6.12 (Dual expression)**.** *A φ-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 φ as*

$$\mathcal{D}_\varphi(\alpha|\beta) = \sup_{f:\mathcal{X}\to\mathbb{R}} \int_{\mathcal{X}} f(x)\mathrm{d}\alpha(x) - \int_{\mathcal{X}} \varphi^{*,\geqslant 0}(f(x))\mathrm{d}\beta(x). \tag{6.8}$$

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

$$\forall\, f \in \mathcal{C}(\mathcal{X}), \quad \mathcal{D}_\varphi^*(f|\beta) = \int_{\mathcal{X}} \varphi^{*,\geqslant 0}(f(x))\mathrm{d}\beta(x). \tag{6.9}$$

Proof. We first consider the superlinear case $\varphi'_\infty = +\infty$, so that $\mathcal{D}_\varphi(\alpha|\beta) = +\infty$ if α does not have a density $\rho \geqslant 0$ with respect to β, $\mathrm{d}\alpha = \rho \mathrm{d}\beta$. 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

Divergence-based adversarial losses.

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

$\min_\theta \max_\xi \int_{\mathcal{Z}} f_\xi(g_\theta(z))\mathrm{d}\zeta(z) - \frac{1}{m}\sum_j \varphi^*(f_\xi(y_j))$. $\varphi_{\mathrm{JS}}(s) = s\log s - (s+1)\log\frac{s+1}{2}, \qquad \varphi_{\mathrm{JS}}^*(u) = -\log(2 - e^u), \quad u < \log 2,$

Dual norms and integral probability metrics.

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

7 Entropic Regularization: Sinkhorn Algorithm

7.1 Entropic Regularization for Discrete Measures

Definition 7.1 (Discrete Shannon–Boltzmann entropy)**.** For a nonnegative matrix P, its Shannon–Boltzmann entropy is $\mathrm{H}(\mathrm{P}) \coloneqq -\sum_{i,j} \mathrm{P}_{i,j}\log(\mathrm{P}_{i,j})$, with the convention $0\log(0)=0$.

$$\mathrm{L}^\varepsilon_{\mathrm{C}}(\mathrm{a},\mathrm{b}) \coloneqq \min_{\mathrm{P}\in\mathrm{U}(\mathrm{a},\mathrm{b})} \langle \mathrm{P},\,\mathrm{C}\rangle - \varepsilon\mathrm{H}(\mathrm{P}). \tag{7.1}$$

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 $\mathrm{U}(\mathrm{a},\mathrm{b})$ is non-empty and compact, and the objective is continuous on it with the convention $0\log 0 = 0$, so a minimizer exists. On the relative interior of the polytope, $-\partial^2\mathrm{H}(\mathrm{P}) = \mathrm{diag}(1/\mathrm{P}_{i,j})$ is positive definite on every non-zero feasible direction. Hence $-\mathrm{H}$ is strictly convex on the polytope and $\langle \mathrm{P},\,\mathrm{C}\rangle - \varepsilon\mathrm{H}(\mathrm{P})$ is strictly convex, which implies uniqueness.
If $\mathrm{a}_i,\mathrm{b}_j>0$ and a minimizer had $\mathrm{P}_{i,j}=0$, then for small $t>0$ the perturbation $\mathrm{P}_t = (1-t)\mathrm{P} + t\,\mathrm{a}\otimes\mathrm{b}$ remains feasible. The directional derivative of the entropic part at $t=0$ is $-\infty$ because the derivative of $r\log r$ at 0 is $-\infty$ along a positive direction. Thus the objective decreases for small t, contradicting optimality. Therefore the minimizer is strictly positive. □

Smoothing effect.

Entropy barriers versus generic LP barriers.

7.2 Sinkhorn's Algorithm

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

$$\forall\,(i,j)\in[\![n]\!]\times[\![m]\!],\quad \mathrm{P}_{i,j} = \mathrm{u}_i\mathrm{K}_{i,j}\mathrm{v}_j \quad \textit{where} \quad \mathrm{K}_{i,j}\coloneqq e^{-\frac{\mathrm{C}_{i,j}}{\varepsilon}}, \tag{7.2}$$

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

Proof. Without loss of generality, we assume $\mathrm{a}_i,\mathrm{b}_j>0$ (otherwise, rows or columns with zero mass are fixed to zero and can be removed). By Proposition 7.2, the minimizer is strictly positive on the remaining support.
We can thus ignore the positivity constraint when introducing two dual variables $\mathrm{f}\in\mathbb{R}^n,\mathrm{g}\in\mathbb{R}^m$ for each marginal constraint so that the Lagrangian of (7.1) reads

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

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

$$\frac{\partial\mathcal{E}(\mathrm{P},\mathrm{f},\mathrm{g})}{\partial\mathrm{P}_{i,j}} = \mathrm{C}_{i,j} + \varepsilon(\log(\mathrm{P}_{i,j})+1) - \mathrm{f}_i - \mathrm{g}_j = 0.$$

which results, in an optimal P coupling of the regularized problem, in the expression $\mathrm{P}_{i,j} = e^{\frac{\mathrm{f}_i+\mathrm{g}_j-\mathrm{C}_{i,j}}{\varepsilon}-1}$ which can be rewritten in the form provided in the proposition using non-negative vectors $\mathrm{u}_i\coloneqq e^{\mathrm{f}_i/\varepsilon-1}$ and $\mathrm{v}_j\coloneqq e^{\mathrm{g}_j/\varepsilon}$. □

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

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

where $\odot$ corresponds to the entry-wise multiplication of vectors.
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

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

initialized with an arbitrary positive vector, for instance $\mathrm{v}^{(0)}=\mathbb{1}_m$. The division operator used above between two vectors is to be understood entry-wise.
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.

7.3 Reformulation using relative entropy

Definition 7.4 (Discrete relative entropy)**.** For nonnegative matrices P,Q of the same size, the generalized relative entropy, or Kullback–Leibler divergence, is

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

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 φ-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.

Proposition 7.5 (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 φ such that $\varphi(1) = 0$, one has indeed by Jensen $\sum_{i,j} \varphi(\mathrm{P}_{i,j}/\mathrm{Q}_{i,j})\mathrm{Q}_{i,j} \geqslant \varphi(\sum_{i,j} \mathrm{P}_{i,j}/\mathrm{Q}_{i,j}\mathrm{Q}_{i,j}) = \varphi(\sum_{i,j} \mathrm{P}_{i,j}) = \varphi(1) = 0$.

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

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

Proposition 7.7 (Convergence with ε)**.** *Assume, after removing zero-mass rows and columns, that* a *and* b *are positive and that* C *is finite. The unique solution* P_ε *of* (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

$$\mathrm{L}_\mathrm{C}^\varepsilon(\mathrm{a},\mathrm{b}) \xrightarrow{\varepsilon\to 0} \mathrm{L}_\mathrm{C}(\mathrm{a},\mathrm{b}).$$

Moreover,

$$\mathrm{P}_\varepsilon \xrightarrow{\varepsilon\to\infty} \mathrm{a}\otimes\mathrm{b}. \tag{7.9}$$

Proof. **Case $\varepsilon \to 0$.** We denote P_ℓ 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, $\mathrm{P}^\star \in \mathrm{U}(\mathrm{a},\mathrm{b})$. Using the equivalent KL-normalized formulation (7.7), optimality of P and P_ℓ 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 ε_ℓ 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 $\mathrm{P}_0^\star$ to this program is unique by strict convexity of $\mathrm{KL}(\cdot|\mathrm{a}\otimes\mathrm{b})$ on the optimal face, one has $\mathrm{P}^\star = \mathrm{P}_0^\star$, 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. □

7.4 General Formulation

$$\mathcal{L}_c^\varepsilon(\alpha,\beta) \coloneqq \min_{\pi\in\mathcal{U}(\alpha,\beta)} \int_{\mathcal{X}\times\mathcal{Y}} c(x,y)\mathrm{d}\pi(x,y) + \varepsilon\,\mathrm{KL}(\pi|\alpha\otimes\beta) \tag{7.11}$$

Definition 7.8 (Relative entropy of measures)**.** For nonnegative measures π and ξ on $\mathcal{X}\times\mathcal{Y}$, the relative entropy is

$$\mathrm{KL}(\pi|\xi) \coloneqq \int_{\mathcal{X}\times\mathcal{Y}} \log\Big(\frac{\mathrm{d}\pi}{\mathrm{d}\xi}(x,y)\Big)\mathrm{d}\pi(x,y) + \int_{\mathcal{X}\times\mathcal{Y}} (\mathrm{d}\xi(x,y) - \mathrm{d}\pi(x,y)). \tag{7.12}$$

By convention, $\mathrm{KL}(\pi|\xi) = +\infty$ if π is not absolutely continuous with respect to ξ.

For fixed balanced marginals, the specific product reference in the entropy only matters up to additive constants, exactly as in Proposition 7.6, provided the alternative reference marginals are mutually absolutely continuous with α and β. Its support and absolute-continuity structure are nevertheless essential: they determine which couplings have finite entropy.

7.5 Path-Space Schrödinger Problem

Unregularized path-space transport. Let $\Omega = C([0,1];\mathcal{X})$ be a path space and denote by
$e_t : \Omega \to \mathcal{X}, \qquad e_t(\omega) = \omega_t \ (e_0)_\sharp M = \alpha, \qquad (e_1)_\sharp M = \beta$

$$\inf_{M\in\mathcal{P}(\Omega)} \left\{ \int_\Omega \mathcal{A}(\omega)\,\mathrm{d}M(\omega) \ ; \ (e_0)_\sharp M = \alpha, \ (e_1)_\sharp M = \beta \right\}. \tag{7.13}$$

For the quadratic Wasserstein geometry on $\mathbb{R}^d$, one takes

$$\mathcal{A}(\omega) = \begin{cases} \displaystyle\int_0^1 \|\dot\omega_t\|^2\,\mathrm{d}t, & \text{if } \omega \text{ is absolutely continuous,} \\ +\infty, & \text{otherwise.} \end{cases}$$

$$c_{\mathcal{A}}(x,y) \coloneqq \inf_{\omega\in\Omega} \{\mathcal{A}(\omega) \ ; \ e_0(\omega) = x, \ e_1(\omega) = y\}. \tag{7.14}$$

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 7.9 (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 π and a measurable selection $(x,y)\mapsto\omega^{x,y}$ with action arbitrarily close to $c_{\mathcal{A}}(x,y)$, the mixture of Dirac path laws $M = \int \delta_{\omega^{x,y}}\,\mathrm{d}\pi(x,y)$ has endpoints α and β and action equal, up to the selected approximation error, to $\int c_{\mathcal{A}}\,\mathrm{d}\pi$. Optimizing over π and letting the approximation error vanish proves equality. If exact minimizing paths are selected for an optimal $\pi^\star$, the displayed $M^\star$ is optimal. □

Entropic path-space problem. Let $\mathcal{R}^\varepsilon\in\mathcal{P}(\Omega)$ be a reference path law, for instance a Brownian or Langevin dynamics at noise level ε. Schrödinger's dynamic problem is the entropy projection

$$\mathrm{SB}_\varepsilon(\alpha,\beta) \coloneqq \inf_{M\in\mathcal{P}(\Omega)} \{\varepsilon\,\mathrm{KL}(M|\mathcal{R}^\varepsilon) \ ; \ (e_0)_\sharp M = \alpha, \ (e_1)_\sharp M = \beta\}. \tag{7.15}$$

Viscous Benamou–Brenier formulations. $\partial_t\rho_t + \mathrm{div}(\rho_t v_t) = \frac{\sigma}{2}\Delta\rho_t$, $\int_0^1\int \frac{1}{2}\|v_t(x)\|^2\rho_t(x)\mathrm{d}x\mathrm{d}t$. $u_t = v_t - \frac{\sigma}{2}\nabla\log\rho_t, \qquad \partial_t\rho_t + \mathrm{div}(\rho_t u_t) = 0.$

$$\int_0^1\int \frac{1}{2}\|v_t\|^2\rho_t\,\mathrm{d}x\,\mathrm{d}t = \int_0^1\int \left(\frac{1}{2}\|u_t\|^2 + \frac{\sigma^2}{8}\|\nabla\log\rho_t\|^2\right)\rho_t\,\mathrm{d}x\,\mathrm{d}t + \frac{\sigma}{2}\left[\int\rho_1\log\rho_1\,\mathrm{d}x - \int\rho_0\log\rho_0\,\mathrm{d}x\right].$$

$$\int_0^1\int \left(\frac{1}{2}\|u_t(x)\|^2 + \frac{\sigma^2}{8}\|\nabla\log\rho_t(x)\|^2\right)\rho_t(x)\mathrm{d}x\mathrm{d}t.$$

Thus the Schrödinger bridge is a least-action interpolation with both transport kinetic energy and a Fisher-information penalty. If one instead writes the viscous equation with diffusion coefficient $\sigma\Delta\rho_t$, the same formula is obtained after replacing σ above by 2σ, so the Fisher coefficient becomes $\sigma^2/2$.

$$\mathcal{R}^\varepsilon(\mathrm{d}\omega) = \int \mathcal{R}^{\varepsilon,x,y}(\mathrm{d}\omega)\,\mathcal{R}^\varepsilon_{01}(\mathrm{d}x,\mathrm{d}y), \qquad \mathcal{R}^\varepsilon_{01} \coloneqq (e_0,e_1)_\sharp\mathcal{R}^\varepsilon,$$

where $\mathcal{R}^{\varepsilon,x,y}$ is the reference bridge conditioned on endpoints (x,y). Similarly, for any feasible M, write
$M(\mathrm{d}\omega) = \int M^{x,y}(\mathrm{d}\omega)\,\pi(\mathrm{d}x,\mathrm{d}y), \qquad \pi \coloneqq (e_0,e_1)_\sharp M.$

Proposition 7.10 (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 π *with finite* $\mathrm{KL}(\pi|\mathcal{R}^\varepsilon_{01})$*, the minimizing path law is the mixture of reference bridges*

$$M^\pi = \int \mathcal{R}^{\varepsilon,x,y}\,\mathrm{d}\pi(x,y). \tag{7.17}$$

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 π-almost every (x,y). Thus, once an endpoint coupling π 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 $+\infty$. □

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 ε. Its endpoint law has a heat-kernel density of the form

$$p_\varepsilon(x,y) \propto \exp\left(-\frac{\|x-y\|^2}{\varepsilon}\right)$$

$$\mathcal{R}^\varepsilon_{01}(\mathrm{d}x,\mathrm{d}y) \propto \exp\left(-\frac{c(x,y)}{\varepsilon}\right)\alpha(\mathrm{d}x)\beta(\mathrm{d}y).$$

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.
$\varepsilon\,\mathrm{KL}(\pi|\mathcal{R}^\varepsilon_{01}) = \int c(x,y)\,\mathrm{d}\pi(x,y) + \varepsilon\,\mathrm{KL}(\pi|\alpha\otimes\beta) + \text{constant}$, where the constant does not depend on π. Hence (7.16) is exactly the continuous Sinkhorn problem (7.11), up to an additive constant in the value.
$M^\star_\varepsilon = \int \mathcal{R}^{\varepsilon,x,y}\,\mathrm{d}\pi^\star_\varepsilon(x,y)$.

Definition 7.11 (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.

$\inf_{X\sim\alpha,\,Y\sim\beta} \mathbb{E}\big(c(X,Y)\big) + \varepsilon\mathcal{I}(X,Y)$.

7.6 Dual of Sinkhorn

Discrete dual.

Proposition 7.12 (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 (f,g) *are linked to scalings* (u,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 (f,g), giving

$$\max_{\mathrm{f},\mathrm{g}} \langle \mathrm{f},\,\mathrm{a}\rangle + \langle \mathrm{g},\,\mathrm{b}\rangle + \varepsilon\min_{\mathrm{P}\geqslant 0}\left(\mathrm{KL}(\mathrm{P}|\mathrm{a}\otimes\mathrm{b}) - \langle\frac{\mathrm{f}\oplus\mathrm{g}-\mathrm{C}}{\varepsilon},\,\mathrm{P}\rangle\right) = \langle \mathrm{f},\,\mathrm{a}\rangle + \langle \mathrm{g},\,\mathrm{b}\rangle - \varepsilon\mathrm{KL}^*\left(\frac{\mathrm{f}\oplus\mathrm{g}-\mathrm{C}}{\varepsilon}|\mathrm{a}\otimes\mathrm{b}\right).$$

One concludes by using (6.9) 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.

$$\mathrm{a}_i - e^{\frac{\mathrm{f}_i}{\varepsilon}}\mathrm{a}_i\sum_j \exp\left(\frac{\mathrm{g}_j-\mathrm{C}_{i,j}}{\varepsilon}\right)\mathrm{b}_j = 0$$

$$\mathrm{f}_i = -\varepsilon\log\sum_j \exp\left(\frac{\mathrm{g}_j-\mathrm{C}_{i,j}}{\varepsilon}\right)\mathrm{b}_j.$$

Definition 7.13 (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 $\min^\varepsilon_{\mathrm{b}}(h) := -\varepsilon\log\sum_j e^{-h_j/\varepsilon}\mathrm{b}_j$. It converges to $\min_j h_j$ as $\varepsilon\to 0$. Given a cost matrix C, the discrete soft c-transforms are

$$\mathrm{f}_i = \min^\varepsilon_{\mathrm{b}}(\mathrm{C}_{i,\cdot} - \mathrm{g}) \tag{7.21}$$

and

$$\mathrm{g}_j = \min^\varepsilon_{\mathrm{a}}(\mathrm{C}_{\cdot,j} - \mathrm{f}). \tag{7.22}$$

$\min^\varepsilon_{\mathrm{b}}(h - \mathrm{cst}) = \min^\varepsilon_{\mathrm{b}}(h) - \mathrm{cst}$

Continuous dual and soft-transforms. For general, not necessarily discrete, measures (α,β), the KL-regularized problem (7.11) has the concave dual

$$\mathcal{L}^\varepsilon_c(\alpha,\beta) = \sup_{f\in\mathcal{C}(\mathcal{X}),\,g\in\mathcal{C}(\mathcal{Y})} \mathcal{D}_\varepsilon(f,g), \tag{7.23}$$

$$\mathcal{D}_\varepsilon(f,g) := \int_{\mathcal{X}} f(x)\mathrm{d}\alpha(x) + \int_{\mathcal{Y}} g(y)\mathrm{d}\beta(y) - \varepsilon\int_{\mathcal{X}\times\mathcal{Y}}\left(e^{\frac{f(x)+g(y)-c(x,y)}{\varepsilon}} - 1\right)\mathrm{d}\alpha(x)\mathrm{d}\beta(y). \tag{7.24}$$

Definition 7.14 (Continuous soft c-transforms)**.** For $f \in \mathcal{C}(\mathcal{X})$ and $g \in \mathcal{C}(\mathcal{Y})$, define

$$f^{c,\varepsilon}(y) := -\varepsilon \log\left(\int_{\mathcal{X}} e^{\frac{f(x)-c(x,y)}{\varepsilon}} \mathrm{d}\alpha(x)\right), \qquad y \in \mathcal{Y}, \tag{7.25}$$

$$g^{\bar{c},\varepsilon}(x) := -\varepsilon \log\left(\int_{\mathcal{Y}} e^{\frac{g(y)-c(x,y)}{\varepsilon}} \mathrm{d}\beta(y)\right), \qquad x \in \mathcal{X}. \tag{7.26}$$

Proposition 7.15 (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 $(f^\star + \lambda, g^\star - \lambda)$, $\lambda \in \mathbb{R}$.*

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 $|g^{\bar{c},\varepsilon}(x) - g^{\bar{c},\varepsilon}(x')| \leqslant \sup_y |c(x, y) - c(x', y)|$. 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. □

7.7 Other Convex Regularizers

Let φ be an entropy function in the sense of Definition 6.8, and recall the φ-divergence $\mathcal{D}_\varphi$ from Definition 6.9. For $\varepsilon > 0$, define the φ-regularized transport value

$$\mathcal{L}^{\varepsilon}_{c,\varphi}(\alpha, \beta) := \min_{\pi \in \mathcal{U}(\alpha,\beta)} \int_{\mathcal{X}\times\mathcal{Y}} c(x, y) \mathrm{d}\pi(x, y) + \varepsilon \mathcal{D}_\varphi(\pi | \alpha \otimes \beta). \tag{7.27}$$

Proposition 7.16 (Dual and density law for φ-regularized OT)**.** *Under the usual Fenchel–Rockafellar qualification assumptions, for instance compact spaces, continuous c, and finite value in (7.27), one has*

$$\mathcal{L}^{\varepsilon}_{c,\varphi}(\alpha, \beta) = \sup_{f \in \mathcal{C}(\mathcal{X}),\, g \in \mathcal{C}(\mathcal{Y})} \int f \mathrm{d}\alpha + \int g \mathrm{d}\beta - \varepsilon \int \varphi^{*,\geqslant 0}\left(\frac{f(x) + g(y) - c(x, y)}{\varepsilon}\right) \mathrm{d}\alpha(x) \mathrm{d}\beta(y). \tag{7.28}$$

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 $\frac{f^\star(x)+g^\star(y)-c(x,y)}{\varepsilon} \in \partial\varphi(r^\star(x, y))$ $\alpha \otimes \beta$-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 π gives

$$\int f \mathrm{d}\alpha + \int g \mathrm{d}\beta + \inf_{\pi \in \mathcal{M}_+(\mathcal{X}\times\mathcal{Y})} \left\{ \int \big(c - (f \oplus g)\big) \mathrm{d}\pi + \varepsilon \mathcal{D}_\varphi(\pi | \alpha \otimes \beta) \right\}.$$

Using the Legendre formula (6.9) 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 φ' gives the density law when φ is differentiable and the optimizer is in the interior of its domain. □

For the KL entropy $\varphi(r) = r \log r - r + 1$, one has $\varphi^{*,\geqslant 0}(s) = e^s - 1$. Taking this parameter to be the Sinkhorn temperature ε in (7.28) recovers exactly the continuous Sinkhorn dual (7.24). Other choices replace the exponential law by another scalar transfer function:

$$\begin{aligned} &\varphi(r) = r \log r - r + 1 \quad &\Rightarrow \quad &r^\star = e^s, \\ &\varphi(r) = r - \log r - 1 \quad &\Rightarrow \quad &r^\star = (1 - s)^{-1} \quad (s < 1), \qquad s := \frac{f^\star \oplus g^\star - c}{\varepsilon}. \\ &\varphi(r) = \tfrac{1}{2}(r - 1)^2 \quad &\Rightarrow \quad &r^\star = (1 + s)_+, \end{aligned}$$

Bregman vs. φ-divergence regularization.

Definition 7.17 (Measure Bregman divergence)**.** If Φ is a differentiable convex functional on a convex class of nonnegative measures and ξ is a reference measure in its domain, the measure Bregman divergence generated by Φ is

$$B_\Phi(\pi | \xi) := \Phi(\pi) - \Phi(\xi) - \int \delta\Phi(\xi) \mathrm{d}(\pi - \xi), \tag{7.29}$$

where $\delta\Phi(\xi)$ is the first variation and the formula is understood whenever the right-hand side is well-defined.

Proposition 7.18 (Dual comparison: Bregman vs. density-ratio penalties)**.** *Fix the marginals α, β and set $\xi := \alpha \otimes \beta$. Let Φ 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 π, as $\varepsilon\Phi(\pi) + \int \big(c - \varepsilon\delta\Phi(\xi)\big)\mathrm{d}\pi - \varepsilon\Phi(\xi) + \varepsilon\int \delta\Phi(\xi)\mathrm{d}\xi$. Introduce dual potentials (f,g) for the two marginal constraints. The inner minimization over nonnegative measures π gives

$$\inf_{\pi\geqslant 0}\left\{\varepsilon\Phi(\pi) + \int \big(c - f\oplus g - \varepsilon\delta\Phi(\xi)\big)\mathrm{d}\pi\right\} = -\varepsilon\Phi^*\left(\delta\Phi(\xi) + \frac{f\oplus g - c}{\varepsilon}\right).$$

Fenchel equality at ξ gives $-\Phi(\xi) + \int \delta\Phi(\xi)\mathrm{d}\xi = \Phi^*(\delta\Phi(\xi))$, which yields (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 7.16. Placing the two formulas side by side shows the structural difference: Bregman regularization translates the reference measure in the dual coordinate $\delta\Phi$, whereas φ-regularization applies a scalar nonlinearity to the density with respect to the moving product reference $\alpha\otimes\beta$. □

When Φ is separable with respect to a fixed dominating measure ξ_0, say $\Phi(\pi) = \int h(\mathrm{d}\pi/\mathrm{d}\xi_0)\mathrm{d}\xi_0$ with $\xi \ll \xi_0$, the Bregman optimality condition becomes

$$h'\left(\frac{\mathrm{d}\pi^\star}{\mathrm{d}\xi_0}\right) = h'\left(\frac{\mathrm{d}\xi}{\mathrm{d}\xi_0}\right) + \frac{f^\star\oplus g^\star - c}{\varepsilon}.$$

This is an additive update in entropy coordinates. The density-ratio formulation instead uses the scalar law associated with φ relative to $\alpha\otimes\beta$.

Proposition 7.19 (KL is the common Bregman and φ case)**.** *Let* ω *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* Φ *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* α,β, *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 $B_\Phi = c\,\mathrm{KL}$. Since $\mathcal{D}_\varphi = c\,\mathrm{KL}$, the function $g(t) = \varphi(t) - c\,t\log t$ generates the zero φ-divergence. Testing on densities for which the ratio p/q takes two values $x < 1 < y$, with weights chosen so that the mean ratio is 1, gives $(y-1)g(x) + (1-x)g(y) = 0$. 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.

7.8 Sinkhorn Divergences

Entropic bias. $\alpha_\varepsilon = \operatorname{argmin}_\beta \mathcal{L}^{\varepsilon}_c(\alpha,\beta)$

Proposition 7.20 (Large-temperature entropic bias)**.** *Assume that* c *is bounded and continuous. Then* $\mathcal{L}^{\varepsilon}_c(\alpha,\beta) \to \iint c(x,y)\mathrm{d}\alpha(x)\mathrm{d}\beta(y)$ *as* $\varepsilon\to+\infty$.

Proof. Let $(f_\varepsilon, g_\varepsilon)$ be optimal dual potentials, normalized by $\int g_\varepsilon\mathrm{d}\beta = 0$. The soft c-transform equation gives

$$f_\varepsilon(x) = -\varepsilon\log\int \exp\left(\frac{g_\varepsilon(y) - c(x,y)}{\varepsilon}\right)\mathrm{d}\beta(y).$$

For bounded c, the oscillations of normalized entropic potentials are bounded uniformly in ε by the oscillation of c. Hence the log-sum-exp expansion is uniform: $f_\varepsilon(x) = -\int(g_\varepsilon(y) - c(x,y))\mathrm{d}\beta(y) + O(\varepsilon^{-1}) = \int c(x,y)\mathrm{d}\beta(y) + O(\varepsilon^{-1})$. At optimality the exponential penalty in the dual integrates to zero, so $\mathcal{L}^{\varepsilon}_c(\alpha,\beta) = \int f_\varepsilon\mathrm{d}\alpha + \int g_\varepsilon\mathrm{d}\beta = \iint c(x,y)\mathrm{d}\alpha(x)\mathrm{d}\beta(y) + O(\varepsilon^{-1})$. This proves the limit. □

$$\alpha_\varepsilon \to \min_\beta \langle \int c(x,\cdot)\mathrm{d}\alpha(x),\, \beta\rangle = \delta_{y^\star(\alpha)} \quad \text{where} \quad y^\star(\alpha) = \operatorname*{argmin}_y \int c(x,y)\mathrm{d}\alpha(x).$$

Sinkhorn divergences.

Definition 7.21 (Sinkhorn divergence)**.** For $\varepsilon > 0$, the debiased Sinkhorn divergence associated with the entropic OT value $\mathcal{L}_c^\varepsilon$ is

$$\bar{\mathcal{L}}_c^\varepsilon(\alpha, \beta) \coloneqq \mathcal{L}_c^\varepsilon(\alpha, \beta) - \frac{1}{2}\mathcal{L}_c^\varepsilon(\alpha, \alpha) - \frac{1}{2}\mathcal{L}_c^\varepsilon(\beta, \beta). \tag{7.31}$$

Lemma 7.22 (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. □

Proposition 7.23 (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_ε (equivalently the Sinkhorn fixed point on f_ε) reads

$$\begin{aligned} f_\varepsilon &= -\varepsilon \log \int \exp\left(\frac{g_\varepsilon(y) - c(\cdot, y)}{\varepsilon}\right) \mathrm{d}\beta(y) = -\varepsilon \log \int \left(1 + \frac{g_\varepsilon(y) - c(\cdot, y)}{\varepsilon} + o(1/\varepsilon)\right) \mathrm{d}\beta(y) \\ &= -\varepsilon \log \left(1 + \frac{1}{\varepsilon} \int (g_\varepsilon(y) - c(\cdot, y)) \mathrm{d}\beta(y) + o(1/\varepsilon)\right) = -\int g_\varepsilon \mathrm{d}\beta + \int c(\cdot, y) \mathrm{d}\beta(y) + o(1). \end{aligned}$$

Plugging this relation in the dual expression (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 (α, β), (α, α) and (β, β) gives

$$\bar{\mathcal{L}}_c^\varepsilon(\alpha, \beta) \to \int c \,\mathrm{d}\alpha \otimes \mathrm{d}\beta - \frac{1}{2} \int c \,\mathrm{d}\alpha \otimes \mathrm{d}\alpha - \frac{1}{2} \int c \,\mathrm{d}\beta \otimes \mathrm{d}\beta = -\frac{1}{2} \int c \,\mathrm{d}(\alpha - \beta) \otimes \mathrm{d}(\alpha - \beta).$$

□

Proposition 7.24 (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 α and β. We denote by $f_{\alpha,\alpha} = g_{\alpha,\alpha}$ (one can assume they are equal by symmetry) the solution for the problem between α and itself.
Using the suboptimal function $(f_{\alpha,\alpha}, g_{\beta,\beta})$ in the dual maximization problem, and using relation (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 β, so the previous inequality equivalently reads

$$\frac{1}{\varepsilon}\bar{\mathcal{L}}_c^\varepsilon(\alpha, \beta) \geqslant 1 - \langle e^{\frac{f_{\alpha,\alpha} \oplus g_{\beta,\beta} - c}{\varepsilon}},\, \alpha \otimes \beta\rangle = 1 - \langle \tilde{\alpha},\, \tilde{\beta}\rangle_k$$

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, $\langle \tilde{\alpha},\, \tilde{\beta}\rangle_k \coloneqq \int k(x, y) \mathrm{d}\tilde{\alpha}(x) \mathrm{d}\tilde{\beta}(y)$. The self Sinkhorn fixed point equation, once exponentiated, reads pointwise $e^{f_{\alpha,\alpha}(x)/\varepsilon} \int k(x, y) \mathrm{d}\tilde{\alpha}(y) = 1$ for α-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 $1 - \langle \tilde{\alpha},\, \tilde{\beta}\rangle_k \geqslant 0$. □

8 Entropic Regularization: Convergence

8.1 Sinkhorn Convergence: Bregman View

Alternating KL projections.

Definition 8.1 (Bregman divergence)**.** Let Φ be a differentiable strictly convex function on a convex domain Ω. The Bregman divergence generated by Φ is $B_\Phi(P|Q) \coloneqq \Phi(P) - \Phi(Q) - \langle \nabla\Phi(Q),\, P - Q\rangle$. 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.

Linear tilts and Gibbs references.

Proposition 8.2 (Linear tilts of Bregman penalties)**.** *Let Φ be differentiable and strictly convex, and let B_Φ be its Bregman divergence. Fix a reference point Q in the interior of the domain. Assume that there exists Q^{C} such that $\nabla\Phi(\mathrm{Q}^{\mathrm{C}}) = \nabla\Phi(\mathrm{Q}) - \mathrm{C}/\varepsilon$. Then, for all 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 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 ε 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 $\mathrm{K}^\varepsilon_{\mathrm{a},\mathrm{b}} := (\mathrm{a} \otimes \mathrm{b}) \odot e^{-\mathrm{C}/\varepsilon}$. $\langle \mathrm{P}, \mathrm{C}\rangle + \varepsilon \mathrm{KL}(\mathrm{P}|\mathrm{a} \otimes \mathrm{b}) = \varepsilon \mathrm{KL}(\mathrm{P}|\mathrm{K}^\varepsilon_{\mathrm{a},\mathrm{b}}) + \mathrm{cst}$. Thus the unique solution P_ε of (7.1) is the KL projection of the tilted Gibbs reference onto $\mathrm{U}(\mathrm{a}, \mathrm{b})$:

$$\mathrm{P}_\varepsilon = \mathrm{Proj}^{\mathrm{KL}}_{\mathrm{U}(\mathrm{a},\mathrm{b})}(\mathrm{K}^\varepsilon_{\mathrm{a},\mathrm{b}}) := \operatorname*{argmin}_{\mathrm{P} \in \mathrm{U}(\mathrm{a},\mathrm{b})} \mathrm{KL}(\mathrm{P}|\mathrm{K}^\varepsilon_{\mathrm{a},\mathrm{b}}). \tag{8.1}$$

Cyclic projection convergence.

Proposition 8.3 (Cyclic Bregman projections on affine constraints)**.** *Let Φ be a Legendre strictly convex generator on a finite-dimensional convex domain, and let $\mathcal{C}_1, \mathcal{C}_2$ be affine constraint sets whose intersection meets the domain. Define $P_{k+1} = \mathrm{Proj}^{B_\Phi}_{\mathcal{C}_2} \mathrm{Proj}^{B_\Phi}_{\mathcal{C}_1}(P_k)$, 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 $\mathcal{C}_1 \cap \mathcal{C}_2$. 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 $P^+ = \mathrm{Proj}^{B_\Phi}_{\mathcal{C}}(P)$ and $\mathcal{C}$ is affine, the first-order optimality condition for minimizing $R \mapsto B_\Phi(R|P)$ over $R \in \mathcal{C}$ is $\langle \nabla\Phi(P^+) - \nabla\Phi(P), R - P^+\rangle = 0 \qquad \forall R \in \mathcal{C}$, because $R - P^+$ ranges over the tangent linear space of $\mathcal{C}$. Taking $R = Q \in \mathcal{C}$ cancels the last term and gives $B_\Phi(Q|P) = B_\Phi(Q|P^+) + B_\Phi(P^+|P) \qquad \forall Q \in \mathcal{C}$.
Let $(Z_\ell)_\ell$ be the half-step sequence obtained by alternating projections onto $\mathcal{C}_1$ and $\mathcal{C}_2$, so that $Z_{2k} = P_k$ and $Z_{2k+2} = P_{k+1}$. Fix $Q \in \mathcal{C}_1 \cap \mathcal{C}_2$. 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 Φ 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 $\mathcal{C}_1$ and $\mathcal{C}_2$, every cluster point belongs to $\mathcal{C}_1 \cap \mathcal{C}_2$.
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_ℓ gives $\nabla\Phi(\bar{P}) - \nabla\Phi(P_0) \in N_{\mathcal{C}_1} + N_{\mathcal{C}_2} = N_{\mathcal{C}_1 \cap \mathcal{C}_2}$, 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 $R \in \mathcal{C}_1 \cap \mathcal{C}_2$. 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. □

$$\mathcal{C}^1_{\mathrm{a}} := \{\mathrm{P} \;;\; \mathrm{P}\mathbb{1}_m = \mathrm{a}\} \quad \text{and} \quad \mathcal{C}^2_{\mathrm{b}} := \left\{\mathrm{P} \;;\; \mathrm{P}^\top \mathbb{1}_n = \mathrm{b}\right\}$$

$$\mathrm{P}^{(\ell+1)} := \mathrm{Proj}^{\mathrm{KL}}_{\mathcal{C}^1_{\mathrm{a}}}(\mathrm{P}^{(\ell)}) \quad \text{and} \quad \mathrm{P}^{(\ell+2)} := \mathrm{Proj}^{\mathrm{KL}}_{\mathcal{C}^2_{\mathrm{b}}}(\mathrm{P}^{(\ell+1)}). \tag{8.2}$$

Row and column scalings.

Proposition 8.4 (KL projections are scalings)**.** *One has* $\mathrm{Proj}^{\mathrm{KL}}_{\mathcal{C}^1_{\mathrm{a}}}(\mathrm{P}) = \mathrm{diag}\left(\frac{\mathrm{a}}{\mathrm{P}\mathbb{1}_m}\right)\mathrm{P}$ *and* $\mathrm{Proj}^{\mathrm{KL}}_{\mathcal{C}^2_{\mathrm{b}}}(\mathrm{P}) = \mathrm{P}\,\mathrm{diag}\left(\frac{\mathrm{b}}{\mathrm{P}^\top \mathbb{1}_n}\right)$.

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 $\log(p/q) + \lambda\mathbb{1} = 0 \quad \Longrightarrow \quad p = uq$ where $u = e^{-\lambda} > 0$. 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 $p = sq/\sum_i q_i$. □

$$\mathrm{P}^{(2\ell)} := \mathrm{diag}(\mathrm{u}^{(\ell)})\mathrm{K}\,\mathrm{diag}(\mathrm{v}^{(\ell)}),$$

$$\begin{aligned} \mathrm{P}^{(2\ell+1)} &:= \mathrm{diag}(\mathrm{u}^{(\ell+1)})\mathrm{K}\,\mathrm{diag}(\mathrm{v}^{(\ell)}) \\ \text{and} \quad \mathrm{P}^{(2\ell+2)} &:= \mathrm{diag}(\mathrm{u}^{(\ell+1)})\mathrm{K}\,\mathrm{diag}(\mathrm{v}^{(\ell+1)}) \end{aligned}$$

8.2 Sinkhorn Convergence: Monotone Point of View

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 α-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. □

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.

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, $f \leqslant g \Rightarrow \mathcal{T}(f) \leqslant \mathcal{T}(g)$, $\mathcal{T}(f+\lambda) = \mathcal{T}(f) + \lambda \quad \forall \lambda \in \mathbb{R}$. 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.*

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. □

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. We state it for balanced entropic OT, which is the specialization needed here.

Proposition 8.9 (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 8.10 (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. 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 8.9. 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 8.11 (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 Δ_k be its dual gap. Define the entropy-corrected lower bound $L_{\varepsilon,k} := \mathcal{D}_{\varepsilon,k} - \varepsilon\mathrm{H}(\mathrm{a}) - \varepsilon\mathrm{H}(\mathrm{b})$, $\mathrm{H}(\mathrm{a}) = -\sum_i a_i \log a_i$. Then $0 \leqslant \mathrm{L_C}(\mathrm{a},\mathrm{b}) - L_{\varepsilon,k} \leqslant \varepsilon\log(nm) + \Delta_k$. Under Proposition 8.10, the intermediate condition is $k+1 \geqslant 2CR^2/(\varepsilon\delta)$. With the above choice of ε, 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_ε be the optimum of the entropy-regularized objective $\langle \mathrm{P}, \mathrm{C}\rangle - \varepsilon\mathrm{H}(\mathrm{P})$. Since $0 \leqslant \mathrm{H}(\mathrm{P}) \leqslant \log(nm)$ for any coupling matrix, $\mathrm{L_C}(\mathrm{a},\mathrm{b}) - \varepsilon\log(nm) \leqslant E_\varepsilon \leqslant \mathrm{L_C}(\mathrm{a},\mathrm{b})$. The corrected iterate satisfies $L_{\varepsilon,k} = E_\varepsilon - \Delta_k$, which gives the displayed value bound. The final iteration estimate follows by combining $\Delta_k \leqslant CR^2/(\varepsilon(k+1))$ with the target $\Delta_k \leqslant \delta/2$. □

8.4 Sinkhorn Convergence: Linear Hilbert Metric Rate

Definition 8.12 (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, $\|z\|_V = \max_i z_i - \min_i z_i$.

Proposition 8.13 (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 $\mathbb{R}^n/\operatorname{Span}(\mathbb{1}_n)$, 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 $\mathbb{R}^n/\operatorname{Span}(\mathbb{1}_n)$ is finite-dimensional and all finite-dimensional normed spaces are complete. □

Theorem 8.14 (Birkhoff contraction theorem). *Let* $\mathrm{K} \in \mathbb{R}_{+,*}^{n\times m}$, *then for* $(\mathrm{v},\mathrm{v}') \in (\mathbb{R}_{+,*}^m)^2$

$$d_{\mathcal{H}}(\mathrm{Kv},\mathrm{Kv}') \leqslant \lambda(\mathrm{K}) d_{\mathcal{H}}(\mathrm{v},\mathrm{v}') \text{ where } \begin{cases} \lambda(\mathrm{K}) \coloneqq \frac{\sqrt{\eta(\mathrm{K})}-1}{\sqrt{\eta(\mathrm{K})}+1} < 1 \\ \eta(\mathrm{K}) \coloneqq \max\limits_{i,j,k,\ell} \frac{\mathrm{K}_{i,k}\mathrm{K}_{j,\ell}}{\mathrm{K}_{j,k}\mathrm{K}_{i,\ell}}. \end{cases}$$

Proof. For a positive linear map A on a cone, define its projective diameter $\Delta(A) \coloneqq \sup_{u,v>0} d_{\mathcal{H}}(Au, Av)$. 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. □

Theorem 8.15 (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 $\mathrm{P}^{(\ell)} \coloneqq \mathrm{diag}(\mathrm{u}^{(\ell)}) \mathrm{K} \,\mathrm{diag}(\mathrm{v}^{(\ell)})$. *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

$$d_{\mathcal{H}}(\mathrm{u}^{(\ell+1)}, \mathrm{u}^\star) = d_{\mathcal{H}}\left(\frac{\mathrm{a}}{\mathrm{Kv}^{(\ell)}}, \frac{\mathrm{a}}{\mathrm{Kv}^\star}\right) = d_{\mathcal{H}}(\mathrm{Kv}^{(\ell)}, \mathrm{Kv}^\star) \leqslant \lambda(\mathrm{K}) d_{\mathcal{H}}(\mathrm{v}^{(\ell)}, \mathrm{v}^\star).$$

where we used Theorem 8.14. This shows (8.4). One also has, using the triangular inequality,

$$\begin{aligned} d_{\mathcal{H}}(\mathrm{u}^{(\ell)}, \mathrm{u}^\star) &\leqslant d_{\mathcal{H}}(\mathrm{u}^{(\ell+1)}, \mathrm{u}^{(\ell)}) + d_{\mathcal{H}}(\mathrm{u}^{(\ell+1)}, \mathrm{u}^\star) \leqslant d_{\mathcal{H}}\left(\frac{\mathrm{a}}{\mathrm{Kv}^{(\ell)}}, \mathrm{u}^{(\ell)}\right) + \lambda(\mathrm{K}) d_{\mathcal{H}}(\mathrm{u}^{(\ell)}, \mathrm{u}^\star) \\ &= d_{\mathcal{H}}\left(\mathrm{a}, \mathrm{u}^{(\ell)} \odot (\mathrm{Kv}^{(\ell)})\right) + \lambda(\mathrm{K}) d_{\mathcal{H}}(\mathrm{u}^{(\ell)}, \mathrm{u}^\star), \end{aligned}$$

which gives the first part of (8.5) since $\mathrm{u}^{(\ell)} \odot (\mathrm{Kv}^{(\ell)}) = \mathrm{P}^{(\ell)}\mathbb{1}_m$ (the second one being similar).
The proof of (8.6) follows from. □

Dual-potential form of the contraction. $\|f_k - f^\star\|_V = O(\lambda^k)$ and $\|g_k - g^\star\|_V = O(\lambda^k)$

$$\|\log \frac{\mathrm{d}\pi_k}{\mathrm{d}\pi^\star}\|_\infty = \|(f_k - f^\star) \oplus (g_k - g^\star)\|_\infty \leqslant \|f_k - f^\star\|_V + \|g_k - g^\star\|_V$$

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 8.14 and $R \coloneqq \sup c - \inf c$,
$\lambda = \frac{\sqrt{\eta}-1}{\sqrt{\eta}+1} \leqslant \tanh(R/(2\varepsilon)) < 1$. $\|f_k - f^\star\|_V \leqslant \frac{\|\log \frac{\mathrm{d}\pi_{k,1}}{\mathrm{d}\alpha}\|_\infty}{1-\lambda}$

8.5 Entropic Optimal Transport between Gaussians

Proposition 8.16 (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 β. Completing the square in y evaluates the integral as a positive constant times the exponential of a quadratic polynomial in x. Taking $-\varepsilon \log$ therefore gives a quadratic polynomial.
For $g_0 = 0$, let $Y \sim \beta$. The Gaussian identity

$$\mathbb{E} \exp\left(-\frac{\|x-Y\|^2}{\varepsilon}\right) = \det\left(\mathrm{Id} + \frac{2\mathbf{\Sigma}_\beta}{\varepsilon}\right)^{-1/2} \exp\left(-\langle x - \mathbf{m}_\beta, (\varepsilon \mathrm{Id} + 2\mathbf{\Sigma}_\beta)^{-1}(x - \mathbf{m}_\beta)\rangle\right)$$

gives the displayed expression. □

Proposition 8.17 (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 $K_\varepsilon = \mathbf{\Sigma}_\alpha^{1/2} U\,\mathrm{diag}(s_i) V^\top \mathbf{\Sigma}_\beta^{1/2}$, $\quad s_i = \frac{\sqrt{\varepsilon^2+16\sigma_i^2}-\varepsilon}{4\sigma_i}$. The optimal value is*

$$\|\mathbf{m}_\alpha - \mathbf{m}_\beta\|^2 + \mathrm{tr}(\mathbf{\Sigma}_\alpha) + \mathrm{tr}(\mathbf{\Sigma}_\beta) + \sum_i \Big(-2\sigma_i s_i - \frac{\varepsilon}{2}\log(1-s_i^2)\Big).$$

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 $\|\mathbf{m}_\alpha - \mathbf{m}_\beta\|^2 + \mathrm{tr}(\mathbf{\Sigma}_\alpha) + \mathrm{tr}(\mathbf{\Sigma}_\beta) - 2\,\mathrm{tr}(K)$, while $\mathrm{KL}(\pi|\alpha\otimes\beta) = -\frac{1}{2}\log\det(\mathrm{Id} - SS^\top)$. By von Neumann's trace inequality, the minimizer aligns S with the singular vectors of $\mathbf{\Sigma}_\alpha^{1/2}\mathbf{\Sigma}_\beta^{1/2}$, so $S = U\,\mathrm{diag}(s_i)V^\top$. The problem separates into scalar minimizations $\min_{0\leqslant s<1} -2\sigma_i s - \frac{\varepsilon}{2}\log(1-s^2)$. The first-order condition is $2\sigma_i = \varepsilon s/(1-s^2)$, whose positive solution is the displayed s_i. Substitution gives the value formula. Since $s_i \to 1$ and $\varepsilon\log(1-s_i^2) \to 0$ as $\varepsilon \downarrow 0$, the spectral sum converges to $-2\sum_i \sigma_i$. The identity

$$\sum_i \sigma_i = \mathrm{tr}\Big((\mathbf{\Sigma}_\alpha^{1/2}\mathbf{\Sigma}_\beta\mathbf{\Sigma}_\alpha^{1/2})^{1/2}\Big)$$

then gives the Bures–Wasserstein covariance contribution. □

Corollary 8.18 (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 Σ, then the debiased Sinkhorn divergence (7.31) is $\mathcal{L}^\varepsilon_{\|\cdot-\cdot\|^2}(\alpha,\beta) = \|\mathbf{m}_\alpha - \mathbf{m}_\beta\|^2 + \mathcal{B}_\varepsilon(\mathbf{\Sigma}_\alpha, \mathbf{\Sigma}_\beta)^2$, 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 2.28.

Proof. Proposition 8.17 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 (α,α) and (β,β) 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: $\mathrm{tr}\,\mathbf{\Sigma}_\alpha + \mathrm{tr}\,\mathbf{\Sigma}_\beta - \frac{1}{2}(2\,\mathrm{tr}\,\mathbf{\Sigma}_\alpha) - \frac{1}{2}(2\,\mathrm{tr}\,\mathbf{\Sigma}_\beta) = 0$, while the polarization of the squared mean terms leaves exactly $\|\mathbf{m}_\alpha - \mathbf{m}_\beta\|^2$. 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 $\mathrm{tr}\,\Sigma + \mathrm{tr}\,\Lambda - 2\sum_i \sigma_i(\Sigma,\Lambda) = \mathcal{B}(\Sigma,\Lambda)^2$, which is the Bures formula (2.16). □

Proposition 8.19 (One-dimensional Gaussian Sinkhorn rate)**.** *Consider $\alpha = \beta = \mathcal{N}(0,1)$ on $\mathbb{R}$ with $c(x,y) = (x-y)^2$. If a dual potential has the form $g_q(y) = qy^2 + \mathrm{cst}$, then one soft transform has quadratic coefficient $T_\varepsilon(q) = 1 - \frac{1}{1-q+\varepsilon/2}$, $\quad q < 1+\varepsilon/2$, 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 $A_\star^2 - \frac{\varepsilon}{2}A_\star - 1 = 0$, $\quad A_\star := 1 - q_\star + \frac{\varepsilon}{2} = \frac{\varepsilon+\sqrt{\varepsilon^2+16}}{4}$.*

$$\rho_\varepsilon = A_\star^{-4} = \left(\frac{4}{\varepsilon+\sqrt{\varepsilon^2+16}}\right)^4.$$

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 $T_\varepsilon'(q_\star)^2 = A_\star^{-4}$. □

8.6 Sample Complexity

Proposition 8.20 (Empirical OT has dimension-dependent value rates)**.** *Let α and β 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, $\left|\mathcal{W}_p(\hat{\alpha}_n, \hat{\beta}_m) - \mathcal{W}_p(\alpha,\beta)\right| \leqslant \mathcal{W}_p(\hat{\alpha}_n, \alpha) + \mathcal{W}_p(\hat{\beta}_m, \beta)$. 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 8.21 (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 8.22 (Sinkhorn divergences interpolate the rates)**.** *Assume that α and β 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 ε, while the prefactor deteriorates as $\varepsilon \to 0$.

Proof. The proof follows the empirical-process argument. 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. □

9 Generalized Wasserstein Distances

9.1 Unbalanced OT

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 ψ_1, ψ_2 are convex entropy functions. Exact conservation $(\pi_1,\pi_2) = (\alpha,\beta)$ is replaced by a cost for changing the marginals. $\mathrm{UW}_{c,\tau}(\alpha,\beta) = \inf_{\pi \geqslant 0} \int c\mathrm{d}\pi + \tau\mathcal{D}_{\bar{\psi}_1}(\pi_1|\alpha) + \tau\mathcal{D}_{\bar{\psi}_2}(\pi_2|\beta)$.

Proposition 9.1 (Small-transport-scale limit for marginal penalties)**.** *Assume that α,β are finite measures on a compact metric space $\mathcal{X}$, that c is continuous, $c \geqslant 0$, and $c(x,y) = 0$ if and only if $x = y$. Assume also that the marginal divergences are nonnegative, weak-* lower semicontinuous, and have weak-* compact sublevel sets on $\mathcal{M}_+(\mathcal{X})$. Then*

$$\lim_{\tau\downarrow 0} \frac{1}{\tau}\,\mathrm{UW}_{c,\tau}(\alpha,\beta) = \inf_{\rho\in\mathcal{M}_+(\mathcal{X})} \mathcal{D}_{\bar{\psi}_1}(\rho|\alpha) + \mathcal{D}_{\bar{\psi}_2}(\rho|\beta).$$

The right-hand side is the infimal gluing divergence obtained by matching the two measures through a common zero-transport marginal ρ. 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

$$\int \mathfrak{m}_{\bar{\psi}_1,\bar{\psi}_2}(a(x),b(x))\mathrm{d}\lambda(x), \qquad \mathfrak{m}_{\bar{\psi}_1,\bar{\psi}_2}(a,b) \coloneqq \inf_{r\geqslant 0} a\,\bar{\psi}_1(r/a) + b\,\bar{\psi}_2(r/b),$$

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 ρ.

For the lower bound, let $\tau_n \downarrow 0$ and let π_n be almost minimizing plans with bounded scaled values $\tau_n^{-1}\,\mathrm{UW}_{c,\tau_n}(\alpha,\beta)$. Since the divergences are nonnegative, $\int c\mathrm{d}\pi_n = O(\tau_n)$, hence $\int c\mathrm{d}\pi_n \to 0$. The bounded scaled values also put the two marginals in compact divergence sublevel sets. Since a coupling has the same total mass as each marginal, the couplings are tight on

$\mathcal{X}\times\mathcal{X}$. Up to subsequences, $\pi_n \rightharpoonup \pi_0$. Lower semicontinuity of the transport cost yields $\int c\mathrm{d}\pi_0 = 0$, so π_0 is concentrated on the diagonal. Its two marginals are therefore equal to a common measure ρ. 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 ρ 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 ρ decouples into the scalar envelope $\mathfrak{m}_{\bar\psi_1,\bar\psi_2}$. For KL, no singular part is admissible when α and β are dominated by λ. The pointwise objective is $r\log(r/a) - r + a + r\log(r/b) - r + b$. Its optimality condition is $\log(r/a)+\log(r/b) = 0$, hence $r = \sqrt{ab}$, and the minimum is $a + b - 2\sqrt{ab} = (\sqrt{a}-\sqrt{b})^2$. □

Proposition 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 (6.9) 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.

$$\int c(x,y)\mathrm{d}\pi(x,y) + \mathcal{D}_{\psi_1}(\pi_1|\alpha) + \mathcal{D}_{\psi_2}(\pi_2|\beta)$$
$$= \int \left(c(x,y) + \psi_1\left(\frac{\mathrm{d}\pi_1}{\mathrm{d}\alpha}(x)\right)\frac{\mathrm{d}\alpha}{\mathrm{d}\pi_1}(x) + \psi_2\left(\frac{\mathrm{d}\pi_2}{\mathrm{d}\beta}(y)\right)\frac{\mathrm{d}\beta}{\mathrm{d}\pi_2}(y)\right)\mathrm{d}\pi(x,y).$$

$$L_c(r,s) \coloneqq c + r\psi_1(1/r) + s\psi_2(1/s), \tag{9.2}$$

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
$\mathrm{UW}_c(\alpha,\beta) = \inf_{\pi\geqslant 0}\int L_{c(x,y)}(F(x),G(y))\mathrm{d}\pi(x,y) + \psi_1(0)\alpha^\perp(\mathcal{X}) + \psi_2(0)\beta^\perp(\mathcal{Y})$.

$$H_c(r,s) \coloneqq \inf_{\theta>0}\theta L_c(r/\theta, s/\theta), \tag{9.3}$$

$$\mathrm{HW}_c(\alpha,\beta) = \inf_{\pi\geqslant 0}\int H_{c(x,y)}(F(x),G(y))\mathrm{d}\pi(x,y) + \psi_1(0)\alpha^\perp(\mathcal{X}) + \psi_2(0)\beta^\perp(\mathcal{Y}). \tag{9.4}$$

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 π 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 π by the rescaled measure $\tilde\pi = \theta\pi$ and the densities by F/θ and G/θ 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 $\mathfrak{C}[\mathcal{X}] \coloneqq (\mathcal{X}\times\mathbb{R}_+)/\sim$, where all points $(x,0)$ are identified at the apex. For an exponent $p\geqslant 1$, define
$\Delta((x,r),(y,s)) \coloneqq H_{c(x,y)}(r^p,s^p)^{1/p}$.

- $\mathcal{D}_\psi = \mathrm{KL}$, $p=2$, and $c(x,y) = -\log\cos^2(d(x,y)\wedge\pi/2)$ give the Hellinger–Kantorovich or Wasserstein–Fisher–Rao cone metric $\Delta((x,r),(y,s))^2 = r^2+s^2-2rs\cos(d(x,y)\wedge\pi/2)$.
- $\mathcal{D}_\psi = \mathrm{KL}$, $p=2$, and $c(x,y) = d(x,y)^2$ give the Gaussian Hellinger cone metric $\Delta((x,r),(y,s))^2 = r^2+s^2-2rse^{-d(x,y)^2/2}$.
- $\mathcal{D}_\psi = \mathrm{TV}$, $p=1$, and $c(x,y) = d(x,y)$ give the partial-transport cone cost $\Delta((x,r),(y,s)) = r+s-(r\wedge s)(2-d(x,y))_+$.

$$\mathrm{CW}(\alpha,\beta) = \inf_{\gamma\in\mathcal{M}_+(\mathfrak{C}[\mathcal{X}]^2)}\left\{\int \Delta((x,r),(y,s))^p\mathrm{d}\gamma\ ;\ \int r^p\mathrm{d}\gamma_1(\cdot,r) = \alpha, \int s^p\mathrm{d}\gamma_2(\cdot,s) = \beta\right\}.$$

Theorem 9.4 (Cone formulation of unbalanced OT)**.** *One has* UW = HW = CW. *If* Δ *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 γ 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 pth powers are the two density factors appearing in H_c.
If Δ 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. $\inf_{\pi\in\mathcal{M}_+(\mathcal{X}\times\mathcal{Y})} \int c\mathrm{d}\pi + \mathcal{D}_{\psi_1}(\pi_1|\alpha) + \mathcal{D}_{\psi_2}(\pi_2|\beta) + \varepsilon\mathcal{D}_\varphi(\pi|\alpha\otimes\beta)$.

$$\sup_{f,g} -\mathcal{D}^*_{\psi_1}(-f|\alpha) - \mathcal{D}^*_{\psi_2}(-g|\beta) - \varepsilon\mathcal{D}^*_\varphi\left(\frac{f\oplus g - c}{\varepsilon}\middle|\alpha\otimes\beta\right).$$

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

$$\sup_{f,g} -\tau\int(e^{-f/\tau}-1)\mathrm{d}\alpha - \tau\int(e^{-g/\tau}-1)\mathrm{d}\beta - \varepsilon\iint\left(e^{(f\oplus g-c)/\varepsilon}-1\right)\mathrm{d}\alpha\mathrm{d}\beta,$$

$$f(x) = -\frac{\tau\varepsilon}{\tau+\varepsilon}\log\int_{\mathcal{Y}}\exp\left(\frac{g(y)-c(x,y)}{\varepsilon}\right)\mathrm{d}\beta(y),$$
$$g(y) = -\frac{\tau\varepsilon}{\tau+\varepsilon}\log\int_{\mathcal{X}}\exp\left(\frac{f(x)-c(x,y)}{\varepsilon}\right)\mathrm{d}\alpha(x).$$

$$u_i \leftarrow \left(\frac{\mathrm{a}_i}{(Kv)_i}\right)^\rho, \qquad v_j \leftarrow \left(\frac{\mathrm{b}_j}{(K^\top u)_j}\right)^\rho, \qquad \mathrm{P} = \mathrm{diag}(u)K\,\mathrm{diag}(v).$$

9.2 Sliced Wasserstein Distances

Definition 9.5 (Sliced Wasserstein distance)**.** Let σ be the uniform probability measure on the sphere $\mathbb{S}^{d-1}$. The sliced p-Wasserstein distance is

$$\mathrm{SW}_p(\alpha,\beta)^p \coloneqq \int_{\mathbb{S}^{d-1}} \mathcal{W}_p\left((P_\theta)_\sharp\alpha, (P_\theta)_\sharp\beta\right)^p \mathrm{d}\sigma(\theta). \tag{9.5}$$

Proposition 9.6 (Metric properties of sliced Wasserstein)**.** *For $p\geqslant 1$, SW_p is a distance on $\mathcal{P}_p(\mathbb{R}^d)$. Moreover, SW_p metrizes weak convergence together with convergence of the pth moment. Finally, $\mathrm{SW}_p(\alpha,\beta)\leqslant\mathcal{W}_p(\alpha,\beta)$, 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 θ and then Minkowski's inequality in $L^p(\mathbb{S}^{d-1})$.
If $\mathrm{SW}_p(\alpha,\beta)=0$, then $(P_\theta)_\sharp\alpha = (P_\theta)_\sharp\beta$ for almost every direction. By continuity of characteristic functions this holds for all directions, and the Cramér–Wold theorem implies $\alpha=\beta$. This proves separation.
The bound $\mathrm{SW}_p\leqslant\mathcal{W}_p$ follows because P_θ is 1-Lipschitz, so $\mathcal{W}_p((P_\theta)_\sharp\alpha,(P_\theta)_\sharp\beta)\leqslant\mathcal{W}_p(\alpha,\beta)$ for every θ. For $p=2$, using any coupling π between α and β, $\int_{\mathbb{S}^{d-1}}\int|\langle x-y,\,\theta\rangle|^2\mathrm{d}\pi(x,y)\mathrm{d}\sigma(\theta) = \frac{1}{d}\int\|x-y\|^2\mathrm{d}\pi(x,y)$. Optimizing over π gives the sharper inequality.
The weak-convergence statement follows from the same Cramér–Wold mechanism plus the moment condition: convergence in SW_p gives convergence of almost all one-dimensional projections and tightness of the pth moments; conversely, weak convergence with pth-moment convergence implies convergence of projected $\mathcal{W}_p$ distances and dominated convergence on the sphere. □

Definition 9.7 (Max-sliced Wasserstein)**.** The max-sliced distance replaces the average over directions by the most discriminating one: $\mathrm{MaxSW}_p(\alpha,\beta)\coloneqq\sup_{\theta\in\mathbb{S}^{d-1}}\mathcal{W}_p((P_\theta)_\sharp\alpha,(P_\theta)_\sharp\beta)$.

Subspace-sliced variants.

Definition 9.8 (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 9.9 (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 α and β 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. $\pi_\theta=\frac{1}{n}\sum_{i=1}^n\delta_{(x_i,y_{\sigma_\theta(i)})}$ $\mathrm{MSWGG}_2(\alpha,\beta)^2\coloneqq\min_{\theta\in\mathbb{S}^{d-1}}\int\|x-y\|^2\mathrm{d}\pi_\theta(x,y)$. $\mathcal{W}_2^2(\alpha,\beta)\leqslant\int\|x-y\|^2\mathrm{d}\pi_\theta(x,y)$, $\mathcal{W}_2^2(\alpha,\beta)\leqslant\mathrm{MSWGG}_2(\alpha,\beta)^2$.

9.3 Vector Quantiles and Linear Optimal Transport

Vector quantiles. Assume that ρ is absolutely continuous. For a target law μ with finite second moment, its vector quantile relative to ρ is the Brenier map
$T_\mu=\nabla\varphi_\mu$, $(T_\mu)_\sharp\rho=\mu$, $\min_{T_\sharp\rho=\mu}\int\|x-T(x)\|^2\mathrm{d}\rho(x)$.
Linearized Wasserstein coordinates.

$$\alpha\mapsto T_\alpha-\mathrm{Id}\in L^2(\rho;\mathbb{R}^d), \qquad \mathrm{LOT}_\rho(\alpha,\beta)=\|T_\alpha-T_\beta\|_{L^2(\rho)}. \tag{9.6}$$

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 a family $(\alpha_s)_s$ with weights $(\lambda_s)_s$, the linearized barycenter is obtained by averaging maps,
$\bar{T}=\sum_s\lambda_s T_{\alpha_s}$, $\bar{\alpha}_{\mathrm{LOT}}=\bar{T}_\sharp\rho$. 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.

Proposition 9.10 (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 ρ are regular. Then, for α, β in a sufficiently small regular neighborhood of ρ, $\mathcal{W}_2(\alpha,\beta) \leqslant \mathrm{LOT}_\rho(\alpha,\beta)$ and $\mathrm{LOT}_\rho(\alpha,\beta) \leqslant C\,\mathcal{W}_2(\alpha,\beta)^\eta$ 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 α and β. 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. 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

Definition 9.11 (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 $0 \preceq M \preceq N \quad\Longrightarrow\quad \gamma(M) \leqslant \gamma(N)$.

Definition 9.12 (Spectral Wasserstein distance)**.** Let γ be a monotone spectral gauge. For a coupling $\pi \in \mathcal{U}(\alpha,\beta)$, define its displacement covariance $M_\pi := \int (x-y)(x-y)^\top \mathrm{d}\pi(x,y)$. The spectral Wasserstein distance associated with γ is

$$\mathcal{W}_\gamma(\alpha,\beta)^2 := \inf_{\pi \in \mathcal{U}(\alpha,\beta)} \gamma(M_\pi). \tag{9.7}$$

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.

$$\mathcal{W}_{2,A}(\alpha,\beta)^2 := \inf_{\pi \in \mathcal{U}(\alpha,\beta)} \int (x-y)^\top A (x-y) \mathrm{d}\pi(x,y) = \mathcal{W}_2((A^{1/2})_\sharp \alpha, (A^{1/2})_\sharp \beta)^2. \tag{9.8}$$

$$\mathcal{B}_\gamma := \{A \succeq 0 \,;\, \mathrm{tr}(AM) \leqslant \gamma(M) \quad \text{for all } M \succeq 0\}, \tag{9.9}$$

Proposition 9.13 (Robust representation and metric equivalence)**.** *Assume, for simplicity, that the measures are compactly supported and that γ is closed and finite on the positive semidefinite cone. Then $\mathcal{W}_\gamma(\alpha,\beta)^2 = \sup_{A \in \mathcal{B}_\gamma} \mathcal{W}_{2,A}(\alpha,\beta)^2$. If there exist constants $0 < a \leqslant b < +\infty$ such that $a\mathrm{Id} \in \mathcal{B}_\gamma$ and $\mathcal{B}_\gamma \subset \{A \,;\, 0 \preceq A \preceq b\mathrm{Id}\}$, equivalently $a\,\mathrm{tr}(M) \leqslant \gamma(M) \leqslant b\,\mathrm{tr}(M) \qquad (M \succeq 0)$, then $\sqrt{a}\,\mathcal{W}_2(\alpha,\beta) \leqslant \mathcal{W}_\gamma(\alpha,\beta) \leqslant \sqrt{b}\,\mathcal{W}_2(\alpha,\beta)$. In particular, $\mathcal{W}_\gamma$ is a distance. When γ is the restriction of a norm to the positive semidefinite cone, these bounds hold automatically in finite dimension, so $\mathcal{W}_\gamma$ is equivalent to $\mathcal{W}_2$ on measures with finite second moments.*

Proof. Using the polar representation of γ, $\mathcal{W}_\gamma(\alpha,\beta)^2 = \inf_{\pi \in \mathcal{U}(\alpha,\beta)} \sup_{A \in \mathcal{B}_\gamma} \mathrm{tr}(AM_\pi)$. The coupling set is convex and compact for weak convergence under compact support. Since γ is a finite gauge on a finite-dimensional cone and vanishes only at the origin, it is equivalent to the trace norm on the slice $\mathrm{tr}(M) = 1$, 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 9.14 (Subspace robust Wasserstein)**.** For $1 \leqslant k \leqslant d$, the Paty–Cuturi subspace robust Wasserstein distance is

$$\mathrm{SRW}_{2,k}(\alpha,\beta) := \sup_{U^\top U = \mathrm{Id}_k} \mathcal{W}_2((U^\top)_\sharp \alpha, (U^\top)_\sharp \beta) = \sup_{P^2 = P = P^\top,\ \mathrm{tr}(P) = k} \mathcal{W}_{2,P}(\alpha,\beta).$$

$\gamma_k(M) = \sum_{\ell=1}^k \lambda_\ell(M)$, $\mathcal{B}_{\gamma_k} = \{A \,;\, 0 \preceq A \preceq \mathrm{Id},\ \mathrm{tr}(A) \leqslant k\}$. 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.

10 Generalized OT Problems

10.1 OT Barycenters

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.

Given a set of input measures $(\beta_s)_s$ defined on some space $\mathcal{X}$, the barycenter problem becomes

$$\min_{\alpha\in\mathcal{M}_+^1(\mathcal{X})} \sum_{s=1}^{S} \lambda_s \mathcal{L}_c(\alpha,\beta_s). \tag{10.2}$$

Proposition 10.1 (Convexity of the OT cost). *The map $(\alpha,\beta)\mapsto\mathcal{L}_c(\alpha,\beta)$ is convex.*

Proof. Let (α_0,β_0) and (α_1,β_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 $\eta\to0$ gives the claim. □

One-dimensional case.

Proposition 10.2 (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: $\mathcal{C}_{\alpha^\star}^{-1}(r)=\sum_{s=1}^S\lambda_s\mathcal{C}_{\beta_s}^{-1}(r), \qquad r\in[0,1]$.*

Proof. The one-dimensional formula (2.11) gives $\sum_s\lambda_s\mathcal{W}_2^2(\alpha,\beta_s)=\int_0^1\sum_s\lambda_s\left|\mathcal{C}_\alpha^{-1}(r)-\mathcal{C}_{\beta_s}^{-1}(r)\right|^2\mathrm{d}r$. The minimization decouples pointwise in r. For each fixed r, the minimizer of $z\mapsto\sum_s\lambda_s|z-\mathcal{C}_{\beta_s}^{-1}(r)|^2$ is the weighted average $\sum_s\lambda_s\mathcal{C}_{\beta_s}^{-1}(r)$. This function is nondecreasing because it is a positive weighted sum of nondecreasing quantile functions, hence it is a valid quantile function. □

Gaussian case.

Sinkhorn for barycenters.

$$\min_{\mathrm{a}\in\Sigma_n} \sum_{s=1}^{S} \lambda_s \mathrm{L}_{\mathrm{C}_s}^{\varepsilon}(\mathrm{a},\mathrm{b}_s) \tag{10.3}$$

$$\min_{(\mathrm{P}_s)_s}\left\{\varepsilon\sum_s\lambda_s\mathrm{KL}(\mathrm{P}_s|\mathrm{K}_s)\ ;\ \forall\, s,\ {\mathrm{P}_s}^\top\mathbb{1}_n=\mathrm{b}_s,\quad \mathrm{P}_1\mathbb{1}_{n_1}=\ldots=\mathrm{P}_S\mathbb{1}_{n_S}\right\}, \tag{10.4}$$

where $\mathrm{K}_s\coloneqq e^{-\mathrm{C}_s/\varepsilon}$. The barycenter a is implicitly encoded in the common row marginal $\mathrm{a}=\mathrm{P}_1\mathbb{1}_{n_1}=\ldots=\mathrm{P}_S\mathbb{1}_{n_S}$.

$$\mathrm{P}_s=\mathrm{diag}(\mathrm{u}_s)\mathrm{K}_s\,\mathrm{diag}(\mathrm{v}_s), \tag{10.5}$$

$$\forall\, s\in[\![1,S]\!], \qquad \mathrm{v}_s^{(\ell+1)}\coloneqq\frac{\mathrm{b}_s}{{\mathrm{K}_s}^\top\mathrm{u}_s^{(\ell)}}, \tag{10.6}$$

$$\forall\, s\in[\![1,S]\!], \qquad \mathrm{u}_s^{(\ell+1)}\coloneqq\frac{\mathrm{a}^{(\ell+1)}}{\mathrm{K}_s\mathrm{v}_s^{(\ell+1)}}, \tag{10.7}$$

$$\text{where}\quad \mathrm{a}^{(\ell+1)}\coloneqq\prod_s(\mathrm{K}_s\mathrm{v}_s^{(\ell+1)})^{\lambda_s}. \tag{10.8}$$

Proposition 10.3 (Dual of entropic barycenters). *The optimal scalings in (10.5) can be written as $(\mathrm{u}_s,\mathrm{v}_s)=(e^{\mathrm{f}_s/\varepsilon},e^{\mathrm{g}_s/\varepsilon})$, where $(\mathrm{f}_s,\mathrm{g}_s)_s$ solve the dual problem*

$$\max_{(\mathrm{f}_s,\mathrm{g}_s)_s}\left\{\sum_s\lambda_s\left(\langle\mathrm{g}_s,\,\mathrm{b}_s\rangle-\varepsilon\langle\mathrm{K}_s e^{\mathrm{g}_s/\varepsilon},\,e^{\mathrm{f}_s/\varepsilon}\rangle\right)\ ;\ \sum_s\lambda_s\mathrm{f}_s=0\right\}. \tag{10.9}$$

Proof. Introduce Lagrange multipliers in (10.4):

$$\min_{(\mathrm{P}_s)_s,\mathrm{a}}\max_{(\mathrm{f}_s,\mathrm{g}_s)_s}\sum_s\lambda_s\Big(\varepsilon\mathrm{KL}(\mathrm{P}_s|\mathrm{K}_s)+\langle\mathrm{a}-\mathrm{P}_s\mathbb{1}_{n_s},\,\mathrm{f}_s\rangle+\langle\mathrm{b}_s-{\mathrm{P}_s}^\top\mathbb{1}_n,\,\mathrm{g}_s\rangle\Big).$$

Strong duality holds, so one can exchange the minimum and maximum. The minimization with respect to a gives the constraint $\sum_s\lambda_s\mathrm{f}_s=0$, and the minimization with respect to P_s gives the Legendre transform of $\mathrm{KL}(\cdot|\mathrm{K}_s)$:

$$\max_{(\mathrm{f}_s,\mathrm{g}_s)_s}\sum_s\lambda_s\left[\langle\mathrm{g}_s,\,\mathrm{b}_s\rangle-\varepsilon\mathrm{KL}^*\left(\frac{\mathrm{f}_s\oplus\mathrm{g}_s}{\varepsilon}\middle|\mathrm{K}_s\right)\right],\qquad\sum_s\lambda_s\mathrm{f}_s=0.$$

The separable conjugate is

$$\mathrm{KL}^*(\mathrm{U}|\mathrm{K})=\sum_{i,j}\mathrm{K}_{i,j}\big(e^{\mathrm{U}_{i,j}}-1\big), \tag{10.10}$$

because for $k>0$, $\sup_{r\geqslant0}ur-\big(r\log(r/k)-r+k\big)=k(e^u-1)$, and the case $k=0$ follows by lower semicontinuity. Dropping constants independent of $(\mathrm{f}_s,\mathrm{g}_s)_s$ gives (10.9). The coordinate maximization in g_s gives (10.6); the block maximization in all $(\mathrm{f}_s)_s$ gives the common marginal (10.8) and then (10.7). □

10.2 Multimarginal OT

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
$\inf_{\pi\in\mathcal{U}(\alpha_1,\ldots,\alpha_S)}\int_{\mathcal{X}_1\times\ldots\times\mathcal{X}_S}c(x_1,\ldots,x_S)\mathrm{d}\pi(x_1,\ldots,x_S)$, where $\mathcal{U}(\alpha_1,\ldots,\alpha_S)$ is the set of probability measures whose s-th marginal is α_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. $c_{\mathrm{bar}}(x_1,\dots,x_S)=\min_{x\in\mathbb{R}^d}\sum_{s=1}^S\lambda_s\|x-x_s\|^2$.

Proposition 10.4 (Multi-marginal formula for quadratic barycenters)**.** *Let $\beta_1,\dots,\beta_S\in\mathcal{P}_2(\mathbb{R}^d)$ and $\lambda\in\Sigma_S$. Define $B(x_1,\dots,x_S)=\sum_{s=1}^S\lambda_s x_s,\qquad c_{\mathrm{bar}}(x_1,\dots,x_S)=\min_x\sum_s\lambda_s\|x-x_s\|^2$. If $\pi^\star$ solves the multi-marginal OT problem with marginals $(\beta_s)_s$ and cost c_{bar}, then $\alpha^\star=B_\sharp\pi^\star$ is a Wasserstein barycenter. Conversely, every barycenter is obtained this way from an optimal multi-marginal plan.*

Proof. For any candidate barycenter α and couplings $\pi_s\in\mathcal{U}(\alpha,\beta_s)$, glue the couplings along their common α marginal to obtain a joint law of $(X,Y_1,\dots,Y_S)$. Conditioning on $(Y_s)_s$ and minimizing over X gives $\sum_s\lambda_s\mathbb{E}\|X-Y_s\|^2\geqslant\mathbb{E}\min_x\sum_s\lambda_s\|x-Y_s\|^2=\mathbb{E}c_{\mathrm{bar}}(Y_1,\dots,Y_S)$. Taking the infimum over the couplings gives that the barycenter value is at least the multi-marginal value. Conversely, from an optimal multi-marginal plan $\pi^\star$, set $X=B(Y_1,\dots,Y_S)$. The couplings between X and each Y_s are feasible for the barycenter problem and attain exactly the multi-marginal cost, proving equality and the formula. If $\alpha^\star$ is any barycenter, choose optimal couplings between $\alpha^\star$ and each β_s and glue them along the common $\alpha^\star$ marginal. Since the barycenter and multi-marginal values are equal, the conditional minimization inequality above must be an equality. Thus $X=B(Y_1,\dots,Y_S)$ almost surely for the induced optimal multi-marginal plan, and $\alpha^\star=B_\sharp\pi^\star$. □

Corollary 10.5 (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 α with mean $\mathbf{m}$ and covariance $\mathbf{\Sigma}$, Gelbrich's inequality, proved later in Theorem 14.9, 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. $\inf_{\pi\in\mathcal{U}(\alpha_1,\dots,\alpha_S)}\int c\mathrm{d}\pi+\varepsilon\,\mathrm{KL}(\pi|\alpha_1\otimes\cdots\otimes\alpha_S)$.

$$\mathrm{d}\pi^\star(x_1,\dots,x_S)=\exp\left(\frac{\sum_s f_s(x_s)-c(x_1,\dots,x_S)}{\varepsilon}\right)\prod_s\mathrm{d}\alpha_s(x_s),$$

10.3 Metric learning and inverse OT

Metric learning and derivatives of OT.

Proposition 10.6 (Derivative with respect to the cost)**.** *In the discrete setting, assume that the optimal coupling for $\mathrm{L_C(a,b)}$ is unique and denote it by $\mathrm{P}^\star(\mathrm{C})$. Then $\mathrm{C}\mapsto\mathrm{L_C(a,b)}$ is differentiable at C and $\nabla_{\mathrm{C}}\mathrm{L_C(a,b)}=\mathrm{P}^\star(\mathrm{C})$.*

Proof. The value is the minimum of affine functions of C, $\mathrm{L_C(a,b)}=\min_{\mathrm{P}\in\mathrm{U(a,b)}}\langle\mathrm{C},\mathrm{P}\rangle$. 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 $\{\mathrm{P}^\star(\mathrm{C})\}$, hence the value is differentiable with the displayed gradient. □

Thus, if the cost is parameterized as C_θ, gradients of losses involving OT values are obtained by backpropagating through the inner product $\langle\mathrm{P}^\star(\mathrm{C}_\theta),\partial_\theta\mathrm{C}_\theta\rangle$. 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.

Inverse Optimal Transport. $\inf_{\pi\in\mathcal{U}(\alpha,\beta)}\int c(x,y)\mathrm{d}\pi(x,y)$. 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 $G_\varepsilon(s)=\sup_{\pi\in\mathcal{U}(\alpha,\beta)}\int s\mathrm{d}\pi-\varepsilon\,\mathrm{KL}(\pi|\alpha\otimes\beta)$. $\mathcal{L}_\varepsilon(c;\widehat{\pi})=G_\varepsilon(-c)+G_\varepsilon^*(\widehat{\pi})+\int c\mathrm{d}\widehat{\pi}$ 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. 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.

$C_\theta=\sum_{r=1}^R\theta_r C^{(r)},\qquad\theta\in\Theta$, where Θ is convex and the matrices $C^{(r)}$ encode features, graph distances or a Mahalanobis parameterization.

Proposition 10.7 (Convex dual-gap formulation of inverse OT)**.** *Let $\widehat{P}\in\mathrm{U(a,b)}$ be an observed coupling and let C_θ depend affinely on $\theta\in\Theta$, where Θ is convex. The condition that $\widehat{P}$ is optimal for the cost C_θ is equivalent to the existence of dual potentials (f,g) such that $f_i+g_j\leqslant(C_\theta)_{i,j}$ and $\sum_{i,j}\widehat{P}_{i,j}\big((C_\theta)_{i,j}-f_i-g_j\big)=0$.*

$$\min_{\theta\in\Theta,f,g}\ R(\theta)+\lambda\sum_{i,j}\widehat{P}_{i,j}\big((C_\theta)_{i,j}-f_i-g_j\big)\quad\text{subject to}\quad f_i+g_j\leqslant(C_\theta)_{i,j}\quad\forall i,j.\tag{10.11}$$

Proof. For a fixed cost C_θ, Kantorovich duality gives $\min_{P\in\mathrm{U(a,b)}}\langle C_\theta,P\rangle=\max_{f_i+g_j\leqslant(C_\theta)_{i,j}}\langle f,\mathrm{a}\rangle+\langle g,\mathrm{b}\rangle$. Since $\widehat{P}$ has marginals (a,b), every dual feasible pair satisfies $\langle C_\theta,\widehat{P}\rangle-\langle f,\mathrm{a}\rangle-\langle g,\mathrm{b}\rangle=\sum_{i,j}\widehat{P}_{i,j}\big((C_\theta)_{i,j}-f_i-g_j\big)\geqslant0$. It vanishes if and only if $\widehat{P}$ reaches the dual value and is therefore optimal. If C_θ is affine and Θ and R are convex, the constraints and objective in (10.11) are convex, proving the relaxation claim. □

$\widehat{P}_{i,j}\approx a_i b_j\exp\left(\frac{f_i+g_j-(C_\theta)_{i,j}}{\varepsilon}\right)$,

10.4 Weak Optimal Transport

Barycentric projection of a coupling.

Definition 10.8 (Barycentric projection of a coupling)**.** Let $\alpha,\beta\in\mathcal{P}_2(\mathbb{R}^d)$ and let $\pi\in\mathcal{U}(\alpha,\beta)$. Disintegrate π with respect to its first marginal as $\pi(\mathrm{d}x,\mathrm{d}y)=\pi_x(\mathrm{d}y)\alpha(\mathrm{d}x)$. The barycentric projection of π is the map

$$\bar{T}_\pi(x):=\int_{\mathbb{R}^d}y\mathrm{d}\pi_x(y),\qquad\bar{\beta}_\pi:=(\bar{T}_\pi)_\sharp\alpha.\tag{10.12}$$

Proposition 10.9 (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 α and $\bar{\beta}_\pi$. Equivalently, $\bar{T}_\pi$ is a quadratic optimal transport map from α to the projected target $\bar{\beta}_\pi$.*

Proof. By Theorem 3.22, π is concentrated on a c-cyclically monotone set Γ for $c(x,y) = \|x-y\|^2$. For the quadratic cost, and since it is enough to check cyclic permutations, this means that every finite cycle $(x_i,y_i)_{i=1}^m \subset \Gamma$ satisfies $\sum_{i=1}^m \langle x_i,\, y_i\rangle \geqslant \sum_{i=1}^m \langle x_i,\, y_{i+1}\rangle, \qquad y_{m+1} = y_1$. After changing the disintegration on an α-negligible set, π_x is supported on the section $\Gamma_x = \{y\ ;\ (x,y)\in\Gamma\}$ for α-a.e. x. Choose $x_1,\dots,x_m$ in this full-measure set and independently sample $Y_i \sim \pi_{x_i}$. Applying the cyclic inequality to (x_i,Y_i) and taking expectations gives $\sum_{i=1}^m \langle x_i,\, \bar{T}_\pi(x_i)\rangle \geqslant \sum_{i=1}^m \langle x_i,\, \bar{T}_\pi(x_{i+1})\rangle$. Thus $(\mathrm{Id},\bar{T}_\pi)_\sharp\alpha$ is concentrated on a cyclically monotone graph. By the cyclic-monotonicity characterization of quadratic optimality, this plan is optimal between its two marginals, namely α and $\bar{\beta}_\pi$. □

$$\mathrm{WOT}_C(\alpha,\beta) \coloneqq \inf_{\pi\in\mathcal{U}(\alpha,\beta)} \int C(x,\pi_x)\mathrm{d}\alpha(x). \tag{10.13}$$

Proposition 10.10 (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 ν, with the standard qualification assumptions ensuring Fenchel–Rockafellar duality. For $g\in\mathcal{C}(\mathcal{Y})$ define the weak C-transform*

$$g^C(x) \coloneqq \inf_{\nu\in\mathcal{P}(\mathcal{Y})} \left\{ C(x,\nu) - \int g(y)\mathrm{d}\nu(y) \right\}.$$

Then

$$\mathrm{WOT}_C(\alpha,\beta) = \sup_{g\in\mathcal{C}(\mathcal{Y})} \left\{ \int g^C(x)\mathrm{d}\alpha(x) + \int g(y)\mathrm{d}\beta(y) \right\}.$$

When $C(x,\nu) = \int c(x,y)\mathrm{d}\nu(y)$, this reduces to the usual Kantorovich dual with $g^C(x) = \inf_y(c(x,y)-g(y))$.

Proof. For any coupling π and any $g\in\mathcal{C}(\mathcal{Y})$, the definition of g^C gives $C(x,\pi_x) \geqslant g^C(x) + \int g(y)\mathrm{d}\pi_x(y)$. After integration with respect to α, the second term becomes $\int g\mathrm{d}\beta$ because the second marginal of π is β. This proves weak duality.
For the 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 10.11 (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 π, 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 π be any coupling and disintegrate it as $\pi_x\alpha$. By Jensen's inequality, $\|x-\bar{T}_\pi(x)\|^2 \leqslant \int \|x-y\|^2\mathrm{d}\pi_x(y)$. Integrating in x gives $\int C_{\mathrm{bar}}(x,\pi_x)\mathrm{d}\alpha(x) \leqslant \int \|x-y\|^2\mathrm{d}\pi(x,y)$. Taking the infimum over π proves the claim. □

The barycentric cost is the canonical example to keep in mind: admissibility still constrains the full conditional laws to have second marginal β, but the objective only charges the displacement from x to $\bar{T}_\pi(x)$ and ignores the conditional variance around this barycenter.

11 Beyond Comparing Measures

11.1 Vector and Matrix-Valued Measures

Positive vector-valued measures.

Definition 11.1 (Positive vector-valued measure)**.** A positive $\mathbb{R}_+^m$-valued measure on $\mathcal{X}$ is a tuple $\mu = (\mu^1,\dots,\mu^m) \in \mathcal{M}_+(\mathcal{X};\mathbb{R}_+^m)$, where each component μ^k is a nonnegative finite measure.

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 $V_{t,\ell}^k$ is the momentum of channel k in spatial direction ℓ.

$$\mathcal{W}_\Phi^2(\mu_0,\mu_1) \coloneqq \inf_{u,V} \int_0^1 \int_{\mathcal{X}} \Phi(u_t(x),V_t(x))\,\mathrm{d}x\,\mathrm{d}t \tag{11.1}$$

$$\partial_t u_t + \nabla_x\cdot V_t = 0, \qquad (\nabla_x\cdot V_t)^k = \sum_{\ell=1}^d \partial_{x_\ell} V_{t,\ell}^k. \tag{11.2}$$

A simple quadratic family is obtained from a mobility matrix $\mathsf{M}(u)\in\mathbb{S}_+^m$, where $\mathbb{S}_+^m$ denotes the cone of real symmetric positive semidefinite matrices:
$\Phi_{\mathsf{M}}(u,V) = \sum_{\ell=1}^d V_\ell^\top \mathsf{M}(u)^\dagger V_\ell$, $\mathsf{M}_{\mathrm{diag}}(u) = \mathrm{diag}(u_1,\dots,u_m)$, $\mathsf{M}_\kappa(u) = \mathrm{diag}(u) + \kappa\Big(\sum_{k=1}^m u_k\Big) qq^\top$

Proposition 11.2 (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*

$$\mathcal{W}_{\mathrm{diag}}^2(\mu_0,\mu_1) = \sum_{k:m_k>0} m_k\, \mathcal{W}_2^2\!\left(\frac{\mu_0^k}{m_k},\frac{\mu_1^k}{m_k}\right),$$

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. □

Positive matrix-valued measures.

Definition 11.3 (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 $\mathrm{tr}\, A(x)$ is the scalar amount of mass at x, while, wherever $\mathrm{tr}\, A(x) > 0$, the normalized matrix $A(x)/\mathrm{tr}\, A(x)$ 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.

$$\mathcal{W}_{\mathrm{mat}}^2(\mathcal{A}_0, \mathcal{A}_1) := \inf_{A,P} \int_0^1 \int_{\mathcal{X}} \sum_{\ell=1}^{d} \mathrm{tr}\left(P_{t,\ell}^\top A_t^\dagger P_{t,\ell}\right) \mathrm{d}x\mathrm{d}t \tag{11.3}$$

$$\partial_t A_t + \nabla_x \cdot P_t = 0, \qquad \nabla_x \cdot P_t = \sum_{\ell=1}^{d} \partial_{x_\ell} P_{t,\ell}. \tag{11.4}$$

Proposition 11.4 (Diagonal matrix subproblem)**.** *Assume that the endpoints are diagonal in a fixed orthonormal basis, $\mathcal{A}_i = \mathrm{diag}(\mu_i^1, \dots, \mu_i^m)$, $i = 0, 1$, 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, $A_t = \mathrm{diag}(u_t^1, \dots, u_t^m)$, $P_{t,\ell} = \mathrm{diag}(V_{t,\ell}^1, \dots, V_{t,\ell}^m)$, 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 11.2.

Proof. The continuity equation (11.4) is diagonal entry by diagonal entry and gives $\partial_t u^k + \nabla \cdot V^k = 0$. Moreover, $\sum_{\ell=1}^{d} \mathrm{tr}\left(P_{t,\ell}^\top A_t^\dagger P_{t,\ell}\right) = \sum_{k=1}^{m} \frac{|V_t^k|^2}{u_t^k}$ with the same scalar perspective convention as before. The admissible set and the action are therefore exactly those of the diagonal vector model. □

11.2 Gromov–Wasserstein

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 $\mathrm{D} \in \mathbb{R}^{n\times n}$ and $\mathrm{D}' \in \mathbb{R}^{m\times m}$ 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.

$$\mathrm{GW}((\mathrm{a}, \mathrm{D}), (\mathrm{b}, \mathrm{D}'))^p := \min_{\mathrm{P}\in \mathrm{U}(\mathrm{a},\mathrm{b})} \mathcal{E}_{\mathrm{D},\mathrm{D}'}(\mathrm{P}) := \sum_{i,j,i',j'} \Delta(\mathrm{D}_{i,i'}, \mathrm{D}'_{j,j'})^p \mathrm{P}_{i,j} \mathrm{P}_{i',j'}, \tag{11.5}$$

where $p \geqslant 1$ and Δ is a distance on $\mathbb{R}$, typically $\Delta(u, v) = |u - v|$.
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.

General setting.

Definition 11.5 (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 α 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.

$$\mathcal{GW}(\mathbb{X}, \mathbb{Y})^p := \min_{\pi\in\mathcal{U}(\alpha,\beta)} \int_{\mathcal{X}^2\times\mathcal{Y}^2} \Delta(d_{\mathcal{X}}(x, x'), d_{\mathcal{Y}}(y, y'))^p \mathrm{d}\pi(x, y)\mathrm{d}\pi(x', y'). \tag{11.6}$$

Proposition 11.6 (Euclidean GW is controlled by Wasserstein)**.** *Let α, β 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 π be any coupling between α and β. 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 π proves the claim. □

Definition 11.7 (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 $\varphi : \mathrm{supp}(\alpha) \to \mathrm{supp}(\beta)$ such that $\varphi_\sharp \alpha = \beta$, $\varphi(\mathrm{supp}(\alpha)) = \mathrm{supp}(\beta)$, and $d_{\mathcal{Y}}(\varphi(x), \varphi(x')) = d_{\mathcal{X}}(x, x')$ for all $x, x' \in \mathrm{supp}(\alpha)$.

Theorem 11.8 (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 π is an optimal plan, then $d_{\mathcal{X}}(x,x')=d_{\mathcal{Y}}(y,y')$ holds $\pi\otimes\pi$-almost everywhere. By continuity, this equality holds on $\operatorname{supp}(\pi)^2$. $d_\pi((x,y),(x',y')) \coloneqq \frac{1}{2}d_{\mathcal{X}}(x,x')+\frac{1}{2}d_{\mathcal{Y}}(y,y')$. 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 ψ 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 π be an optimal coupling between $\mathbb{X}$ and $\mathbb{Y}$, and ξ an optimal coupling between $\mathbb{Y}$ and $\mathbb{Z}=(Z,d_Z,\gamma)$. By the gluing lemma, take σ on $\mathcal{X}\times\mathcal{Y}\times Z$ whose $(\mathcal{X},\mathcal{Y})$ and $(\mathcal{Y},Z)$ marginals are π and ξ. Let $\rho=(P_{\mathcal{X},Z})_\sharp\sigma$, and write $\bar\sigma=\sigma\otimes\sigma$ for the product law of two independent triples (x,y,z) and (x',y',z'). Then ρ is feasible between $\mathbb{X}$ and $\mathbb{Z}$, and

$$\begin{aligned}\mathcal{GW}(\mathbb{X},\mathbb{Z}) &\leqslant \left(\int |d_{\mathcal{X}}(x,x')-d_Z(z,z')|^p\,\mathrm{d}\bar\sigma\right)^{1/p}\\ &\leqslant \left(\int |d_{\mathcal{X}}(x,x')-d_{\mathcal{Y}}(y,y')|^p\,\mathrm{d}\bar\sigma\right)^{1/p}+\left(\int |d_{\mathcal{Y}}(y,y')-d_Z(z,z')|^p\,\mathrm{d}\bar\sigma\right)^{1/p}\\ &= \mathcal{GW}(\mathbb{X},\mathbb{Y})+\mathcal{GW}(\mathbb{Y},\mathbb{Z}),\end{aligned}$$

where the second inequality uses the pointwise triangle inequality followed by Minkowski's inequality. Symmetry and non-negativity are immediate. □

Proposition 11.9 (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 $Z=\mathcal{X}_0\times\mathcal{X}_1$, $d_t((x_0,x_1),(x_0',x_1'))\coloneqq(1-t)d_{\mathcal{X}_0}(x_0,x_0')+td_{\mathcal{X}_1}(x_1,x_1')$, $\mathbb{X}_t=(Z,d_t,\pi^\star)$. 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 $0\leqslant s\leqslant t\leqslant1$, $D\leqslant\mathcal{GW}(\mathbb{X}_0,\mathbb{X}_s)+\mathcal{GW}(\mathbb{X}_s,\mathbb{X}_t)+\mathcal{GW}(\mathbb{X}_t,\mathbb{X}_1)\leqslant sD+\mathcal{GW}(\mathbb{X}_s,\mathbb{X}_t)+(1-t)D$, so $\mathcal{GW}(\mathbb{X}_s,\mathbb{X}_t)\geqslant(t-s)D$. This proves equality. □

Proposition 11.10 (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 $p\geqslant1$. For each $x\in\mathcal{X}$ and $y\in\mathcal{Y}$, define the distance-profile measures on $\mathbb{R}_+$ by $\alpha_x\coloneqq(d_{\mathcal{X}}(x,\cdot))_\sharp\alpha$, $\beta_y\coloneqq(d_{\mathcal{Y}}(y,\cdot))_\sharp\beta$. 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 π to a coupling between α_x and β_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 π and then the p-th root proves the claim. □

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

$$\min_{\mathrm{P}\in\mathrm{U(a,b)}}\ \mathcal{E}_{\mathrm{D},\mathrm{D}'}(\mathrm{P})-\varepsilon\mathrm{H}(\mathrm{P}). \tag{11.7}$$

$$\mathrm{P}^{(\ell+1)}\coloneqq\min_{\mathrm{P}\in\mathrm{U(a,b)}}\ \langle\mathrm{P},\,\mathrm{C}^{(\ell)}\rangle-\varepsilon\mathrm{H}(\mathrm{P}),\qquad \mathrm{C}^{(\ell)}\coloneqq\mathrm{D}^{\odot2}\mathrm{a}\,\mathbb{1}_m^\top+\mathbb{1}_n(\mathrm{D}'^{\odot2}\mathrm{b})^\top-2\mathrm{D}\,\mathrm{P}^{(\ell)}\,\mathrm{D}'^{\top}. \tag{11.8}$$

Adding features.

$$\mathrm{FGW}_{\lambda,p}((\mathrm{a},\mathrm{D}),(\mathrm{b},\mathrm{D}'))^p\coloneqq\min_{\mathrm{P}\in\mathrm{U(a,b)}}(1-\lambda)\sum_{i,j}M_{ij}\mathrm{P}_{ij}+\lambda\sum_{i,j,i',j'}\Delta(\mathrm{D}_{ii'},\mathrm{D}'_{jj'})^p\mathrm{P}_{ij}\mathrm{P}_{i'j'}.$$

Hausdorff and Gromov–Hausdorff viewpoints. If A,B are compact subsets of a common metric space (Z,d_Z), their Hausdorff distance is

$$d_{\mathrm{H}}^Z(A,B)=\max\left\{\sup_{a\in A}\inf_{b\in B}d_Z(a,b),\sup_{b\in B}\inf_{a\in A}d_Z(a,b)\right\}.$$

$d_{\mathrm{GH}}(\mathcal{X},\mathcal{Y})=\inf_{Z,\varphi,\psi}d_{\mathrm{H}}^Z(\varphi(\mathcal{X}),\psi(\mathcal{Y}))$.

11.3 Quantum Optimal Transport

Finite-dimensional states and couplings.

Definition 11.11 (Hermitian and density matrices)**.** Let $\mathbb{H}_n$ be the real vector space of $n\times n$ Hermitian matrices, $\mathbb{H}_n^+=\{A\in\mathbb{H}_n\ ;\ A\succeq0\}$, $\mathbb{H}_n^{+,1}=\{A\in\mathbb{H}_n^+\ ;\ \operatorname{tr}(A)=1\}$. 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

$$\operatorname{tr}(F\ \mathrm{Tr}_B\,T)=\operatorname{tr}((F\otimes\mathrm{Id}_m)T),\qquad \operatorname{tr}(G\ \mathrm{Tr}_A\,T)=\operatorname{tr}((\mathrm{Id}_n\otimes G)T), \tag{11.9}$$

for all $F\in\mathbb{H}_n$ and $G\in\mathbb{H}_m$. Thus $\mathrm{Tr}_B(T)\in\mathbb{H}_n^+$ and $\mathrm{Tr}_A(T)\in\mathbb{H}_m^+$ play exactly the role of the two marginals of a classical coupling.

Definition 11.12 (Finite-dimensional quantum OT)**.** Let $A \in \mathbb{H}_n^{+,1}$, $B \in \mathbb{H}_m^{+,1}$ and let $C \in \mathbb{H}_{nm}$ be a Hermitian cost observable. The quantum OT value is the semidefinite program

$$\mathrm{QOT}_C(A,B) \coloneqq \min_{T \in \mathbb{H}_{nm}^+} \{\mathrm{tr}(CT)\ ;\ \mathrm{Tr}_B(T) = A,\ \mathrm{Tr}_A(T) = B\}\,. \tag{11.10}$$

Proposition 11.13 (Quantum Kantorovich duality)**.** *For $A \in \mathbb{H}_n^{+,1}$ and $B \in \mathbb{H}_m^{+,1}$, the dual of (11.10) is*

$$\mathrm{QOT}_C(A,B) = \max_{F \in \mathbb{H}_n,\ G \in \mathbb{H}_m} \{\mathrm{tr}(FA) + \mathrm{tr}(GB) :\ F \otimes \mathrm{Id}_m + \mathrm{Id}_n \otimes G \preceq C\}\,. \tag{11.11}$$

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

$$\mathrm{tr}(CT) + \mathrm{tr}(F(A - \mathrm{Tr}_B T)) + \mathrm{tr}(G(B - \mathrm{Tr}_A T)) = \mathrm{tr}(FA) + \mathrm{tr}(GB) + \mathrm{tr}((C - F \otimes \mathrm{Id}_m - \mathrm{Id}_n \otimes G)T)\,,$$

where (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. 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. □

Entropic regularization and Bregman iterations.

Definition 11.14 (von Neumann quantum entropy)**.** For a density matrix or positive semidefinite matrix T, the shifted von Neumann entropy functional used here is

$$H(T) = \mathrm{tr}(T(\log T - \mathrm{Id}))\,, \qquad \nabla H(T) = \log T,$$

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

$$\mathrm{QOT}_C^\varepsilon(A,B) = \min_{T \succeq 0} \{\mathrm{tr}(CT) + \varepsilon H(T) :\ \mathrm{Tr}_B(T) = A, \mathrm{Tr}_A(T) = B\}\,. \tag{11.12}$$

Proposition 11.15 (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. □

$$D_H(T|K) = \mathrm{tr}(T(\log T - \log K) - T + K)\,.$$

$\mathcal{M}_A = \{T \succeq 0 : \mathrm{Tr}_B(T) = A\}, \qquad \mathcal{M}_B = \{T \succeq 0 : \mathrm{Tr}_A(T) = B\}.$

Proposition 11.16 (Exact Bregman projections)**.** *Assume $A, B \succ 0$ 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 $\mathrm{Tr}_B\, T_e(F^+,G) = A$. 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 $\mathrm{tr}(CT) + \varepsilon H(T) = \varepsilon D_H(T|K) - \varepsilon\,\mathrm{tr}(K)$ holds, so the primal minimizer is the constrained Bregman projection of K up to an additive constant. For the projection of a positive definite matrix S onto $\mathcal{M}_A$, the affine set contains the positive definite point $A \otimes \mathrm{Id}_m/m$; the entropy derivative $\log T - \log S$ is singular at the boundary, so the projection lies in the interior of the positive cone. Its Lagrangian first variation is $\log T - \log S - \Lambda \otimes \mathrm{Id}_m = 0$ for a Hermitian multiplier Λ. Hence $T = \exp(\log S + \Lambda \otimes \mathrm{Id}_m)$. If $S = T_e(F,G)$, this is again of the form $T_e(F + \varepsilon\Lambda, G)$. The multiplier is fixed by the marginal equation $\mathrm{Tr}_B(T) = A$. 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. □

$\mathrm{Tr}_B\, T_e(F,G) = A, \qquad \mathrm{Tr}_A\, T_e(F,G) = B$

Gurvits scaling and quantum Sinkhorn.

$$T_s(F,G) = \exp\left(\frac{Z}{2\varepsilon}\right)\exp(-C/\varepsilon)\exp\left(\frac{Z}{2\varepsilon}\right) = (U\otimes V)K(U\otimes V), \qquad Z = F\otimes \mathrm{Id}_m + \mathrm{Id}_n\otimes G, \tag{11.15}$$

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
$U\,\mathcal{K}(V^2)\,U = A, \qquad V\,\mathcal{K}^\star(U^2)\,V = B.$

$$\begin{aligned} R_V &= \mathcal{K}(V^2), \quad & U &\leftarrow R_V^{-1/2}\big(R_V^{1/2}AR_V^{1/2}\big)^{1/2}R_V^{-1/2},\\ S_U &= \mathcal{K}^\star(U^2), \quad & V &\leftarrow S_U^{-1/2}\big(S_U^{1/2}BS_U^{1/2}\big)^{1/2}S_U^{-1/2}. \end{aligned} \tag{11.16}$$

These inverse square roots are well-defined when $K \succ 0$ and $U,V,A,B \succ 0$. This is Gurvits/operator 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.

12 Dynamic Optimal Transport

12.1 Evolutions over the Space of Measures

Lagrangian and Eulerian descriptions.

$$\frac{\mathrm{d}x(t)}{\mathrm{d}t} = v_t(x(t)), \tag{12.1}$$

$$\frac{\partial \alpha_t}{\partial t} + \operatorname{div}(v_t\alpha_t) = 0. \tag{12.2}$$

$$\int_0^1\int_{\mathbb{R}^d} \left(\partial_t\varphi(t,x) + \langle v_t(x),\, \nabla_x\varphi(t,x)\rangle\right)\mathrm{d}\alpha_t(x)\mathrm{d}t = 0. \tag{12.3}$$

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 (α_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 φ 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

$$\partial_t\alpha_t + \operatorname{div}(\alpha_t v_t) = 0. \tag{12.4}$$

$\mathcal{H}_\alpha = \{v\ ;\ \operatorname{div}(\alpha v) = 0\}$.

Dacorogna–Moser inversion.

$$v_t = -\frac{1}{\alpha_t}\nabla\Delta^{-1}(\partial_t\alpha_t), \tag{12.5}$$

Least-square inversion and gradient structure.

$$\min_v \frac{1}{2}\int_0^1\int_{\mathbb{R}^d}\|v_t(x)\|^2\,\mathrm{d}\alpha_t(x)\mathrm{d}t \quad \text{subject to} \quad \partial_t\alpha_t + \operatorname{div}(\alpha_t v_t) = 0. \tag{12.6}$$

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 φ_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 φ_t for the continuity equation. The constrained problem has the formal saddle formulation

$$\min_{v}\max_{\varphi}\int_0^1\left[\frac{1}{2}\int_{\mathbb{R}^d}\|v_t(x)\|^2\,\mathrm{d}\alpha_t(x)+\int_{\mathbb{R}^d}\varphi_t(x)\big(\operatorname{div}(\alpha_t v_t)(x)+\partial_t\alpha_t(x)\big)\mathrm{d}x\right]\mathrm{d}t.$$

Integrating by parts in the divergence term gives, for each t,

$$\int\left(\frac{1}{2}\|v_t\|^2-\langle\nabla\varphi_t,\,v_t\rangle\right)\mathrm{d}\alpha_t+\int\varphi_t\,\partial_t\alpha_t.$$

The pointwise minimizer in v_t is therefore $v_t=\nabla\varphi_t$. Substituting this into the constraint $\partial_t\rho_t+\operatorname{div}(\rho_t v_t)=0$ gives the weighted Poisson equation in (12.7). The inverse notation is just a shorthand for solving this equation on zero-mean right-hand sides, modulo additive constants. □

12.2 Benamou–Brenier dynamic formulation of OT

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 (α_t,v_t) solving $\partial_t\alpha_t+\nabla\cdot(\alpha_t v_t)=0$ with $\alpha_{t=0}=\alpha_0$ and $\alpha_{t=1}=\alpha_1$. If α_0 has a density and T is the optimal Monge map $T_\sharp\alpha_0=\alpha_1$, 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 (α_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 π: sample $(X,Y)\sim\pi$ and move along the straight path $\gamma_{X,Y}(t)=(1-t)X+tY$. This path measure has action $\int\|x-y\|^2\mathrm{d}\pi(x,y)$; projecting path velocities onto their conditional mean at time t gives an admissible Eulerian velocity with no larger action, so the dynamic value is no larger than the Kantorovich value.

Conversely, for a smooth deterministic path, take the flow T_t defined by $\dot T_t=v_t\circ T_t$ and $T_0=\mathrm{Id}$. Then $\alpha_t=(T_t)_\sharp\alpha_0$ and $(T_1)_\sharp\alpha_0=\alpha_1$. Jensen's inequality gives $\|T_1(x)-x\|^2\leqslant\int_0^1\|v_t(T_t(x))\|^2\mathrm{d}t$. After integration with respect to α_0, the Monge cost is bounded above by the dynamic action. For general finite-energy solutions of the continuity equation, the superposition principle lifts the curve to a probability measure on absolutely continuous paths; applying Jensen's inequality pathwise gives a coupling of the endpoints whose quadratic cost is no larger than the action. Thus the Kantorovich value is bounded above by the dynamic value. □

Although (12.8) is not jointly convex in (α_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

$$\mathcal{W}_2^2(\alpha_0,\alpha_1)=\inf_{\substack{\partial_t\rho_t+\operatorname{div}m_t=0\\ \rho_{t=0}\mathrm{d}x=\alpha_0,\ \rho_{t=1}\mathrm{d}x=\alpha_1}}\int_0^1\int_{\mathbb{R}^d}\frac{\|m_t(x)\|^2}{\rho_t(x)}\,\mathrm{d}x\,\mathrm{d}t,\tag{12.10}$$

Extensions of the dynamic formulation.

Dynamic unbalanced OT.

$$\partial_t\rho_t+\nabla\cdot m_t=s_t,\qquad\int_0^1\int\left(\frac{|m_t|^2}{\rho_t}+\kappa^2\frac{s_t^2}{\rho_t}\right)\mathrm{d}x\,\mathrm{d}t,$$

$$\mathcal{A}_\kappa(\rho,m,s)\coloneqq\int\left(\frac{\|\dot m\|^2}{\dot\rho}+\kappa^2\frac{\dot s^2}{\dot\rho}\right)\mathrm{d}\lambda,\qquad(\dot\rho,\dot m,\dot s)=\left(\frac{\mathrm{d}\rho}{\mathrm{d}\lambda},\frac{\mathrm{d}m}{\mathrm{d}\lambda},\frac{\mathrm{d}s}{\mathrm{d}\lambda}\right),$$

where λ dominates ρ 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 12.4 (Static/dynamic equivalence for unbalanced OT)**.** *Fix the action above and let CW_κ be the cone value of Theorem 9.4 with the cone metric normalized to the same growth scale κ. For nonnegative finite measures α_0,α_1 on $\mathbb{R}^d$, the dynamic value*

$$\mathrm{WFR}_\kappa^2(\alpha_0,\alpha_1)\coloneqq\inf_{\substack{\partial_t\rho_t+\nabla\cdot m_t=s_t\\ \rho_0=\alpha_0,\ \rho_1=\alpha_1}}\int_0^1\mathcal{A}_\kappa(\rho_t,m_t,s_t)\,\mathrm{d}t\tag{12.11}$$

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 CW_κ of Theorem 9.4. This is the standard static/dynamic identification for the Hellinger–Kantorovich and Wasserstein–Fisher–Rao metrics.

Projecting a cone curve back to the base space with the weight r^2 produces a measure curve ρ_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 (ρ_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. □

13 Wasserstein Gradient Flows

13.1 Minimizing Movements and Wasserstein Gradients

$$\alpha_{t+\tau} := \arg\min_{\alpha} \frac{1}{2\tau}\mathcal{W}_2(\alpha_t,\alpha)^2 + f(\alpha). \tag{13.1}$$

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:
$x(t+\tau) = (\mathrm{Id} + \tau\nabla h)^{-1}(x(t))$. $x(t+\tau) = (\mathrm{Id} - \tau\nabla h)(x(t)) = x(t) - \tau\nabla h(x(t))$.

$$\dot{x}(t) = -\nabla h(x(t)). \tag{13.2}$$

Wasserstein gradient formula. $f((1-\tau)\alpha + \tau\beta) = f(\alpha + \tau\rho) = f(\alpha) + \tau\int[\delta f(\alpha)](x)\mathrm{d}\rho(x) + o(\tau)$.

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 α is the gradient vector field $\nabla_{\mathcal{W}} f(\alpha) = \nabla_x \delta f(\alpha)$.

$$\frac{\partial \alpha_t}{\partial t} + \mathrm{div}(-\nabla_{\mathcal{W}} f(\alpha_t)\alpha_t) = 0. \tag{13.3}$$

Proposition 13.2 (Formal Wasserstein gradient)**.** *Assume that f admits a smooth first variation $\delta f(\alpha)$ and that α 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, $(\mathrm{Id}+\tau v)_\sharp\alpha = \alpha - \tau\,\mathrm{div}(\alpha v) + o(\tau)$. Using the definition of the first variation, $f((\mathrm{Id}+\tau v)_\sharp\alpha) = f(\alpha) - \tau\int\delta f(\alpha)\,\mathrm{div}(\alpha v)\mathrm{d}x + o(\tau)$. An integration by parts, with either compact support or vanishing boundary flux, gives $-\int\delta f(\alpha)\,\mathrm{div}(\alpha v)\mathrm{d}x = \int\langle\nabla\delta f(\alpha), v\rangle\mathrm{d}\alpha$. By definition of the Riesz representative for the $L^2(\alpha)$ metric, this representative is $\nabla\delta f(\alpha)$. □

From the JKO step to the velocity field. $\min_v \frac{1}{2\tau}\tau^2\|v\|^2_{L^2(\alpha_t)} + f((\mathrm{Id}+\tau v)_\sharp\alpha_t)$. $(\mathrm{Id}+\tau v)_\sharp\alpha_t = \alpha_t - \tau\,\mathrm{div}(v\alpha_t) + o(\tau)$
$f((\mathrm{Id}+\tau v)_\sharp\alpha_t) = f(\alpha_t) - \tau\int\delta f(\alpha_t)\,\mathrm{div}(v\alpha_t)\mathrm{d}x + o(\tau) = f(\alpha_t) + \tau\int\langle\nabla_x\delta f(\alpha_t)(x), v(x)\rangle\mathrm{d}\alpha_t(x) + o(\tau)$ $\min_v f(\alpha_t) + \tau\int\Big[\frac{1}{2}\|v(x)\|^2 + \langle\nabla_{\mathcal{W}} f(\alpha_t)(x), v(x)\rangle\Big]\mathrm{d}\alpha_t(x) + o(\tau)$.

Discrete evolutions. If $f(\alpha)$ can be evaluated on discrete distributions and $\nabla_{\mathcal{W}}$ 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 $\frac{1}{n}\sum_i\|\dot{x}_i\|^2$.

Linear Functionals.

$$f(\alpha) = \int h(x)\mathrm{d}\alpha(x). \tag{13.5}$$

$\frac{\partial\alpha_t}{\partial t} + \mathrm{div}(-\nabla h\alpha_t) = 0$.

Shannon Neg-Entropy.

$$f(\alpha) = \int \log\left(\frac{\mathrm{d}\alpha}{\mathrm{d}x}(x)\right)\mathrm{d}\alpha(x). \tag{13.6}$$

$\partial_t\alpha_t = \Delta\alpha_t$.

$$f(\alpha) = \int g\left(\frac{\mathrm{d}\alpha}{\mathrm{d}x}\right)\mathrm{d}x, \tag{13.7}$$

$\frac{\partial\alpha_t}{\partial t} = \Delta(P(\alpha_t))$, where the pressure P satisfies $P'(s) = sg''(s)$.

Interaction Energies.

$$f(\alpha) := \iint k(x,y)\mathrm{d}\alpha(x)\mathrm{d}\alpha(y). \tag{13.8}$$

For a symmetric kernel k:
$\delta f(\alpha)(x) = 2\int k(x,y)\mathrm{d}\alpha(y)$, $\nabla_{\mathcal{W}} f(\alpha)(x) = 2\int\nabla_x k(x,y)\mathrm{d}\alpha(y)$. 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 β, then

$$\|\alpha-\beta\|_k^2 = \iint k\mathrm{d}\alpha\mathrm{d}\alpha - 2\int\left(\int k(x,y)\mathrm{d}\beta(y)\right)\mathrm{d}\alpha(x) + \text{constant}.$$

Thus MMD-type training energies are exactly an interaction energy plus a linear potential; the teacher distribution appears through the potential $x \mapsto -2\int k(x,y)\mathrm{d}\beta(y)$. The corresponding empirical Wasserstein gradient flow is
$\dot{x}_i(t) = -\frac{2}{n}\sum_j\nabla_x k(x_i(t), x_j(t)) + 2\int\nabla_x k(x_i(t), y)\mathrm{d}\beta(y)$.

Stochastic particles and McKean–Vlasov limits. $\mathrm{d}X_t = b(X_t)\mathrm{d}t + \sqrt{2}\sigma\mathrm{d}B_t$, and the one-particle law $\alpha_t = \rho_t \mathrm{d}x$ directly satisfies the linear Fokker–Planck equation
$\partial_t \rho_t = -\operatorname{div}(b\rho_t) + \sigma^2 \Delta \rho_t$. $\mathrm{d}X_i^n(t) = b(X_i^n(t), \mu_t^n)\mathrm{d}t + \sqrt{2}\sigma \mathrm{d}B_i(t), \qquad \mu_t^n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i^n(t)}$. For finite n, the empirical law μ_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 μ_t^n converges in probability to this law.
$\partial_t \rho_t = -\operatorname{div}\big(b(x,\rho_t)\rho_t\big) + \sigma^2 \Delta \rho_t$. $b(x,\rho) = -\nabla \frac{\delta \mathcal{E}}{\delta \rho}(x)$, $\mathcal{E}(\rho) + \sigma^2 \int \rho \log \rho \, \mathrm{d}x$.

13.2 Geodesic Convexity and Convergence

Geodesics and convexity. $\alpha_t = ((1-t)P_0 + tP_1)_\sharp \pi^\star, \qquad t \in [0,1]$, 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 $((1-t)\mathrm{Id} + tT)_\sharp \alpha_0$.

Definition 13.3 (Geodesic convexity)**.** A functional f on $\mathcal{P}_2(\mathbb{R}^d)$ is geodesically convex if for every $\mathcal{W}_2$ geodesic $(\alpha_t)_t$, $f(\alpha_t) \leqslant (1-t)f(\alpha_0) + tf(\alpha_1)$. It is λ-geodesically convex if the right-hand side is improved by $-\frac{\lambda}{2}t(1-t)\,\mathcal{W}_2^2(\alpha_0,\alpha_1)$.

Proposition 13.4 (Basic geodesically convex energies)**.** *1. If h is convex, then $\alpha \mapsto \int h \mathrm{d}\alpha$ is geodesically convex; if h is λ-strongly convex, it is λ-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 λ-geodesically convex when V is λ-strongly convex.

Proof. Along a Monge geodesic $X_t = (1-t)X_0 + tX_1$, convexity of h gives $h(X_t) \leqslant (1-t)h(X_0) + th(X_1)$, 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 λ-geodesically convex linear potential. □

Convergence of the flow.

Proposition 13.5 (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 α_t to $\alpha^\star$, then $f(\alpha_t) - f(\alpha^\star) \leqslant -\frac{\mathrm{d}}{\mathrm{d}t}\frac{1}{2}\mathcal{W}_2^2(\alpha_t,\alpha^\star)$, and consequently $f(\alpha_t) - f(\alpha^\star) \leqslant \frac{\mathcal{W}_2^2(\alpha_0,\alpha^\star)}{2t}$. If f is λ-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 $f(\alpha_t) - f(\alpha^\star) \leqslant \int \langle v_t(x), T_t(x) - x\rangle \, \mathrm{d}\alpha_t(x)$. $\frac{\mathrm{d}}{\mathrm{d}t}\frac{1}{2}\mathcal{W}_2^2(\alpha_t,\alpha^\star) = \int \langle x - T_t(x), v_t(x)\rangle \, \mathrm{d}\alpha_t(x)$, 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 λ-geodesically convex, the Wasserstein analogue of strong convexity gives the slope inequality $\int \|\nabla_{\mathcal{W}} f(\alpha_t)\|^2 \mathrm{d}\alpha_t \geqslant 2\lambda\big(f(\alpha_t) - f(\alpha^\star)\big)$. $\frac{\mathrm{d}}{\mathrm{d}t}\big(f(\alpha_t) - f(\alpha^\star)\big) \leqslant -2\lambda\big(f(\alpha_t) - f(\alpha^\star)\big)$, and Gronwall's lemma gives the exponential rate. □

Proposition 13.6 (Convex examples covered by the theory)**.** *The hypotheses of Proposition 13.5 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 λ-strongly convex, then f is λ-geodesically convex and the flow enjoys the exponential rate of Proposition 13.5.
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 λ-strongly convex, then the relative entropy $\mathrm{KL}(\alpha|\gamma)$ is λ-geodesically convex. Its flow is the Fokker–Planck equation with invariant law γ.

Proof. Let $(\alpha_t)_t$ be the McCann interpolation between α_0 and α_1, written with an optimal coupling as $X_t = (1-t)X_0 + tX_1$. For a linear energy, Jensen's inequality gives $h(X_t) \leqslant (1-t)h(X_0) + th(X_1)$, 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 λ-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 λ-geodesically convex linear potential generated by V. Proposition 13.5 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.

$$\mathrm{Ent}_{\mathfrak{m}}(\alpha) := \begin{cases} \displaystyle\int_{\mathcal{X}} \rho \log \rho \, \mathrm{d}\mathfrak{m}, & \text{if } \alpha = \rho\,\mathfrak{m}, \\ +\infty, & \text{otherwise.} \end{cases}$$

Theorem 13.7 (Ricci curvature and entropy convexity)**.** *Let (M, g) be a smooth compact connected Riemannian manifold without boundary, and let $\mathfrak{m} = \mathrm{vol}_g$. For $\lambda \in \mathbb{R}$, the lower Ricci bound $\mathrm{Ric}_g \geqslant \lambda g$ holds if and only if $\mathrm{Ent}_{\mathfrak{m}}$ is λ-geodesically convex on $(\mathcal{P}_2(M), \mathcal{W}_2)$.*

13.3 Training Two-Layer MLPs as Wasserstein Flows

$x = (u, v) \in \mathbb{R}^d \times \mathbb{R}^{d'}$, where u is the inner weight and v is the outer vector weight. For a scalar nonlinearity σ, define the vector-valued feature
$\psi(x, z) = v\,\sigma(\langle u,\, z\rangle) \in \mathbb{R}^{d'}$. $G_X(z) = \frac{1}{n}\sum_{i=1}^n \psi(x_i, z)$, $\quad G_\alpha(z) = \int \psi(x, z)\mathrm{d}\alpha(x)$, $\quad \alpha = \frac{1}{n}\sum_i \delta_{x_i}$. Let ρ be a probability distribution on data-label pairs $(z, y) \in \mathbb{R}^d \times \mathbb{R}^{d'}$. The population risk is
$f(\alpha) = \int \ell(G_\alpha(z), y)\mathrm{d}\rho(z, y)$,

$$\dot{x}_i = -n\nabla_{x_i} F(X), \qquad F(X) = f\left(\frac{1}{n}\sum_i \delta_{x_i}\right),$$

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 ℓ is differentiable in its first variable. The first variation is

$$\delta f(\alpha)(x) = \int \langle \nabla_1 \ell(G_\alpha(z), y),\, \psi(x, z)\rangle \mathrm{d}\rho(z, y), \tag{13.9}$$

$\nabla_{\mathcal{W}} f(\alpha)(x) = \nabla_x \delta f(\alpha)(x) = \int [D_x \psi(x, z)]^\top \nabla_1 \ell(G_\alpha(z), y)\mathrm{d}\rho(z, y)$. 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:

$$f(\alpha) = \frac{1}{2}\iint k(x, x')\mathrm{d}\alpha(x)\mathrm{d}\alpha(x') + \int g(x)\mathrm{d}\alpha(x) + \frac{1}{2}\int \|y\|^2 \mathrm{d}\rho(z, y), \tag{13.10}$$

$$k(x, x') = \int \langle \psi(x, z),\, \psi(x', z)\rangle \mathrm{d}\rho(z, y), \qquad g(x) = -\int \langle y,\, \psi(x, z)\rangle \mathrm{d}\rho(z, y). \tag{13.11}$$

$\delta f(\alpha)(x) = \int k(x, x')\mathrm{d}\alpha(x') + g(x)$, $\quad \nabla_{\mathcal{W}} f(\alpha)(x) = \int \nabla_x k(x, x')\mathrm{d}\alpha(x') + \nabla_x g(x)$.
Classical convexity and stationarity. $F(\alpha) = \frac{1}{2}\iint k(x, x')\mathrm{d}\alpha(x)\mathrm{d}\alpha(x') + \int V(x)\mathrm{d}\alpha(x) + C$, $Q((1-s)\alpha + s\beta) \leqslant (1-s)Q(\alpha) + sQ(\beta)$, $\quad Q(\alpha) = \frac{1}{2}\iint k\mathrm{d}\alpha\mathrm{d}\alpha$.

Proposition 13.8 (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 ρ_∞ 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 β, the variational inequality $\int \delta F(\alpha_\infty)(x)\mathrm{d}(\beta - \alpha_\infty)(x) \geqslant 0$. Then α_∞ 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 β. Classical convexity of F in the affine variable α 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. □

Proposition 13.9 (Formal global optimality for two-homogeneous mean-field flows)**.** *Assume that the feature is positively two-homogeneous in the neuron variable, $\psi(\lambda x, z) = \lambda^2 \psi(x, z)$ $\quad (\lambda > 0)$, and that $f(\alpha) = J(G_\alpha)$ with J convex and differentiable as a functional of the predictor. Let α be a smooth stationary point of the Wasserstein flow, so that $\nabla_x \delta f(\alpha)(x) = 0$ on $\mathrm{supp}(\alpha)$. Assume also full directional support: for every nonzero direction ω, the support of α intersects the ray $\{\lambda\omega : \lambda > 0\}$. Then α is a global minimizer of f over the mean-field model class.*

Proof. Write

$$h_\alpha(x) = \delta f(\alpha)(x) = \langle \nabla J(G_\alpha), \psi(x, \cdot)\rangle_\rho .$$

By two-homogeneity of ψ, one has $h_\alpha(\lambda x) = \lambda^2 h_\alpha(x)$. Normalize a nonzero direction ω and choose $r_\omega > 0$ with $r_\omega \omega \in \mathrm{supp}(\alpha)$. Stationarity gives a zero radial derivative at this point: $0 = \frac{\mathrm{d}}{\mathrm{d}r} h_\alpha(r\omega)\Big|_{r=r_\omega} = 2r_\omega h_\alpha(\omega)$. Hence $h_\alpha(\omega) = 0$ for every direction ω, and by homogeneity $h_\alpha(x) = 0$ for every x.
For any competitor β, 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. □

14 Generative Models via Transportation

14.1 Generative Models via Flow Matching

Stochastic interpolant. We assume that α_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.

$$\forall t \in [0, 1], \quad \alpha_t := (P_t)_\sharp \pi. \tag{14.1}$$

A common case is $d' = 2d$. We write $(x, y) \in \mathbb{R}^{d'} = \mathbb{R}^d \times \mathbb{R}^d$ and assume $P_0(x, y) = x$ and $P_1(x, y) = y$, so that π is a probabilistic coupling between α_0 and α_1, i.e. π has marginals (α_0, α_1).
If $\pi = \alpha \otimes \beta$ and $\alpha = \frac{1}{n}\sum_i \delta_{x_i}$, $\beta = \frac{1}{m}\sum_j \delta_{y_j}$, then α_t consists of $n \times m$ Dirac masses
$\alpha_t = \frac{1}{nm}\sum_{i,j} \delta_{P_t(x_i, y_j)}$. 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.

Flow matching formula.

Proposition 14.1 (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 α_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 (α_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 φ,

$$\int_{\mathbb{R}^d} \varphi(z)\,\mathrm{d}\alpha_t(z) = \int_{\mathbb{R}^{d'}} \varphi(P_t(u))\,\mathrm{d}\pi(u). \tag{14.4}$$

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,

$$\int \langle m(z),\, v_t(z)\rangle\,\mathrm{d}\alpha_t(z) = \int \langle m(P_t(u)),\, [\partial_t P_t](u)\rangle\,\mathrm{d}\pi(u). \tag{14.5}$$

The weak form of $\partial_t \alpha_t + \operatorname{div}(\alpha_t v_t) = 0$ is that, for every smooth scalar test function φ,

$$\frac{\mathrm{d}}{\mathrm{d}t}\int \varphi(z)\,\mathrm{d}\alpha_t(z) - \int \langle v_t(z),\, \nabla\varphi(z)\rangle\,\mathrm{d}\alpha_t(z) = 0. \tag{14.6}$$

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
$\int_{\mathbb{R}^d} \langle \psi(z),\, \mathrm{d}m_t(z)\rangle \coloneqq \int_{\mathbb{R}^{d'}} \langle \psi(P_t(u)),\, [\partial_t P_t](u)\rangle \mathrm{d}\pi(u)$ $\mathrm{d}m_t(z) = v_t(z)\mathrm{d}\alpha_t(z)$, $v_t = \frac{\mathrm{d}m_t}{\mathrm{d}\alpha_t}$. In the language of Lebesgue decomposition, the flux measure m_t has only an absolutely continuous part with respect to α_t and no singular part; the conditional expectation is precisely this density. Equivalently, disintegrating π 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.
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 $(T_t)_\sharp \alpha_0 = \alpha_t$.

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 α_t is a convolution of rescaled versions of α_0 and α_1. The flow-matching problem (14.2) becomes
$\min_{(v_t)_t} \int_{\mathbb{R}^d\times\mathbb{R}^d} \|v_t((1-t)x + ty) - (y - x)\|^2 \,\mathrm{d}\alpha_0(x)\mathrm{d}\alpha_1(y)$.

Proposition 14.2 (Tweedie identity)**.** *Let W be a random vector in $\mathbb{R}^d$ with density β. For $\sigma > 0$, observe $Z = W + \sigma\,\varepsilon$, where $\varepsilon \sim \mathcal{N}(0, I_d)$ is independent of W. Denote by $\beta_\sigma = \beta * \mathcal{N}\big(0, \sigma^2 I_d\big)$ the density of Z. Then $\mathbb{E}\big[W \mid Z = z\big] = z + \sigma^2\,\nabla \log \beta_\sigma(z)$ 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 φ_σ 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 14.3 (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 $Z_t = (1 - t)\,X + t\,Y$, $\alpha_t = \mathrm{Law}(Z_t)$. 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 $v^\star(x, t) = -\frac{1}{1-t}\,x - \frac{t}{1-t}\,\nabla \log \alpha_t(x)$ $(x \in \mathbb{R}^d,\ t \in (0, 1))$. 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 14.2. 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 14.2 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. □

When is the induced map optimal? Integrating the learned velocity gives a deterministic map from α_0 to α_1, but this map is not automatically the Brenier optimal map. It is optimal only in special cases where the accumulated flow remains the gradient of a convex potential.

Proposition 14.4 (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^{FM} from α_0 to α_t is

$$T_t^{\mathrm{FM}} = \Sigma_0^{1/2}\Big((1-t)^2\mathrm{Id} + t^2\Sigma_0^{-1/2}\Sigma_1\Sigma_0^{-1/2}\Big)^{1/2}\Sigma_0^{-1/2}. \tag{14.11}$$

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 $\widetilde{\Sigma}_t = (1-t)^2\mathrm{Id} + t^2C$. Because Id and C commute, the affine flow map in whitened coordinates is simply $\widetilde{T}_t = \widetilde{\Sigma}_t^{1/2}$. Indeed, $\frac{\mathrm{d}}{\mathrm{d}t}\widetilde{\Sigma}_t^{1/2} = \big(tC - (1-t)\mathrm{Id}\big)\widetilde{\Sigma}_t^{-1/2}$, which is exactly the equation $\dot{\widetilde{T}}_t = \widetilde{A}_t\widetilde{T}_t$ with $\widetilde{T}_0 = \mathrm{Id}$. Returning to the original coordinates gives (14.11), and $t = 1$ gives (14.12).

Both T_1^{FM} and T^{OT} push $\mathcal{N}(0, \Sigma_0)$ to $\mathcal{N}(0, \Sigma_1)$. The Brenier map between nondegenerate Gaussians is the unique symmetric positive definite linear map with this property. Hence $T_1^{\mathrm{FM}} = T^{\mathrm{OT}}$ if and only if T_1^{FM} is symmetric positive definite. The map T_1^{FM} is similar to $C^{1/2}$, so if it is symmetric then it is automatically positive definite. Since $C^{1/2}$ is symmetric positive definite, $(T_1^{\mathrm{FM}})^\top = \Sigma_0^{-1/2}C^{1/2}\Sigma_0^{1/2}$. Thus symmetry of T_1^{FM} is equivalent to $\Sigma_0 C^{1/2} = C^{1/2}\Sigma_0$, 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 Σ_0 and Σ_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. □

Proposition 14.4 explains why the statement "flow matching gives an optimal map" is fragile. The same terminal map (14.12) is obtained for any scalar schedule $Z_t = a_tX_0 + b_tX_1$ with the same endpoints, because after whitening the covariance path remains $a_t^2\mathrm{Id} + b_t^2C$.

Variations on the interpolant. $v_t(z) = \lambda'(t)\, v_{\lambda(t)}^{\mathrm{lin}}(z), \qquad v_r^{\mathrm{lin}}(z) = \mathbb{E}[Y - X \mid (1-r)X + rY = z]$. $Z_t = a_tX + b_tY, \qquad Y \sim \mathcal{N}(0, \sigma^2\mathrm{Id})$, 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 τ decreases.

$a_t = (1-t)(1-2t), \qquad b_t = t,$

14.2 One-Step Generative Models

Training a one-step flow. Let ζ be a simple latent distribution and let $\alpha_\theta = (G_\theta)_\sharp\zeta$ be the model distribution. Assume that the target data distribution is β. A Wasserstein-flow construction chooses a discrepancy $\mathcal{E}_\beta(\alpha)$,

$$\partial_t\mu_t + \mathrm{div}(\mu_t w_t) = 0, \qquad w_t(x) = -\nabla\delta_\alpha\mathcal{E}_\beta(\mu_t)(x). \tag{14.14}$$

$$\min_\eta \int_0^1 \int \|U_\eta(t,x) - w_t(x)\|^2\,\mathrm{d}\mu_t(x)\,\mathrm{d}t. \tag{14.15}$$

$$\alpha_\theta^+ = (\mathrm{Id} + \tau U_\eta)_\sharp\alpha_\theta, \qquad \text{or equivalently} \qquad G_\theta^+(z) = G_\theta(z) + \tau U_\eta(G_\theta(z)).$$

Self-corrected drifting fields.

$$B_\varepsilon[\nu](x) \coloneqq \frac{\int (y-x)K_\varepsilon(x,y)\,\mathrm{d}\nu(y)}{\int K_\varepsilon(x,y)\,\mathrm{d}\nu(y)}. \tag{14.16}$$

For the Gaussian kernel $K_\varepsilon(x,y) = \exp(-\|x-y\|^2/(2\varepsilon))$, this normalized field is a score of a smoothed density:

$$B_\varepsilon[\nu](x) = \varepsilon\nabla \log\left(\int K_\varepsilon(x,y)\,\mathrm{d}\nu(y)\right). \tag{14.17}$$

$$u_t(x) = B_\varepsilon[\beta](x) - B_\varepsilon[\mu_t](x) = \varepsilon\nabla \log \frac{\int K_\varepsilon(x,y)\,\mathrm{d}\beta(y)}{\int K_\varepsilon(x,y)\,\mathrm{d}\mu_t(y)}. \tag{14.18}$$

Proposition 14.5 (Drifting as a time-dependent Wasserstein gradient)**.** *Let μ_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*

$$\mathcal{R}_t(\alpha|\mu_t) \coloneqq -\int \varphi_t(x)\,\mathrm{d}\alpha(x) + \int \varphi_t(x)\,\mathrm{d}\mu_t(x). \tag{14.19}$$

Here μ_t and φ_t are frozen when taking the first variation with respect to the first argument α. 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 μ_t and φ_t are fixed in the variation with respect to α, 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. □

14.3 Evolution in Depth of Transformers

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:

$x_i \mapsto x_i + \frac{1}{T}\sum_j \frac{e^{\langle Qx_i, Kx_j\rangle}Vx_j}{\sum_\ell e^{\langle Qx_i, Kx_\ell\rangle}}$, 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:

$x_i \mapsto x_i + \frac{1}{T}\Gamma_\theta[\alpha](x_i)$, $\Gamma_\theta[\alpha](x) \coloneqq \frac{\int e^{\langle Qx,Ky\rangle}Vy\,\mathrm{d}\alpha(y)}{\int e^{\langle Qx,Kz\rangle}\,\mathrm{d}\alpha(z)}$. $\alpha_{t+\tau} = (\mathrm{Id} + \tau\Gamma_{\theta_t}[\alpha_t])_\sharp\alpha_t$. $\partial_t\alpha_t + \operatorname{div}(\alpha_t\Gamma_{\theta_t}[\alpha_t]) = 0$.

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

$\Phi_\alpha(x) = \int \exp(\langle Qx, Ky\rangle)\mathrm{d}\alpha(y)$, $\quad U_\alpha(x) = \log\Phi_\alpha(x)$. $\nabla_x U_\alpha(x) = \frac{\int Q^\top Ky\,\exp(\langle Qx,Ky\rangle)\mathrm{d}\alpha(y)}{\int \exp(\langle Qx,Kz\rangle)\mathrm{d}\alpha(z)} = \Gamma_\theta[\alpha](x)$. This is an instantaneous gradient in x. It is not, however, the gradient of the first variation of a fixed functional of α, because the potential U_α itself depends on the current measure through the same attention normalization.

14.4 Flows over the Gaussian Manifold

Gaussianity preservation.

Proposition 14.6 (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 α_t remains Gaussian and its moments solve $\dot{\mathbf{m}}_t = b_t$, $\quad \dot{\mathbf{\Sigma}}_t = A_t\mathbf{\Sigma}_t + \mathbf{\Sigma}_t A_t^\top$. 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. $\mathcal{G} = \{\mathcal{N}(\mathbf{m},\mathbf{\Sigma}) : \mathbf{m}\in\mathbb{R}^d,\ \mathbf{\Sigma}\succ 0\}$ 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 $\mathcal{M}$.

$\alpha^{k+1} \in \operatorname*{argmin}_{\alpha\in\mathcal{M}} \frac{1}{2\tau}\mathcal{W}_2^2(\alpha,\alpha^k) + f(\alpha)$. 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 14.7 (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 14.6 gives (14.20). □

Gaussian-preserving gradient flows.

Proposition 14.8 (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*

$$\begin{aligned}
\mathrm{KL}(\mu|\gamma) \quad &: \quad h(C) = 2(\mathrm{Id} - C),\\
\tfrac{1}{2}\mathcal{I}(\mu|\gamma) \quad &: \quad h(C) = 2(C^{-1} - C),\\
\mathcal{W}_2^2(\mu, \gamma) \quad &: \quad h(C) = 4(C^{1/2} - C),\\
\mathrm{MMD}_k^2(\mu, \gamma), \quad k(x,y) = \langle x,\, y\rangle^2 \quad &: \quad h(C) = 8(C - C^2),\\
S_\varepsilon(\mu, \gamma) \quad &: \quad h(C) = 4\left(C + \tfrac{\varepsilon^2}{16}\mathrm{Id}\right)^{1/2} - 2\left(C^2 + \tfrac{\varepsilon^2}{16}\mathrm{Id}\right)^{1/2} - 2C - \tfrac{\varepsilon}{2}\mathrm{Id},\\
\mathrm{SW}_2^2(\mu, \gamma) \quad &: \quad h(C) = V(C)C + CV(C),
\end{aligned}$$

where S_ε is the debiased Sinkhorn divergence for the quadratic cost $\|x - y\|^2$ and KL regularization strength ε, 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 σ. 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 14.6, 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

$$\frac{1}{2}\left(\operatorname{tr}(C) + \operatorname{tr}(C^{-1}) - 2d\right).$$

Using Proposition 14.7 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 γ 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 $\dot{C} = V(C)C + CV(C)$. □

Non-variational Gaussian-preserving flows.

Contractive Gaussian projection.

Theorem 14.9 (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 μ. 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 μ and ν, 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 2.28. □

Theorem 14.10 (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 14.9. If γ is Gaussian and ν 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 γ is Gaussian. Hence, for any competitor η, $F(\mathcal{R}\eta) + \frac{1}{2\tau}\mathcal{W}_2^2(\gamma, \mathcal{R}\eta) \leqslant F(\eta) + \frac{1}{2\tau}\mathcal{W}_2^2(\gamma, \eta)$. Applying this to a minimizer $\eta = \nu$ shows that $\mathcal{R}\nu$ is again a minimizer. Uniqueness forces $\nu = \mathcal{R}\nu$. □