

Duality in Computational Optimal Transport

Lecture 2 — release notes

Romain Lopez

Monday 21 September 2026

These notes record what was derived in the second lecture, together with the background it assumed. §1 sets up the Monge and Kantorovich problems and the Wasserstein distance. §2 adds an entropic penalty, derives the dual, and reads the Sinkhorn iterations off it. §3 turns to the line, where optimal transport reduces to sorting. §4 asks what survives once the ordering is gone, and closes with Brenier’s theorem.

Contents

1	The optimal transport problem	2
1.1	Monge’s formulation	2
1.2	Kantorovich’s relaxation	3
2	Entropic optimal transport	5
2.1	The entropic OT problem	5
2.2	Duality	8
2.3	Sinkhorn iterations	10
3	Optimal transport on the line	12
3.1	Convex costs reward non-crossing	12
3.2	Optimal transport on the line is sorting	12
4	The geometry of optimal plans beyond the line	15
4.1	Cyclical monotonicity	15
4.2	Brenier’s theorem	15

1 The optimal transport problem

Optimal transport asks for the cheapest way to rearrange one distribution of mass into another. This section fixes the two classical formulations — Monge’s, which asks for a map, and Kantorovich’s, which asks for a coupling — and shows that the resulting cost defines a metric on distributions.

1.1 Monge’s formulation

Optimal transport asks for the cheapest way to rearrange one distribution of mass into another. We make this precise starting from the operation that moves mass: the pushforward.

Definition 1.1 (Pushforward). Let $T: \mathbb{R}^d \rightarrow \mathbb{R}^d$ be measurable and $\mu \in \mathcal{P}(\mathbb{R}^d)$. The *pushforward* $T_{\#}\mu$ is the law of $T(X)$ when $X \sim \mu$; equivalently, $(T_{\#}\mu)(A) = \mu(T^{-1}(A))$ for measurable A , or in integrated form

$$\int_{\mathbb{R}^d} g \, d(T_{\#}\mu) = \int_{\mathbb{R}^d} g(T(x)) \, d\mu(x) \quad \text{for all bounded measurable } g.$$

Definition 1.2 (Transport map). A map T *transports* μ to ν if $T_{\#}\mu = \nu$. It relocates the mass sitting at each x to the point $T(x)$, and the constraint $T_{\#}\mu = \nu$ demands that the relocated mass reconstitute ν exactly.

A transport map is thus one admissible rearrangement of μ into ν , and there will in general be many. We may now ask for the cheapest, which first requires saying what movement costs.

Definition 1.3 (Monge problem). Let $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$, and fix a *cost* $c: \mathbb{R}^d \times \mathbb{R}^d \rightarrow [0, \infty)$, where $c(x, y)$ is the price of moving one unit of mass from x to y . The *Monge problem* is to find the cheapest transport map from μ to ν :

$$\inf_T \int_{\mathbb{R}^d} c(x, T(x)) \, d\mu(x) \quad \text{subject to} \quad T_{\#}\mu = \nu, \quad (1)$$

the infimum being over measurable maps $T: \mathbb{R}^d \rightarrow \mathbb{R}^d$.

Two difficulties follow, and between them they motivate everything after this section. The first is that the problem is awkward to optimize; the second, more serious, is that it may have nothing to optimize over at all.

Remark 1.4 (The problem is nonconvex). The operation $T \mapsto T_{\#}\mu$ is nonlinear, and the feasible set $\{T : T_{\#}\mu = \nu\}$ is not convex: averaging two transport maps generally gives a map that no longer pushes μ onto ν . So (1) is a nonconvex problem in an infinite-dimensional variable, with none of the guarantees — a unique optimum, a certificate of optimality, a convergent algorithm — that convexity would supply.

Remark 1.5 (A map need not exist). The second difficulty is more basic: the feasible set can be *empty*. A map sends each point to a single destination, so it can move mass but never divide it; formally, the pushforward of a point mass is again a point mass, $T_{\#}\delta_{x_0} = \delta_{T(x_0)}$.

Take

$$\mu = \delta_0 \quad \text{and} \quad \nu = \frac{1}{2}\delta_1 + \frac{1}{2}\delta_2.$$

Every candidate satisfies $T_{\#}\delta_0 = \delta_{T(0)}$, a single Dirac, which can never equal a two-atom measure. No transport map exists at all, and (1) is an infimum over the empty set.

Nor is this exotic. It occurs whenever μ carries an atom heavier than any atom of ν — in particular for empirical measures supported on fewer points than their target. It is exactly what the Kantorovich relaxation (§1.2) repairs, by letting the mass at 0 split and send half to 1 and half to 2.

1.2 Kantorovich’s relaxation

Kantorovich’s idea is to stop insisting that all the mass at x go to a single destination. Allow it to be *split*, and describe the transport by a joint distribution instead of a map.

Definition 1.6 (Coupling). A *coupling* (or *transport plan*) of $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$ is a probability measure π on $\mathbb{R}^d \times \mathbb{R}^d$ whose marginals are μ and ν :

$$\pi(A \times \mathbb{R}^d) = \mu(A), \quad \pi(\mathbb{R}^d \times B) = \nu(B).$$

The set of couplings is $\Pi(\mu, \nu)$. The number $d\pi(x, y)$ is the amount of mass sent from x to y ; the marginal constraints say “all of μ leaves” and “all of ν arrives”.

Definition 1.7 (Kantorovich problem). Let $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$ and let c be a cost as in Definition 1.3. The *Kantorovich problem* is to minimize the expected cost over all couplings of μ and ν :

$$\text{OT}_c(\mu, \nu) := \inf_{\pi \in \Pi(\mu, \nu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} c(x, y) d\pi(x, y), \quad (2)$$

where $\Pi(\mu, \nu)$ is the set of couplings of μ and ν .

Remark 1.8 (Mass may now split). This is the entire difference from (1). A map sends all the mass at x to the single point $T(x)$; a coupling may divide it, sending part to one destination and part to another. The pair $\mu = \delta_0$, $\nu = \frac{1}{2}\delta_1 + \frac{1}{2}\delta_2$ of Remark 1.5, for which no transport map exists, is now solved by the coupling that sends half the mass at 0 to 1 and half to 2.

Remark 1.9 (Existence of an optimal plan). Unlike Monge, Kantorovich is *always feasible*: the product $\mu \otimes \nu$ is a coupling, so $\Pi(\mu, \nu)$ is never empty. Under mild conditions on c the infimum in (2) is moreover *attained*; see [Santambrogio, 2015, Villani, 2009].

To see that (2) really is a relaxation of (1), observe that every transport map induces a coupling. Given T with $T_{\#}\mu = \nu$, set

$$\pi_T := (\text{id}, T)_{\#}\mu, \quad (3)$$

the *deterministic* coupling supported on the graph of T . Its marginals are μ and $T_{\#}\mu = \nu$, so $\pi_T \in \Pi(\mu, \nu)$, and by the defining identity of the pushforward its cost is $\int c d\pi_T = \int c(x, T(x)) d\mu$

— the Monge cost of T . Every Monge candidate is therefore a Kantorovich candidate of equal cost, and minimizing over the larger class can only do better:

$$\text{OT}_c(\mu, \nu) \leq \inf_T \int c(x, T(x)) \, d\mu.$$

The transport cost can also be used to put a genuine *distance* on the space of distributions, which is what makes it more than a measure of dissimilarity: one can speak of convergence, of interpolation, of one distribution lying between two others. Taking $c(x, y) = |x - y|^p$ and correcting by a p -th root does it.

Definition 1.10 (p -Wasserstein distance). Let $p \geq 1$ and let $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$ have finite p -th moment. For the cost $c(x, y) = |x - y|^p$, the p -Wasserstein distance is

$$\mathcal{W}_p(\mu, \nu) := \text{OT}_c(\mu, \nu)^{1/p}. \quad (4)$$

This defines a metric on the set of such measures.

Proof. (i) Finiteness. We must check $\mathcal{W}_p(\mu, \nu) < \infty$, since an infimum of an unbounded quantity would be of little use. Bound the infimum by the product coupling $\mu \otimes \nu$, under which x and y are independent. Convexity of $t \mapsto t^p$ gives the elementary bound $|x - y|^p \leq (|x| + |y|)^p \leq 2^{p-1}(|x|^p + |y|^p)$, so

$$\mathcal{W}_p^p(\mu, \nu) \leq \int |x - y|^p \, d(\mu \otimes \nu) \leq 2^{p-1} \left(\int |x|^p \, d\mu + \int |y|^p \, d\nu \right) < \infty,$$

the last step being exactly the assumption that μ and ν have finite p -th moment. This is where that hypothesis is used, and the only place.

(ii) Symmetry. Let $S(x, y) := (y, x)$ be the swap. If $\pi \in \Pi(\mu, \nu)$ then $S_\# \pi$ has its two marginals exchanged, so $S_\# \pi \in \Pi(\nu, \mu)$; and S is its own inverse, so $\pi \mapsto S_\# \pi$ is a bijection between the two sets of couplings. It preserves the cost, because the ground distance is symmetric:

$$\int |x - y|^p \, d(S_\# \pi) = \int |y - x|^p \, d\pi = \int |x - y|^p \, d\pi.$$

The two problems therefore range over the same set of values, and their infima agree.

(iii) Vanishing exactly on the diagonal. If $\mu = \nu$, the coupling $(\text{id}, \text{id})_\# \mu$ places all mass on $\{x = y\}$ and has cost 0, so $\mathcal{W}_p(\mu, \mu) = 0$. Conversely suppose $\mathcal{W}_p(\mu, \nu) = 0$. The infimum is attained (Remark 1.9), say by π , so $\int |x - y|^p \, d\pi = 0$. A nonnegative integrand with vanishing integral is zero almost everywhere, so π is concentrated on the diagonal. But a measure on the diagonal has equal marginals, and those marginals are μ and ν ; hence $\mu = \nu$.

(iv) Triangle inequality. We prove this for finitely supported measures, where the construction it needs is an explicit formula. The general case runs the identical argument with conditional probabilities in place of the quotient below; we refer to [Villani, 2009] for it.

Write

$$\mu_1 = \sum_i a_i \delta_{x_i}, \quad \mu_2 = \sum_j b_j \delta_{y_j}, \quad \mu_3 = \sum_k c_k \delta_{z_k},$$

let P be an *optimal* plan from μ_1 to μ_2 and Q an *optimal* plan from μ_2 to μ_3 , so that $P\mathbf{1} = a$, $Q^\top \mathbf{1} = b = Q\mathbf{1}$ and $Q^\top \mathbf{1} = c$. Discarding any j with $b_j = 0$, define a distribution on triples by

$$\sigma_{ijk} := \frac{P_{ij} Q_{jk}}{b_j} \geq 0. \quad (5)$$

This says: route the mass into y_j as P prescribes, then split it onward in exactly the proportions Q_{jk}/b_j that Q prescribes — the two legs are made conditionally independent given the middle point. Both two-index marginals come out right,

$$\sum_k \sigma_{ijk} = \frac{P_{ij}}{b_j} \sum_k Q_{jk} = \frac{P_{ij}}{b_j} b_j = P_{ij}, \quad \sum_i \sigma_{ijk} = \frac{Q_{jk}}{b_j} \sum_i P_{ij} = \frac{Q_{jk}}{b_j} b_j = Q_{jk},$$

so σ carries P and Q simultaneously, and $R_{ik} := \sum_j \sigma_{ijk}$ is a coupling of μ_1 and μ_3 . Bounding the infimum by this particular coupling, and then applying the triangle inequality in $\mathbb{R}^d$ inside the sum,

$$\mathcal{W}_p(\mu_1, \mu_3) \leq \left(\sum_{ijk} \sigma_{ijk} |x_i - z_k|^p \right)^{1/p} \leq \left(\sum_{ijk} \sigma_{ijk} (|x_i - y_j| + |y_j - z_k|)^p \right)^{1/p}. \quad (6)$$

Every sum here is finite, so the right-hand side is an ordinary p -norm on $\mathbb{R}^N$, with N the number of triples (i, j, k) . Indeed, setting

$$u_{ijk} := \sigma_{ijk}^{1/p} |x_i - y_j|, \quad v_{ijk} := \sigma_{ijk}^{1/p} |y_j - z_k|,$$

the right-hand side of (6) is exactly $\|u + v\|_p$, and Minkowski's inequality in $\mathbb{R}^N$ gives $\|u + v\|_p \leq \|u\|_p + \|v\|_p$, that is

$$\mathcal{W}_p(\mu_1, \mu_3) \leq \left(\sum_{ijk} \sigma_{ijk} |x_i - y_j|^p \right)^{1/p} + \left(\sum_{ijk} \sigma_{ijk} |y_j - z_k|^p \right)^{1/p}.$$

Both terms collapse. Summing out the index that does not appear, and using the marginal identities computed above,

$$\sum_{ijk} \sigma_{ijk} |x_i - y_j|^p = \sum_{ij} \left(\sum_k \sigma_{ijk} \right) |x_i - y_j|^p = \sum_{ij} P_{ij} |x_i - y_j|^p = \mathcal{W}_p^p(\mu_1, \mu_2),$$

the last equality because P was chosen optimal; the second term is $\mathcal{W}_p^p(\mu_2, \mu_3)$ in the same way. $\square$

2 Entropic optimal transport

2.1 The entropic OT problem

Two finite families of points carry the mass: the *sources* $x_1, \dots, x_n$ and the *targets* $y_1, \dots, y_m$. Their weights are probability vectors, living on the simplices

$$\Delta_n := \left\{ a \in \mathbb{R}_{\geq 0}^n : \sum_{i=1}^n a_i = 1 \right\}, \quad \Delta_m := \left\{ b \in \mathbb{R}_{\geq 0}^m : \sum_{j=1}^m b_j = 1 \right\},$$

so that a_i is the mass sitting at x_i and b_j the mass wanted at y_j .

Definition 2.1 (Transport polytope). Let $a \in \Delta_n$ and $b \in \Delta_m$. The *transport polytope* of a and b is

$$U(a, b) := \{P \in \mathbb{R}_{\geq 0}^{n \times m} : P\mathbf{1}_m = a, P^\top \mathbf{1}_n = b\},$$

the set of plans with marginals a and b : all of a leaves, and all of b arrives.

A *cost* $c(x, y)$ prices the move of one unit of mass from x to y ; collect its values into the *cost matrix* $C \in \mathbb{R}^{n \times m}$ with $C_{ij} = c(x_i, y_j)$, and write $\langle C, P \rangle := \sum_{ij} C_{ij} P_{ij}$ for the entrywise (Frobenius) inner product.

Definition 2.2 (Kantorovich problem). Let $a \in \Delta_n$, $b \in \Delta_m$ and $C \in \mathbb{R}^{n \times m}$. The *Kantorovich problem* is the linear program

$$\text{OT}_C(a, b) := \min_{P \in U(a, b)} \langle C, P \rangle. \quad (7)$$

Proposition 2.3 (The Kantorovich problem is well posed). *For every $a \in \Delta_n$, $b \in \Delta_m$ and $C \in \mathbb{R}^{n \times m}$, the transport polytope $U(a, b)$ is nonempty, convex and compact, and the minimum in (7) is attained: there is at least one $P^* \in U(a, b)$ with $\langle C, P^* \rangle = \text{OT}_C(a, b)$.*

Proof. The objective $P \mapsto \langle C, P \rangle$ is linear, hence continuous.

The polytope is an intersection of closed sets — the $n + m$ hyperplanes $\{\sum_j P_{ij} = a_i\}$ and $\{\sum_i P_{ij} = b_j\}$, and the nm half-spaces $\{P_{ij} \geq 0\}$ — so it is closed. It is bounded in the infinity norm: the entries being nonnegative, every one of them obeys $0 \leq P_{ij} \leq \sum_{j'} P_{ij'} = a_i \leq 1$, so $\|P\|_\infty \leq 1$ on $U(a, b)$. Closed and bounded in $\mathbb{R}^{n \times m}$, hence compact.

A continuous function on $U(a, b)$ therefore attains its minimum, provided only that $U(a, b)$ is not empty. And it is not: the *independent coupling* $ab^\top$, with entries $(ab^\top)_{ij} = a_i b_j \geq 0$, has row sums $\sum_j a_i b_j = a_i \sum_j b_j = a_i$ and column sums $\sum_i a_i b_j = b_j \sum_i a_i = b_j$, so $ab^\top \in U(a, b)$. Whatever the cost, one may always ship mass in proportion to demand. $\square$

Remark 2.4 (The minimiser need not be unique). Proposition 2.3 gives existence and nothing more. If the cost matrix C is constant, every $P \in U(a, b)$ has the same cost, so every feasible plan is optimal at once. Recovering uniqueness is one of the things the entropic penalty will buy.

We now introduce the *entropy* of a transport plan,

$$H(P) := - \sum_{ij} P_{ij} (\log P_{ij} - 1),$$

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

Definition 2.5 (Entropic optimal transport). Let $a \in \Delta_n$, $b \in \Delta_m$, $C \in \mathbb{R}^{n \times m}$ and $\varepsilon > 0$. The *entropic optimal transport* problem is

$$\text{OT}_C^\varepsilon(a, b) := \min_{P \in U(a, b)} \langle C, P \rangle - \varepsilon H(P). \quad (8)$$

Remark 2.6 (The two extremes of ε). As $\varepsilon \rightarrow 0$ the penalty vanishes and the minimiser of (8) converges to a minimiser of the linear program (7) — specifically to the one of largest entropy among them, which is a genuine selection, since that program's optimum need not be unique (Remark 2.4). As $\varepsilon \rightarrow \infty$ the penalty dominates, and the optimal plan degenerates to the independent coupling $ab^\top$ of Proposition 2.3, which ignores the cost entirely.

Proposition 2.7 (A unique interior solution). *For every $a \in \Delta_n$, $b \in \Delta_m$, $C \in \mathbb{R}^{n \times m}$ and $\varepsilon > 0$, the minimum in (8) is attained at exactly one $P^\star \in U(a, b)$; and if $a, b > 0$, then $P_{ij}^\star > 0$ for every i, j , so $P^\star$ lies in the relative interior of the polytope.*

Proof. (i) **Existence and uniqueness.** The objective separates across the entries,

$$\langle C, P \rangle - \varepsilon H(P) = \sum_{ij} \phi_{ij}(P_{ij}), \quad \phi_{ij}(z) := C_{ij} z + \varepsilon z(\log z - 1),$$

each ϕ_{ij} being continuous on $[0, \infty)$ with the convention $0 \log 0 = 0$. Since $\phi_{ij}''(z) = \varepsilon/z > 0$ for $z > 0$, every ϕ_{ij} is strictly convex, and hence so is their sum.

The objective is therefore a continuous, strictly convex function on $U(a, b)$, which by Proposition 2.3 is nonempty, compact and convex. Continuity and compactness give a minimiser, exactly as before; strict convexity makes it unique, since if $P \neq Q$ were both minimal their midpoint would lie in $U(a, b)$ and have strictly smaller value.

(ii) **Interiority.** Assume now $a, b > 0$, and suppose $P_{ij}^\star = 0$ for some pair (i, j) . Consider the path running from $P^\star$ to the independent coupling,

$$P_t := (1 - t) P^\star + t ab^\top, \quad t \in [0, 1],$$

which stays inside $U(a, b)$ because that set is convex, and write

$$F(t) := \sum_{kl} \phi_{kl}((P_t)_{kl})$$

for the objective along it.

Optimality of $P^\star$ says exactly that F is minimised at $t = 0$. Since the objective is convex and $t \mapsto P_t$ is affine, F is convex; and a convex function minimised at the left endpoint of an interval is nondecreasing on it. So we must have

$$F'(t) \geq 0 \quad \text{for every } t \in (0, 1). \tag{9}$$

Now compute F' near 0. Since $\phi'_{kl}(z) = C_{kl} + \varepsilon \log z$,

$$\begin{aligned} F'(t) &= \sum_{kl} \phi'_{kl}((P_t)_{kl}) (a_k b_l - P_{kl}^\star) \\ &= \sum_{kl} (C_{kl} + \varepsilon \log(P_t)_{kl}) (a_k b_l - P_{kl}^\star). \end{aligned}$$

Any term with $P_{kl}^\star > 0$ stays bounded as $t \rightarrow 0^+$: being a convex combination of $P_{kl}^\star$ and $a_k b_l$, both positive, $(P_t)_{kl} \geq \min(P_{kl}^\star, a_k b_l) > 0$ uniformly in t . Any term with $P_{kl}^\star = 0$, such as $P_{ij}^\star$,

does not: there $(P_t)_{kl} = t a_k b_l$, so $\log(P_t)_{kl} \rightarrow -\infty$ while its factor $a_k b_l$ is a fixed positive number. Every divergent term carries the same sign, so

$$F'(t) \rightarrow -\infty \quad \text{as } t \rightarrow 0^+,$$

which contradicts (9). Hence $P_{ij}^* > 0$ for every i, j . $\square$

The entropic plan is therefore *dense* where the optimum of (7) is sparse.

2.2 Duality

Proposition 2.8 (The entropic dual). *Dualising the marginal constraints of (8) yields the unconstrained maximisation*

$$\text{OT}_C^\varepsilon(a, b) = \max_{f \in \mathbb{R}^n, g \in \mathbb{R}^m} \langle a, f \rangle + \langle b, g \rangle - \varepsilon \sum_{ij} \exp\left(\frac{f_i + g_j - C_{ij}}{\varepsilon}\right). \quad (10)$$

The objective is smooth and concave, and the maximum is attained.

Proof. (i) **Introducing the Lagrangian.** Dualise the two marginal constraints only, keeping $P > 0$ as the domain — legitimate because Proposition 2.7 places the minimiser in the interior. Attach a multiplier f_i to each row constraint and g_j to each column constraint; these $n + m$ numbers are the *dual potentials*, one per source and one per target. With $f \in \mathbb{R}^n$ and $g \in \mathbb{R}^m$,

$$L(P, f, g) = \langle C, P \rangle - \varepsilon H(P) - \sum_i f_i((P\mathbf{1})_i - a_i) - \sum_j g_j((P^\top \mathbf{1})_j - b_j).$$

Maximising L over f and g recovers the two constraints we dropped. If P violates a marginal, the corresponding multiplier can be scaled without bound and L is driven to $+\infty$; if P satisfies both, every multiplier term vanishes. Hence

$$\sup_{f, g} L(P, f, g) = \begin{cases} \langle C, P \rangle - \varepsilon H(P), & P \in U(a, b), \\ +\infty, & \text{otherwise,} \end{cases} \quad \text{so} \quad \min_{P > 0} \sup_{f, g} L(P, f, g) = \text{OT}_C^\varepsilon(a, b).$$

That is what the Lagrangian buys: a hard constraint has become a maximisation, and the dual is this same object with the order of min and sup exchanged.

(ii) **Deriving a separable form.** Now write everything out entrywise. We have $\langle C, P \rangle = \sum_{ij} C_{ij} P_{ij}$ and $-\varepsilon H(P) = \varepsilon \sum_{ij} P_{ij}(\log P_{ij} - 1)$, while expanding the marginals gives $(P\mathbf{1})_i = \sum_j P_{ij}$ and $(P^\top \mathbf{1})_j = \sum_i P_{ij}$, so

$$L(P, f, g) = \sum_{ij} C_{ij} P_{ij} + \varepsilon \sum_{ij} P_{ij}(\log P_{ij} - 1) - \sum_{ij} f_i P_{ij} + \sum_i f_i a_i - \sum_{ij} g_j P_{ij} + \sum_j g_j b_j.$$

Abbreviating $s_{ij} := f_i + g_j - C_{ij}$ for the coefficient that appears,

$$L(P, f, g) = \langle a, f \rangle + \langle b, g \rangle + \sum_{ij} \left(\varepsilon P_{ij}(\log P_{ij} - 1) - s_{ij} P_{ij} \right). \quad (11)$$

(iii) **Strong duality applies.** Step (i) wrote the primal as $\min_{P>0} \sup_{f,g} L$. By *Slater's condition*, for a convex programme whose constraints are affine and which has one feasible point in the interior of the domain,

$$\min_{P>0} \sup_{f,g} L(P, f, g) = \max_{f,g} \inf_{P>0} L(P, f, g), \quad (12)$$

with both sides attained. All three hypotheses hold here: the objective is strictly convex by Proposition 2.7(i), the constraints $P\mathbf{1} = a$ and $P^\top \mathbf{1} = b$ are affine, and when $a, b > 0$ the independent coupling $ab^\top$ of Proposition 2.3 is feasible with every entry $a_i b_j > 0$, hence interior to the domain $\{P > 0\}$. The left-hand side of (12) is $\text{OT}_C^\varepsilon(a, b)$; the right-hand side is what we now compute.

(iv) **The inner minimisation.** The sum in (11) decouples across entries, and each term $\varepsilon z(\log z - 1) - s_{ij}z$ is strictly convex on $z > 0$ with derivative $\varepsilon \log z - s_{ij}$. This vanishes at $z = e^{s_{ij}/\varepsilon}$, where the term equals $-\varepsilon z$. Hence

$$\inf_{P>0} L(P, f, g) = \langle a, f \rangle + \langle b, g \rangle - \varepsilon \sum_{ij} e^{s_{ij}/\varepsilon}, \quad \text{attained at } P_{ij} = e^{s_{ij}/\varepsilon}, \quad (13)$$

Maximising this over (f, g) and invoking (12) gives (10). Concavity is then free: by (13) the objective is an infimum of functions affine in (f, g) , and an infimum of affine functions is concave. $\square$

Remark 2.9 (What dualising bought). In (11) the Lagrangian is *separable* in the entries of P : no term couples P_{ij} to P_{kl} . All the coupling lived in the marginal constraints, and dualising is exactly what removed it. That is why step (iv) is a one-variable calculation done once rather than an nm -dimensional one, and it is what makes the algorithm of §2.3 possible.

Remark 2.10 (The sign of $f_i + g_j - C_{ij}$). The dual (10) is *unconstrained*, so for an arbitrary pair (f, g) the quantity $s_{ij} = f_i + g_j - C_{ij}$ may perfectly well be positive. This deserves stressing, because the linear-programming dual carries $f_i + g_j \leq C_{ij}$ as a hard constraint, and that same inequality does reappear here — but as a *conclusion*, and only at the optimum. At a maximiser of (10), differentiating in f_i and setting the result to zero gives

$$\sum_j e^{s_{ij}/\varepsilon} = a_i, \quad (14)$$

which is nothing but the first marginal constraint. Its left-hand side is a sum of strictly positive numbers, so each one of them is at most a_i ; taking the column condition as well,

$$e^{s_{ij}/\varepsilon} \leq \min(a_i, b_j) \leq 1 \quad \implies \quad s_{ij} \leq \varepsilon \log \min(a_i, b_j) \leq 0,$$

strictly negative unless $n = m = 1$. For uniform marginals this reads $s_{ij} \leq -\varepsilon \log n$, so the gap grows with the size of the problem.

The intuition is that s_{ij}/ε is the logarithm of a *probability*: by (13) the plan induced by (f, g) has entries $e^{s_{ij}/\varepsilon}$, and one entry cannot carry more than the single unit of mass being transported. Were s_{ij} positive, that one route would ship more than the whole problem. The normalisation is the ceiling, and nothing has been imposed by hand.

So the two duals reach the same inequality by different means, and the entropic one *measures* slack where the linear program *enforces* it: $s_{ij} = \varepsilon \log P_{ij}^*$, the logarithm of the mass a route actually carries. Letting $\varepsilon \rightarrow 0$ recovers complementary slackness. Where the unregularised optimum puts mass, P_{ij}^* stays bounded away from 0 and $s_{ij} \rightarrow 0$: the constraint goes tight. Where it puts none, P_{ij}^* decays exponentially in $1/\varepsilon$ and s_{ij} tends to a strictly negative limit.

2.3 Sinkhorn iterations

Proposition 2.8 left an unconstrained, smooth, concave maximisation in (f, g) . Maximising it in one block with the other held fixed can be done exactly, so the obvious scheme — alternate the two — is available at once. Write

$$\Phi(f, g) := \langle a, f \rangle + \langle b, g \rangle - \varepsilon \sum_{ij} \exp\left(\frac{f_i + g_j - C_{ij}}{\varepsilon}\right)$$

for the dual objective of (10).

Proposition 2.11 (Block coordinate ascent, in closed form). *Each partial maximisation of Φ has a closed form. For fixed g , the maximiser f^* of $\Phi(\cdot, g)$ is given by*

$$e^{f_i^*/\varepsilon} = \frac{a_i}{\sum_j e^{(g_j - C_{ij})/\varepsilon}}, \quad i = 1, \dots, n, \quad (15)$$

and symmetrically, for fixed f , the maximiser g^* of $\Phi(f, \cdot)$ is given by

$$e^{g_j^*/\varepsilon} = \frac{b_j}{\sum_i e^{(f_i - C_{ij})/\varepsilon}}, \quad j = 1, \dots, m. \quad (16)$$

Proof. Factor the exponential as $e^{(f_i + g_j - C_{ij})/\varepsilon} = e^{f_i/\varepsilon} e^{(g_j - C_{ij})/\varepsilon}$, so that with g held fixed the f -dependent part of Φ separates across i :

$$\Phi(f, g) = \sum_i \left(a_i f_i - \varepsilon e^{f_i/\varepsilon} \sum_j e^{(g_j - C_{ij})/\varepsilon} \right) + \text{terms free of } f.$$

Each summand is strictly concave in f_i , with derivative $a_i - e^{f_i/\varepsilon} \sum_j e^{(g_j - C_{ij})/\varepsilon}$. Setting it to zero gives (15), whose right-hand side is a finite positive number, so f_i^* is well defined. The second claim is the same computation with the roles exchanged. $\square$

Running block coordinate ascent means applying (15) and (16) alternately, and the sums they contain are matrix–vector products in disguise. Write

$$K := \exp(-C/\varepsilon), \quad u_i := e^{f_i/\varepsilon}, \quad v_j := e^{g_j/\varepsilon},$$

all entrywise, and call K the *Gibbs kernel* and u, v the *scalings*. Then $\sum_j e^{(g_j - C_{ij})/\varepsilon} = \sum_j K_{ij} v_j = (Kv)_i$ and likewise $\sum_i e^{(f_i - C_{ij})/\varepsilon} = (K^\top u)_j$. Since (15) and (16) are already written in exponentiated form, they read off directly as the *Sinkhorn iterations*

$$u \leftarrow a \oslash (Kv), \quad v \leftarrow b \oslash (K^\top u), \quad (17)$$

with $\oslash$ entrywise division. Every logarithm and exponential has gone.

Remark 2.12 (Cost per iteration). One round of (17) forms Kv and then $K^\top u$: two matrix–vector products, so $2nm$ multiply–adds and $n + m$ divisions, and $\mathcal{O}(nm)$ time and memory — the only object that must be held is K itself, which can overwrite the cost matrix in place. By contrast the exact linear program (7) costs $\mathcal{O}(n^3 \log n)$ with network-simplex or auction methods. Both half-steps are dense matrix–vector products, so they map directly onto GPU kernels and batch across problem instances.

Theorem 2.13 (Convergence). *The Sinkhorn iterations (17) converge, and their limit (u^*, v^*) corresponds to a maximiser (f^*, g^*) of Φ . The convergence is linear: the Sinkhorn map is a contraction in the Hilbert projective metric.*

Not proved here. The iteration goes back to [Sinkhorn, 1964], and its use for optimal transport to [Cuturi, 2013]; the geometric rate via Hilbert’s projective metric is [Franklin & Lorenz, 1989], and the convergence section of [Peyré & Cuturi, 2019] gives a modern account.

Proposition 2.14 (Recovering the plan). *The unique optimal plan P^* of (8) satisfies*

$$P_{ij}^* = \exp\left(\frac{f_i + g_j - C_{ij}}{\varepsilon}\right) \quad (18)$$

for some $f \in \mathbb{R}^n$ and $g \in \mathbb{R}^m$.

Proof. By Proposition 2.7 the optimum P^* exists, is unique, and has every entry strictly positive, so no nonnegativity constraint is active there: the only active constraints are the $n + m$ marginal equalities, and there are multipliers f, g for which $\partial L / \partial P_{ij} = 0$ at P^* . Since

$$\frac{\partial}{\partial P_{ij}} \left(\langle C, P \rangle - \varepsilon H(P) \right) = C_{ij} + \varepsilon \log P_{ij},$$

that condition reads $C_{ij} + \varepsilon \log P_{ij}^* = f_i + g_j$, which is (18). $\square$

Those multipliers are dual optimal — by (12) the two problems share their optima — so they are what Theorem 2.13 converges to, and in the scalings u, v equation (18) is the matrix $\text{diag}(u) K \text{diag}(v)$ that Algorithm 1 returns.

Algorithm 1 Sinkhorn (entropic OT)

Input: cost C , marginals a, b , regularisation ε , iteration budget L

- 1: $K \leftarrow \exp(-C/\varepsilon); \quad v \leftarrow \mathbf{1}_m$
 - 2: **for** $\ell = 1, \dots, L$ **do**
 - 3: $u \leftarrow a \oslash (Kv)$
 - 4: $v \leftarrow b \oslash (K^\top u)$
 - 5: **end for**
 - 6: **return** $P \leftarrow \text{diag}(u) K \text{diag}(v)$
-

Remark 2.15 (What is actually run). For small ε the entries of $K = \exp(-C/\varepsilon)$ underflow, and with them Kv ; the divisions in Algorithm 1 then fail. Implementations therefore iterate Proposition 2.11 directly, in the log domain, evaluating $\log \sum_j e^{(g_j - C_{ij})/\varepsilon}$ with a stabilised log-sum-exp. The matrix form is the one to reason about; the log form is the one to run.

3 Optimal transport on the line

3.1 Convex costs reward non-crossing

From here on the sources and targets are *real numbers*, $x_1, \dots, x_n \in \mathbb{R}$ and $y_1, \dots, y_m \in \mathbb{R}$, and we take the quadratic cost $c(x, y) = |x - y|^2$, so that $C_{ij} = |x_i - y_j|^2$. The goal is to show that an optimal plan on the line is *monotone*: it never sends a smaller source to a larger target while sending a larger source to a smaller one. Everything rests on one inequality, which says that in one dimension crossing is never worthwhile.

Lemma 3.1 (Uncrossing pays, for the quadratic cost). *Let $x_1 \leq x_2$ and $y_1 \leq y_2$ in $\mathbb{R}$. Then uncrossing costs no more than crossing:*

$$\underbrace{|x_1 - y_1|^2 + |x_2 - y_2|^2}_{\text{uncrossed}} \leq \underbrace{|x_1 - y_2|^2 + |x_2 - y_1|^2}_{\text{crossed}}. \quad (19)$$

Proof. Expand the four squares. Every x_i^2 and every y_j^2 occurs once on each side and cancels, so the crossed cost minus the uncrossed cost is exactly

$$-2x_1y_2 - 2x_2y_1 + 2x_1y_1 + 2x_2y_2 = 2(x_1 - x_2)(y_1 - y_2) = 2(x_2 - x_1)(y_2 - y_1) \geq 0,$$

both factors being nonnegative. $\square$

Remark 3.2 (Any convex cost will do). Nothing above is special to the square. For any convex $h: \mathbb{R} \rightarrow \mathbb{R}$ and any $x_1 \leq x_2, y_1 \leq y_2$,

$$h(x_1 - y_1) + h(x_2 - y_2) \leq h(x_1 - y_2) + h(x_2 - y_1), \quad (20)$$

which covers $|x - y|^p$ for every $p \geq 1$. Everything in B.2 uses only (20), so it holds at that generality.

3.2 Optimal transport on the line is sorting

It is convenient now to name the two *distributions* rather than their weight vectors. Write δ_x for the *Dirac measure* at x — all of the mass placed at the single point x — and take $n = m$ with equal weights $a = b = \frac{1}{n}\mathbf{1}_n$, so that

$$\mu := \frac{1}{n} \sum_{i=1}^n \delta_{x_i}, \quad \nu := \frac{1}{n} \sum_{i=1}^n \delta_{y_i}$$

are two empirical measures on $\mathbb{R}$ with the same number of equally weighted atoms. A plan $P \in U(a, b)$ is then called a *coupling* of μ and ν , and its *support* is the set of pairs (x_i, y_j) with $P_{ij} > 0$.

Theorem 3.3 (Optimal transport on the line is sorting). *Let $x_1 < \dots < x_n$ and $y_1 < \dots < y_n$ in $\mathbb{R}$, let μ, ν be as above, and take the quadratic cost. Then (7) has a unique optimal plan, namely*

$$P^* = \frac{1}{n} I, \quad \text{which matches } x_i \text{ to } y_i,$$

and consequently

$$\text{OT}_C(a, b) = \frac{1}{n} \sum_{i=1}^n |x_i - y_i|^2. \quad (21)$$

Proof. The two steps use different tools. Step (i) uses the cost, through Lemma 3.1, to constrain the support of *any* optimal plan. Step (ii) forgets the cost entirely and counts mass; it needs no optimality at all.

(i) Every optimal plan has monotone support. Let P be optimal, and suppose it carries a *crossing pair*:

$$P_{i_1 j_1} > 0, \quad P_{i_2 j_2} > 0, \quad i_1 < i_2, \quad j_1 > j_2.$$

The competitor. Put $\lambda := \min(P_{i_1 j_1}, P_{i_2 j_2}) > 0$ and

$$Q := P + \lambda(E_{i_1 j_2} + E_{i_2 j_1} - E_{i_1 j_1} - E_{i_2 j_2}),$$

where E_{ij} has a single 1 in position (i, j) and zeros elsewhere.

Q is admissible. Only the entries (i_1, j_1) and (i_2, j_2) decrease, each by λ , so by the choice of λ ,

$$Q \geq 0.$$

Each of the rows i_1, i_2 gains λ in one entry and loses λ in another, and so does each of the columns j_1, j_2 ; the perturbation is a 2×2 cycle. Hence

$$Q\mathbf{1}_m = P\mathbf{1}_m = a, \quad Q^\top \mathbf{1}_n = P^\top \mathbf{1}_n = b, \quad \text{so} \quad Q \in U(a, b).$$

Q is strictly cheaper. Only four entries differ, so

$$\begin{aligned} \langle C, Q \rangle - \langle C, P \rangle &= \lambda \left[|x_{i_1} - y_{j_2}|^2 + |x_{i_2} - y_{j_1}|^2 - |x_{i_1} - y_{j_1}|^2 - |x_{i_2} - y_{j_2}|^2 \right] \\ &= -2\lambda (x_{i_2} - x_{i_1})(y_{j_1} - y_{j_2}), \end{aligned}$$

by Lemma 3.1 applied to $x_{i_1} < x_{i_2}$ and $y_{j_2} < y_{j_1}$. Both factors are strictly positive, so

$$\langle C, Q \rangle < \langle C, P \rangle,$$

contradicting the optimality of P . Therefore

$$P_{ij} > 0, \quad P_{kl} > 0, \quad i < k \quad \implies \quad j \leq l. \quad (22)$$

Concretely, (22) says the nonzeros form a *staircase*. Writing $*$ for a positive entry and leaving zeros blank,

$$P = \begin{pmatrix} * & * & & & \\ & * & * & & \\ & & & * & \\ & & & * & * \\ & & & & * \end{pmatrix} \quad \text{but never} \quad \begin{pmatrix} & * \\ * & \end{pmatrix}.$$

Each row's block starts no earlier than the previous row's block ended. Two consecutive rows may share a column, and a block may jump ahead, but nothing ever steps back: that last 2×2 pattern is precisely the crossing just ruled out.

(ii) **The only $P \in U(a, b)$ with monotone support is $\frac{1}{n}I$.** Let $P \in U(a, b)$ have monotone support, and suppose

$$P_{il} > 0 \quad \text{for some } l > i.$$

Rows after i avoid columns before l . Monotone support and $P_{il} > 0$ force every $P_{kl'} > 0$ with $k > i$ to have $l' \geq l$. Hence

$$P_{kl'} = 0 \quad \text{whenever } k > i \text{ and } l' < l.$$

Count the mass. Every row and every column carries $1/n$. The columns $1, \dots, l-1$ are therefore fed by the rows $1, \dots, i$ alone:

$$\frac{l-1}{n} = \sum_{l' < l} \sum_k P_{kl'} = \sum_{l' < l} \sum_{k \leq i} P_{kl'} \leq \sum_{k \leq i} \sum_j P_{kj} = \frac{i}{n},$$

so $l-1 \leq i$; combined with $l > i$ this leaves $l = i+1$.

Close that case. With $l = i+1$ the chain above is an equality throughout, so its one inequality is tight:

$$\sum_{k \leq i} \sum_{j > i} P_{kj} = 0, \quad \text{in particular} \quad P_{i,i+1} = 0,$$

contradicting $P_{il} > 0$. The same argument applied to $P^\top$ rules out $l < i$.

Hence $P_{ij} > 0$ only when $j = i$, and the row sums give $P = \frac{1}{n}I$.

Combining the two. An optimal plan exists by Proposition 2.3; by (i) it has monotone support; so by (ii) it equals $\frac{1}{n}I$. Since (ii) leaves no other candidate, $\frac{1}{n}I$ is the *unique* optimal plan, and (21) is its cost. $\square$

Remark 3.4 (The cheapest useful algorithm in the course). One-dimensional optimal transport between equal-size samples costs a sort: $\mathcal{O}(n \log n)$, no solver at all. This is the basis of sliced-Wasserstein methods [Bonnel et al., 2015], which reduce high-dimensional optimal transport to averaging this computation over random one-dimensional projections.

For general measures the same picture holds, with sorting replaced by quantile matching. Write $F_\mu(t) = \mu((-\infty, t])$ for the cumulative distribution function and $F_\mu^{-1}(q) = \inf\{t \in \mathbb{R} : F_\mu(t) \geq q\}$ for the *quantile function*; the two obey the equivalence $F_\mu^{-1}(q) \leq t \iff q \leq F_\mu(t)$. We state the result without proof.

Theorem 3.5 (One-dimensional optimal transport). *Let $\mu, \nu \in \mathcal{P}(\mathbb{R})$ and $c(x, y) = h(x - y)$ with h convex, and suppose the cost is finite for some coupling. Then the monotone coupling*

$$(X, Y) = (F_\mu^{-1}(U), F_\nu^{-1}(U)), \quad U \sim \mathcal{U}([0, 1]),$$

is optimal. If moreover μ is atomless, this coupling is the deterministic plan of the nondecreasing map $T = F_\nu^{-1} \circ F_\mu$, which satisfies $T_\# \mu = \nu$. For $c(x, y) = |x - y|^p$,

$$\mathcal{W}_p(\mu, \nu) = \left(\int_0^1 |F_\mu^{-1}(q) - F_\nu^{-1}(q)|^p dq \right)^{1/p}, \quad (23)$$

so $\mathcal{W}_p$ is the L^p distance between quantile functions.

4 The geometry of optimal plans beyond the line

Everything in §3 leaned on the ordering of $\mathbb{R}$. This section asks what is left once that ordering is gone, and answers it twice: first combinatorially, by replacing “no crossing” with a condition on finite cycles, and then — for the quadratic cost — by identifying the optimal map outright.

4.1 Cyclical monotonicity

In one dimension we found that an optimal plan never crosses itself (step (i) of Theorem 3.3), and we proved it by exhibiting a cheaper plan whenever two assignments crossed. But “crossing” presupposes an *order* on the ground space, and $\mathbb{R}^d$ has none for $d \geq 2$. What survives?

What survives is the *move*, not the order. The 2×2 exchange of §3.1 took two assignments and swapped their destinations; in $\mathbb{R}^d$ we may still do that, and more generally permute the destinations of any finite family of assignments. An optimal plan must contain no such improving reshuffle — otherwise we would perform it. Making that precise gives a purely combinatorial certificate of optimality, valid in any dimension and for any cost, of which one-dimensional monotonicity is the special case.

Definition 4.1 (*c-cyclical monotonicity*). A set $S \subseteq \mathbb{R}^d \times \mathbb{R}^d$ is *c-cyclically monotone* if for every $N \in \mathbb{N}$, every family $(x_1, y_1), \dots, (x_N, y_N) \in S$, and every permutation σ of $\{1, \dots, N\}$,

$$\sum_{i=1}^N c(x_i, y_i) \leq \sum_{i=1}^N c(x_i, y_{\sigma(i)}).$$

The assignment already in S beats every reshuffling of the same destinations among the same sources.

Theorem 4.2 (Fundamental theorem of optimal transport). *Let c be continuous and bounded below, and let $\pi \in \Pi(\mu, \nu)$. Then π is optimal for (2) if and only if $\text{supp } \pi$ is c-cyclically monotone.*

For the proof, see [Santambrogio, 2015] or [Ambrosio & Gigli, 2013].

Remark 4.3 (Cyclical monotonicity as a loss function). Theorem 4.2 is purely combinatorial: it mentions no marginals, no dual variables, and no dimension. Because the equivalence runs both ways, checking the condition on a candidate *certifies* its optimality — and, more usefully, the failure of the condition can be *penalised*. That is what the *Monge gap* of [Uscidda & Cuturi, 2023] does: given a candidate map T , it measures how far $\{(x, T(x))\}$ is from being *c-cyclically monotone*, yielding a regulariser that requires no convex potential and no dual variable, and that can be evaluated on samples alone.

4.2 Brenier’s theorem

Theorem 3.3 says that on the line, under the quadratic cost, the optimal plan is carried by a *nondecreasing map*: $x_i \mapsto y_i$. That is a statement about an order, and $\mathbb{R}^d$ carries none for $d \geq 2$. But the same map admits a second description, which does survive. A function $T: \mathbb{R} \rightarrow \mathbb{R}$ is nondecreasing exactly when it is the derivative of a convex function,

$$T = \varphi', \quad \varphi \text{ convex,}$$

since $\varphi(x) = \int_0^x T(s) ds$ is convex precisely when T is nondecreasing. Replace the derivative by a gradient and the statement has meaning in every dimension.

Theorem 4.4 (Brenier [Brenier, 1991]). *Let $c(x, y) = \frac{1}{2}|x - y|^2$ on $\mathbb{R}^d$, and let μ, ν have finite second moments with μ absolutely continuous. Then the optimal coupling is unique and is carried by a map*

$$T = \nabla \Phi, \quad \Phi: \mathbb{R}^d \rightarrow \mathbb{R} \text{ convex,}$$

and this T is also the unique solution of the Monge problem.

The link is Theorem 4.2. Take $c(x, y) = \frac{1}{2}|x - y|^2$ and expand the squares in Definition 4.1. The terms $|x_i|^2$ cancel, and $\sum_i |y_i|^2 = \sum_i |y_{\sigma(i)}|^2$ because σ merely permutes the y 's, so the condition collapses to

$$\sum_i \langle x_i, y_i \rangle \geq \sum_i \langle x_i, y_{\sigma(i)} \rangle \quad \text{for every cycle in the support.}$$

This is precisely *cyclical monotonicity* in the sense of convex analysis, and Rockafellar's theorem [Rockafellar, 1966] says such a set lies in the graph of the subdifferential of a convex function Φ . If μ admits a density that subdifferential collapses to a gradient — a convex function on $\mathbb{R}^d$ is differentiable off a Lebesgue-null set, hence μ -almost everywhere — so the optimal plan is carried by $\nabla \Phi$. For the details see [Santambrogio, 2015] or [Ambrosio & Gigli, 2013].

Remark 4.5 (What the structure buys). Brenier does two things. First, it makes “the Monge map” a well-defined object, so one may speak of the optimal map rather than of an optimal plan. Second, and this is why it matters to us, $\nabla(\text{convex})$ is a *parameterisation one can compute with*: an input-convex neural network [Amos et al., 2017] represents Φ by a network convex in its input, so $T = \nabla \Phi$ is an optimal map by construction, with no constraint left to enforce.

References

- [Ambrosio & Gigli, 2013] L. Ambrosio and N. Gigli. A user's guide to optimal transport. In *Modelling and Optimisation of Flows on Networks*, Lecture Notes in Mathematics 2062, pages 1–155. Springer, 2013.
- [Amos et al., 2017] B. Amos, L. Xu, and J. Z. Kolter. Input convex neural networks. In *International Conference on Machine Learning*, 2017.
- [Bonneel et al., 2015] N. Bonneel, J. Rabin, G. Peyré, and H. Pfister. Sliced and Radon Wasserstein barycenters of measures. *Journal of Mathematical Imaging and Vision*, 51(1):22–45, 2015.
- [Brenier, 1991] Y. Brenier. Polar factorization and monotone rearrangement of vector-valued functions. *Communications on Pure and Applied Mathematics*, 44(4):375–417, 1991.
- [Cuturi, 2013] M. Cuturi. Sinkhorn distances: lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, volume 26, 2013.

- [Franklin & Lorenz, 1989] J. Franklin and J. Lorenz. On the scaling of multidimensional matrices. *Linear Algebra and its Applications*, 114–115:717–735, 1989.
- [Peyré & Cuturi, 2019] G. Peyré and M. Cuturi. Computational optimal transport. *Foundations and Trends in Machine Learning*, 11(5–6):355–607, 2019. <https://arxiv.org/abs/1803.00567>.
- [Rockafellar, 1966] R. T. Rockafellar. Characterization of the subdifferentials of convex functions. *Pacific Journal of Mathematics*, 17(3):497–510, 1966.
- [Santambrogio, 2015] F. Santambrogio. *Optimal Transport for Applied Mathematicians: Calculus of Variations, PDEs, and Modeling*. Progress in Nonlinear Differential Equations and Their Applications 87, Birkhäuser, 2015.
- [Sinkhorn, 1964] R. Sinkhorn. A relationship between arbitrary positive matrices and doubly stochastic matrices. *The Annals of Mathematical Statistics*, 35(2):876–879, 1964.
- [Uscidda & Cuturi, 2023] T. Uscidda and M. Cuturi. The Monge gap: a regularizer to learn all transport maps. In *International Conference on Machine Learning*, 2023. <https://arxiv.org/abs/2302.04953>.
- [Villani, 2009] C. Villani. *Optimal Transport: Old and New*. Grundlehren der mathematischen Wissenschaften 338, Springer, 2009.