Topics in flow-based generative modeling and optimal transport

(work in progress)

Aram-Alexandre Pooladian*

September 3, 2026

Contents

*Institute for Foundations of Data Science, Yale University. aram-alexandre.pooladian@yale.edu

1 Preface and overview

2 Primer on optimal transport

2.1 Transport maps

Take two probability measures ρ and μ, which we will respectively call the *source* and *target* measures. Our first notion of transport is through *maps*.

Definition 2.1 (Transport maps). For $\rho, \mu \in \mathcal{P}(\mathbb{R}^d)$, the set of *transport maps* from ρ to μ is

$$\mathcal{T}(\rho, \mu) := \{T : \mathbb{R}^d \to \mathbb{R}^d \mid X \sim \rho, T(X) \sim \mu\}. \tag{2.1}$$

Equivalently, T is a *pushforward* from ρ to μ, written $T_\sharp \rho = \mu$.

To further motivate this definition, consider the following simple example:

Exercise 2.1. Let $\rho = \mathcal{N}(0, 1)$ and $\mu = \mathcal{N}(m, \sigma^2)$. Two possible transport maps from ρ to μ are

$$T(x) = m \pm \sigma x.$$

Indeed, if $X \sim \rho$, then one easily sees that $T(X) \sim \mu$ (the minus sign makes no difference as the standard Gaussian distribution is symmetric). Similarly, in $\mathbb{R}^d$, we have

$$T(x) = m \pm \Sigma^{1/2} x,$$

as possible maps from $\mathcal{N}(0, I_d)$ to $\mathcal{N}(m, \Sigma)$, where $S \mapsto S^{1/2}$ is the matrix square-root.

2.2 The Monge problem

The minus sign in Example 2.1 should raise a small red flag. While the sign flip is "allowed", in the sense that the resulting random variable is still distributed according to μ, the act of flipping the sign of the (input) random variable seems potentially superfluous.

Exercise 2.2. To see that $T_0(x) = m + \sigma x$ is "more optimal" than if we included the sign flip, compute the cost of displacement with respect to the (squared) Euclidean cost:

$$\mathbb{E}_{X \sim \rho}(X - T_0(X))^2 = m^2 + (1 - \sigma)^2.$$

A similar computation for the seemingly "suboptimal" mapping reads

$$\mathbb{E}_{X \sim \rho}(X - (m - \sigma X))^2 = m^2 + (1 + \sigma)^2.$$

Clearly, the cost of the first map is less than that of the second for all $\sigma > 0$. One can readily extend this argument to $\mathbb{R}^d$.

We now define a class of *optimal* transport maps: maps in $\mathcal{T}(\rho, \mu)$ which minimize the cost of displacement.

Definition 2.2 (Optimal transport map). An *optimal transport map* between $\rho, \mu \in \mathcal{P}_2(\mathbb{R}^d)$ is the solution to

$$T_0 \in \operatorname*{argmin}_{T \in \mathcal{T}(\rho,\mu)} \int \tfrac{1}{2}\|x - T(x)\|^2 \, \mathrm{d}\rho(x)\,, \tag{2.2}$$

where (2.2) is called the *Monge problem*.

2.3 Couplings

What are some caveats to working directly with transport maps?

First, the set $\mathcal{T}(\rho, \mu)$ is *non-convex*, and thus the Monge problem (2.2) is non-convex.

Second, the set of transport maps might not even be larger than the empty set. Indeed, suppose our source distribution was a Dirac mass at a single point

$$\rho = \delta_{x_1}$$

then *any* mapping $x_1 \mapsto T(x_1)$ outputs a single value. This causes problems for existence of a transport map as soon as the target is even a mixture of Diracs, say

$$\mu = \frac{1}{2}\delta_{y_1} + \frac{1}{2}\delta_{y_2}\,.$$

Indeed, a mapping can only send x_1 to one of $\{y_1, y_2\}$.

To remedy the situation, we will now work with *plans* or *couplings*, in place of maps. This gives us the flexibility of "splitting" the mass of a source particle to multiple points in the target space.

Definition 2.3 (Couplings). For $\rho, \mu \in \mathcal{P}_2(\mathbb{R}^d)$, we write $\Gamma(\rho, \mu)$ to be the set of *couplings* between ρ and μ

$$\Gamma(\rho, \mu) := \{\pi \in \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d) \,|\, \pi(A \times \mathbb{R}^d) = \rho(A), \pi(\mathbb{R}^d \times A) = \mu(A)\}\,. \tag{2.3}$$

Notice that the set of couplings always includes the independent coupling

$$\pi_{\text{ind}} = \rho \otimes \mu\,.$$

Of course, *if* there exists a deterministic function $T : \mathbb{R}^d \to \mathbb{R}^d$ such that $T_\sharp \rho = \mu$, then a candidate coupling is

$$\pi_T = (\mathrm{id}, T)_\sharp \rho\,.$$

Thus, the set of couplings is strictly larger than the set of transport maps. Even stronger, one should think of the set of couplings as a "convexification" of the set of transport maps; see the following lemma.

Lemma 2.4. *For two measures ρ, μ, the set of couplings $\Gamma(\rho, \mu)$ is a non-empty, convex, and compact set (with respect to the weak topology).*

Proof. In class. □

2.4 Kantorovich relaxation and dual formulation

2.4.1 The 2-Wasserstein distance and general costs

As with the Monge problem and transport maps, we now have infinitely many couplings to choose from. Taking the optimization route, we arrive at what is known as the (primal) Kantorovich formulation of optimal transport [Kan42], which we use to define the (squared) 2-Wasserstein distance.

Definition 2.5 (2-Wasserstein distance). For $\rho, \mu \in \mathcal{P}_2(\mathbb{R}^d)$, the *optimal transport distance* under the squared-Euclidean cost between ρ and μ is given by

$$\tfrac{1}{2}W_2^2(\rho, \mu) := \inf_{\pi \in \Gamma(\rho,\mu)} \iint \tfrac{1}{2}\|x - y\|^2 \, \mathrm{d}\pi(x, y) \tag{2.4}$$

where we recall that $\Gamma(\rho, \mu)$ is the set of *couplings* between ρ and μ. Throughout these notes, W_2 denotes the usual Wasserstein distance; the quadratic transport cost with $c(x, y) = \frac{1}{2}\|x - y\|^2$ is therefore $\frac{1}{2}W_2^2(\rho, \mu)$.

If we now observe that the functional

$$\pi \mapsto \int \|\cdot\|^2 \, \mathrm{d}\pi$$

is *lower semicontinuous*, and the set of couplings is non-empty, compact, and convex, we arrive at the following theorem that justifies the use of (2.4) as the definition of the 2-Wasserstein distance.

Theorem 2.6. *If $\rho, \mu \in \mathcal{P}_2(\mathbb{R}^d)$, then (2.4) admits at least one minimizer π_0.*

In many applications, the cost of interest will not be the squared Euclidean distance. Following the definition of 2-Wasserstein, we write

$$W_c(\rho, \mu) := \inf_{\pi \in \Gamma(\rho,\mu)} \iint c(x, y) \, \mathrm{d}\pi(x, y) \, ,$$

for suitable costs $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R} \cup \{+\infty\}$. An important class is the set of W_p distances for $p \geq 1$, written

$$W_p^p(\rho, \mu) := \inf_{\pi \in \Gamma(\rho,\mu)} \iint \|x - y\|_2^p \, \mathrm{d}\pi(x, y) \, .$$

(Important: the norm is still the Euclidean norm, just raised to the power p!)

2.4.2 The Wasserstein space is a metric space

Importantly, the 2-Wasserstein distance defines a metric on the space of probability measures with finite second moment $\mathcal{P}_2(\mathbb{R}^d)$.

Proposition 2.7. *The Wasserstein space, denoted* $\mathbb{W} := (\mathcal{P}_2(\mathbb{R}^d), W_2)$*, defines a metric space.*

Exercise 2.3. Use Hölder's inequality to prove that for $1 \leq p \leq q$ it holds that $W_p(\mu,\nu) \leq W_q(\mu,\nu)$.

2.4.3 Duality

Following our intuition from the finite-dimensional programming, we suspect that the Kantorovich problem admits a *dual* formulation.

Proposition 2.8 (Kantorovich dual)**.** *For* $\rho, \mu \in \mathcal{P}_2(\mathbb{R}^d)$*, the dual Kantorovich problem is the following constrained maximization problem*

$$\tfrac{1}{2}W_2^2(\rho,\mu) = \sup_{\substack{f\in L^1(\rho)\\ g\in L^1(\mu)}} \int f \,\mathrm{d}\rho + \int g \,\mathrm{d}\mu \quad \text{s.t.} \quad f(x)+g(y) \leq \tfrac{1}{2}\|x-y\|^2 \tag{2.5}$$

where the constraint holds for (x,y)*-almost everywhere.*

Proof. We start by rewriting the marginal constraints (the set of couplings) in terms of their *Lagrange multipliers*. Indeed the marginal constraint for ρ is equivalent to

$$\iint f(x)\,\mathrm{d}\pi(x,y) - \int f(x)\,\mathrm{d}\rho(x) = 0 \tag{2.6}$$

over all $f \in C_b(\mathbb{R}^d)$, the set of continuous bounded functions on $\mathbb{R}^d$, and analogously for μ. From these definitions, using (f,g) as the two Lagrange multipliers, we have that (2.4) is equivalent to

$$\begin{aligned}\inf_{\pi\geq 0} &\iint \tfrac{1}{2}\|x-y\|^2\,\mathrm{d}\pi(x,y)\\ &+ \sup_{\substack{f\in C_b(\mathbb{R}^d)\\ g\in C_b(\mathbb{R}^d)}} \Big(\int f(x)\,\mathrm{d}(\rho(x)-\pi(x,y)) + \int g(y)\,\mathrm{d}(\mu(y)-\pi(x,y))\Big)\,.\end{aligned}$$

We can move the supremum to the left of the first integral, and group together like-terms to obtain

$$\inf_{\pi\geq 0} \sup_{\substack{f\in C_b(\mathbb{R}^d)\\ g\in C_b(\mathbb{R}^d)}} \int f\,\mathrm{d}\rho + \int g\,\mathrm{d}\mu + \iint \left(\tfrac{1}{2}\|x-y\|^2 - f(x) - g(y)\right)\mathrm{d}\pi(x,y)\,.$$

Taking for granted the interchange of the supremum and infimum, moving the infimum inside yields

$$\sup_{\substack{f\in C_b(\mathbb{R}^d)\\ g\in C_b(\mathbb{R}^d)}} \int f \,\mathrm{d}\rho + \int g \,\mathrm{d}\mu + \inf_{\pi\geq 0} \iint \left(\tfrac{1}{2}\|x-y\|^2 - f(x) - g(y)\right) \mathrm{d}\pi(x,y)\,.$$

To conclude, it suffices to observe that if the integrand on the right is ever negative, one can take π to drive the value of the optimization problem to $-\infty$ (which we do not want). This motivates the desired constraint

$$f(x) + g(y) \leq \tfrac{1}{2}\|x-y\|^2\,,$$

which must hold for all (x,y). If this is true, then the integral on the right is lower-bounded by zero. Finally, since $C_b(\mathbb{R}^d)$ is dense in $L^1(\rho)$ and $L^1(\mu)$, we can extend the supremum to these larger spaces as long as we keep the pointwise constraint. □

We conclude with a definition for the optimal maximizers of (2.5).

Definition 2.9 (Optimal Kantorovich potentials). We write the maximizers to (2.5) as (f_0, g_0), which are called *(optimal) Kantorovich potentials*.

2.4.4 Semi-dual

Looking again at the dual Kantorovich problem, we can make the following adjustments to the objective function: let $m_2^2(\rho) \coloneqq \int \|\cdot\|^2 \,\mathrm{d}\rho$ and similarly for μ, then (2.5) is

$$\tfrac{1}{2}m_2^2(\rho) + \tfrac{1}{2}m_2^2(\mu) + \int (f - \tfrac{1}{2}\|\cdot\|^2)\,\mathrm{d}\rho + \int (g - \tfrac{1}{2}\|\cdot\|^2)\,\mathrm{d}\mu\,,$$

with the constraint written as

$$(\tfrac{1}{2}\|\cdot\|^2 - f)(x) + (\tfrac{1}{2}\|\cdot\|^2 - g)(y) \geq \langle x, y\rangle\,.$$

Making the substitutions

$$(\varphi, \psi) \coloneqq (\tfrac{1}{2}\|\cdot\|^2 - f, \tfrac{1}{2}\|\cdot\|^2 - g)\,,$$

then (2.5) is equivalently written as (up to additive constants)

$$\inf_{\varphi,\psi} \int \varphi \,\mathrm{d}\rho + \int \psi \,\mathrm{d}\mu \quad \text{s.t.} \quad \varphi(x) + \psi(y) \geq \langle x, y\rangle\,. \tag{2.7}$$

Let's fix φ and consider this only as an optimization problem in ψ. This is a minimization problem, and the constraint now reads, for all $y \in \mathbb{R}^d$

$$\psi(y) \geq \langle x, y\rangle - \varphi(x)\,.$$

In particular, we can take the supremum over all x on the right-hand side, which is precisely the definition of the *convex conjugate* of φ.[1] We can swap this into the objective now, since it only makes the objective function smaller—but the objective is now free of ψ entirely! Repeating this procedure leads to yet another formulation of the Kantorovich problem.

Definition 2.10 (Kantorovich semidual). The *semidual* formulation of the Kantorovich problem is

$$\tfrac{1}{2}W_2^2(\rho,\mu) = \tfrac{1}{2}m_2^2(\rho) + \tfrac{1}{2}m_2^2(\mu) - \inf_{\varphi} \mathcal{S}^{\rho\mu}(\varphi)\,, \tag{2.8}$$

where we write

$$\mathcal{S}^{\rho\mu}(\varphi) \coloneqq \int \varphi \,\mathrm{d}\rho + \int \varphi^* \,\mathrm{d}\mu\,, \tag{2.9}$$

and the optimization is taken over proper, lower semicontinuous, convex functions φ.

2.5 Brenier's Theorem and generalizations

The semi-dual formulation (2.9) leads into one of the most influential theorems in optimal transport, due to Brenier [Bre91]. It connects the primal Kantorovich problem, the semidual problem, and the Monge problem, all in one statement.

Theorem 2.11 (Brenier's theorem). *For $\rho, \mu \in \mathcal{P}_2(\mathbb{R}^d)$ with ρ absolutely continuous, the (ρ-a.e. unique) optimal transport map T_0 between ρ and μ exists and is given by the gradient of a proper, lower semicontinuous, convex function, namely*

$$T_0 \coloneqq \nabla \varphi_0 \quad (\rho\text{-a.e.})$$

where φ_0 is the minimizer to (2.8), called a Brenier potential. Moreover, the optimal plan π_0 is given by

$$\pi_0 = (\mathrm{id}, T_0)_\sharp \rho\,. \tag{2.10}$$

The key assumption in Brenier's theorem is that the source distribution must have a density, which rules out the pathological Dirac mass example from earlier. In brief, once there are "enough" points that can move from ρ to μ, we have enough freedom to uniquely and optimally transport mass.

Proof. Let π_0 be the optimal transport plan between ρ and μ, and let φ_0 be the optimal

[1] For a function $f : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$, the convex conjugate is written $f^*(y) \coloneqq \sup_{x \in \mathrm{dom}(f)} \{\langle x, y\rangle - f(x)\}$.

Brenier potential. Rearranging the semi-dual formulation, one arrives at

$$\int \varphi_0(x)\,\mathrm{d}\rho(x) + \int \varphi_0^*(y)\,\mathrm{d}\mu(y) = \int \langle x, y\rangle\,\mathrm{d}\pi_0(x,y)\,.$$

Since $\pi_0 \in \Gamma(\rho,\mu)$, we can further write the above as

$$\int (\varphi_0(x) + \varphi_0^*(y) - \langle x, y\rangle)\,\mathrm{d}\pi_0(x,y) = 0\,. \tag{2.11}$$

Now, suppose $(x_0, y_0) \in \mathrm{supp}(\pi_0)$; (2.11) implies that

$$\varphi_0^*(y_0) = \langle x_0, y_0\rangle - \varphi_0(x_0) \equiv \sup_x \langle x, y_0\rangle - \varphi_0(x)\,.$$

Since φ_0 is differentiable ρ-a.e., the optimality relation tells us that at x_0,

$$y_0 = \nabla\varphi_0(x_0)\,.$$

Thus, the pair $(x_0, \nabla\varphi_0(x_0)) \in \mathrm{supp}(\pi_0)$ is deterministic, which completes the claim. □

2.5.1 Gaussian to Gaussian case

In the case of optimal transport between Gaussians, a closed-form expression is available for the optimal transport cost and transport map.

Theorem 2.12 (Optimal transport between Gaussians)**.** *Let $\rho = \mathcal{N}(m_\rho, \Sigma_\rho)$ and $\mu = \mathcal{N}(m_\mu, \Sigma_\mu)$ with $\Sigma_\rho \succ 0$. Then*

$$\tfrac{1}{2}W_2^2(\rho,\mu) = \tfrac{1}{2}\|m_\mu - m_\rho\|^2 + \tfrac{1}{2}\,\mathrm{tr}\big(\Sigma_\rho + \Sigma_\mu - 2(\Sigma_\rho^{1/2}\Sigma_\mu\Sigma_\rho^{1/2})^{1/2}\big)\,, \tag{2.12}$$

and the optimal transport map is given by

$$T_0(x) = m_\mu + \Sigma_\rho^{-1/2}(\Sigma_\rho^{1/2}\Sigma_\mu\Sigma_\rho^{1/2})^{1/2}\Sigma_\rho^{-1/2}(x - m_\rho)\,.$$

Proof. We first make the ansatz that the optimal transport map will be affine, i.e., of the form $T_0(x) = m_\mu + A(x - m_\rho)$ for some symmetric positive definite matrix A (otherwise Brenier's theorem will be violated). Then the pushforward condition is equivalent to

$$A\Sigma_\rho A^\top = A\Sigma_\rho A = \Sigma_\mu\,,$$

Writing $B = \Sigma_\rho^{1/2} A \Sigma_\rho^{1/2}$, it is easy to verify that we obtain

$$A = \Sigma_\rho^{-1/2}(\Sigma_\rho^{1/2}\Sigma_\mu\Sigma_\rho^{1/2})^{1/2}\Sigma_\rho^{-1/2}\,.$$

For the rest of the proof, we need to show that no other transport map results in a lower cost, which is left as an exercise. □

2.5.2 Optimal transport and Brenier's theorem for general costs

Let $c : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_+$ be a lower semicontinuous function; for simplicity take $c(x,y) = h(x-y)$ for any strictly convex function h. Recall the optimal transport problem in this case is

$$W_c(\rho,\mu) = \inf_{\pi\in\Gamma(\rho,\mu)} \iint c(x,y)\,\mathrm{d}\pi(x,y)\,. \tag{2.13}$$

We also have a dual representation

$$W_c(\rho,\mu) = \sup_{\substack{f\in L^1(\rho)\\ g\in L^1(\mu)}} \mathcal{D}^{\rho\mu}(f,g) \tag{2.14}$$

with

$$\mathcal{D}^{\rho\mu}(f,g) := \int f\,\mathrm{d}\rho + \int g\,\mathrm{d}\mu \quad \text{s.t.} \quad f(x)+g(y)\le c(x,y)\,, \tag{2.15}$$

where again the inequality holds for $\rho\otimes\mu$-a.e.

Pattern-matching with the squared-Euclidean cost (which can be neatly rewritten in terms of the inner product cost), we can equivalently define the *c-transform* of a function as

$$f^c(y) := \inf_x \Big\{ c(x,y) - f(x) \Big\}\,.$$

Note that replacing this in the dual objective only makes the supremum bigger. Repeating this gives

$$f^{cc}(x) := (f^c)^c(x) = \inf_y \Big\{ c(x,y) - f^c(y) \Big\}\,.$$

Then, as in the semi-dual case, we can rewrite the dual problem as

$$W_c(\rho,\mu) = \sup_{f\in L^1(\rho)} \int f\,\mathrm{d}\rho + \int f^c\,\mathrm{d}\mu\,, \tag{2.16}$$

where the constraint is now embedded in the definition of the problem.

Exercise 2.4. Show that the total variation distance can be expressed as a Wasserstein distance for a particular cost.

[GM96] extended Brenier's result to the optimal transport problem for these general costs. We state a version of their result below.

Theorem 2.13 (Gangbo-McCann theorem)**.** *Under suitable conditions on c, for $\rho \in \mathcal{P}_{ac}(\mathbb{R}^d)$ and $\mu \in \mathcal{P}(\mathbb{R}^d)$, the ($\rho$-a.e. unique) optimal transport map T_c between ρ and μ*

exists and is given by

$$T_c(x) := \nabla_x c(x, \cdot)^{-1} \circ \nabla f_c(x) \quad (\rho\text{-a.e.}) \tag{2.17}$$

where f_c is the maximizer to (2.16).

For *translation-invariant* convex costs as we mentioned above, i.e., $c(x, y) = h(x - y)$ with h strictly convex, we obtain the following corollary.

Corollary 2.14 (Translation-invariant convex costs)**.** *Under the assumptions of Theorem 2.13, for $c(x, y) = h(x - y)$ with h strictly convex, the optimal transport map is given by*

$$T_h(x) = x - (\nabla h^*) \circ \nabla f_h(x)\,, \tag{2.18}$$

where h^* is the convex conjugate of h, and f_h is the maximizer to (2.16) under the cost function above.

2.5.3 Paths defined by optimal transport maps

Optimal transport maps define a convenient mapping which sends one probability distribution to another. It is natural to ask how such mappings might give rise to a notion of paths between probability distributions.

For example, taking $x, y \in \mathbb{R}^d$ with the usual (squared)-Euclidean distance, the shortest path from x to y is, for any $t \in [0, 1]$

$$x_t := (1 - t)x + ty\,.$$

If we now replace two points in $\mathbb{R}^d$ with two probability measures with density, we can also ask to define a notion of path or interpolation between them. For $\mu, \nu \in \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$, one such interpolation is

$$\tilde{\mu}_t = t\mu + (1 - t)\nu\,. \tag{2.19}$$

This formulation, however, does not take advantage of optimal transport maps. A more natural interpolation is defined at the level of the *random variables*, where for $X_0 \sim \mu$ and optimal transport map $T^{\mu\nu}$

$$X_t = (1 - t)X_0 + tT^{\mu\nu}(X_0)\,,$$

where we define $\mu_t := \mathrm{Law}(X_t)$. At the level of the measures, this displacement becomes

$$\mu_t = (T_t^{\mu\nu})_\sharp \mu := ((1 - t)\mathrm{id} + tT^{\mu\nu})_\sharp \mu\,. \tag{2.20}$$

For example, if μ, ν are both Gaussian measures, then one can check that (2.20) and (2.19) are not the same: μ_t is always a Gaussian for any $t \in [0, 1]$ whereas $\tilde{\mu}_t$ is a mixture of Gaussians.

2.6 The univariate case

We now consider the special case where the source and target measures are *univariate*. Similar to the Gaussian case, we also have "closed-form" expressions for maps of a certain kind, as well as a delicate isometry property as a result.

2.6.1 Closed-form expression of maps

Recall that for a probability measure ρ, its *cumulative distribution function* (CDF) is

$$F_\rho(t) = \int_{-\infty}^{t} \mathrm{d}\rho(x) = \mathbb{P}_\rho(X \leq t)\,, \tag{2.21}$$

where $F_\rho : \mathbb{R} \to [0,1]$. By design, F_ρ is a nondecreasing function. The (pseudo)inverse of a CDF, called the *(generalized) quantile*, is

$$F_\rho^{-1}(u) := \inf\{x \in \mathbb{R} : F_\rho(x) \geq u\}\,, \tag{2.22}$$

where $F_\rho^{-1} : [0,1] \to \mathbb{R} \cup \{+\infty\}$. Note that when ρ has a nice density, the generalized quantile is simply the quantile function of ρ. Moreover, F_ρ^{-1} is also a nondecreasing function.

From here, we can prove the following result, which gives a general recipe for obtaining closed-form expressions of optimal transport maps between univariate probability measures.

Theorem 2.15 (Optimal transport maps in 1D)**.** *Let $\rho, \mu \in \mathcal{P}(\mathbb{R})$ with densities. Then the Brenier map from ρ to μ is given by*

$$T^{\rho\to\mu} = F_\mu^{-1} \circ F_\rho\,. \tag{2.23}$$

Proof. First, one should check that $F_\mu^{-1} \circ F_\rho$ is a valid pushforward map from ρ to μ. This follows from two facts:

- for $X \sim \rho$, $F_\rho(X) \sim \mathrm{Unif}([0,1])$,
- for $U \sim \mathrm{Unif}([0,1])$, it holds that $F_\mu^{-1}(U) \sim \mu$.

To be a Brenier map, it remains to check that $F_\mu^{-1} \circ F_\rho$ is the gradient of a convex function, as such maps are unique. Over $\mathbb{R}$, this is simply asking that the function is itself nondecreasing, which is easy to verify. □

Exercise 2.5. Fill in the remaining details of Theorem 2.15.

2.6.2 Isometry property

From Theorem 2.15, we know that the cost of transport is

$$\tfrac{1}{2}W_2^2(\rho,\mu) = \tfrac{1}{2}\|\mathrm{id} - F_\mu^{-1}\circ F_\rho\|_{L^2(\rho)}^2 = \tfrac{1}{2}\|F_\rho^{-1} - F_\mu^{-1}\|_{L^2([0,1])}^2\,,$$

where the second equality is nothing but a change-of-variables. This is just to say that, in 1D, the squared 2-Wasserstein distance is isometric to $L^2([0,1])$. Said differently, $\rho \mapsto F_\rho^{-1}$ embeds $(\mathcal{P}_2(\mathbb{R}), W_2)$ isometrically into $(L^2([0,1]), \|\cdot\|_{L^2})$. As we'll see shortly, this property is deeply rooted in the univariate nature of the probability measures in question.

2.6.3 Sliced Wasserstein distances

We now describe the sliced Wasserstein distance, an offshoot of univariate optimal transport. For $\theta \in \mathbb{S}^{d-1}$, let $\pi_\theta : \mathbb{R}^d \to \mathbb{R}$ denote the projection $\pi_\theta(x) := \langle x, \theta\rangle$, and set $\rho_\theta := (\pi_\theta)_\sharp\rho \in \mathcal{P}_2(\mathbb{R})$. The definition is given below

Definition 2.16 (Sliced Wasserstein distance). For $\rho, \mu \in \mathcal{P}_2(\mathbb{R}^d)$, the *sliced Wasserstein distance* is

$$\tfrac{1}{2}\,\mathrm{SW}_2^2(\rho,\mu) := \int_{\mathbb{S}^{d-1}} \tfrac{1}{2}W_2^2(\rho_\theta,\mu_\theta)\;\mathrm{d}\sigma_{d-1}(\theta)\,, \tag{2.24}$$

where σ_{d-1} is the uniform probability measure on $\mathbb{S}^{d-1}$. Equivalently, by the 1D isometry (Section 2.6.2),

$$\tfrac{1}{2}\,\mathrm{SW}_2^2(\rho,\mu) = \int_{\mathbb{S}^{d-1}} \tfrac{1}{2}\|F_{\rho_\theta}^{-1} - F_{\mu_\theta}^{-1}\|_{L^2([0,1])}^2\;\mathrm{d}\sigma_{d-1}(\theta)\,.$$

Proposition 2.17. *SW_2 is a metric on $\mathcal{P}_2(\mathbb{R}^d)$ that induces the same topology as W_2.*

2.7 On the stability of optimal transport maps

Note that a trivial coupling argument always implies that

$$\tfrac{1}{2}W_2^2(\rho,\mu) \le \tfrac{1}{2}\|T^{\gamma\to\rho} - T^{\gamma\to\mu}\|_{L^2(\gamma)}^2\,,$$

for any probability measures γ with density. (Here the transport maps need not be optimal.) The right-hand side is known as a *linearization* of optimal transport (with respect to γ). A question of theoretical, statistical, and computational relevance is therefore the *stability* of (optimal) transport maps: for what $\alpha \in (0,1]$ does the following hold

$$\|T^{\gamma\to\rho} - T^{\gamma\to\mu}\|_{L^2(\gamma)} \le C\,W_2^\alpha(\rho,\mu)\,, \tag{2.25}$$

where $C > 0$ is some constant that ought to depend on γ, and mildly on properties of ρ and μ (e.g., the size of their support).

In the univariate case, (2.25) is an equality with $C = \alpha = 1$.

2.7.1 A first instability result

Our goal here is to provide a setting where the $\frac{1}{2}$-Hölder exponent is tight for the stability of optimal transport maps. The original example is presented in [Gig11], though we present the proof from [MDC20].

Let $\rho = \mathrm{Unif}(B(0,1))$ in $\mathbb{R}^2$, and write $\mu_\theta = \frac{1}{2}(\delta_{e_\theta} + \delta_{-e_\theta})$, where $e_\theta := (\cos(\theta), \sin(\theta))$. For $\theta = 0$, this returns two atoms along the horizontal axis of the ball; we call this discrete measure μ_0. For $0 < \theta \ll \pi/2$, μ_θ is merely a rotation of μ_0 along the perimeter of the ball. For all $\theta \in [0, \pi/2]$, the optimal transport map is given by

$$T_\theta(x) := T^{\rho \to \mu_\theta}(x) := e_\theta \operatorname{sign}(\langle x, e_\theta \rangle) \tag{2.26}$$

The first step is to verify that

$$\tfrac{1}{2} W_2^2(\mu_\theta, \mu_0) \leq \tfrac{1}{2}\theta^2 \,.$$

This follows from the fact that a transport map from μ_0 to μ_θ is given by the standard rotation matrix, and then a Taylor expansion. Next, lower-bound the quantity in question:

$$\|T_\theta - T_0\|_{L^2(\rho)}^2 \geq \int_{D_\theta} \|e_\theta + e_0\|^2 \, \mathrm{d}\rho(x) \geq 2\rho(D_\theta) = \frac{\theta}{\pi} \,, \tag{2.27}$$

where $D_\theta := \{x \in B(0,1) \mid x^\top e_\theta < 0 \,,\; x^\top e_0 > 0\}$. So it follows that

$$\|T_\theta - T_0\|_{L^2(\rho)}^2 \gtrsim \theta \gtrsim W_2(\mu_\theta, \mu_0) \,,$$

which completes the claim.

2.7.2 Some positive stability results

We state the following results without proof. The first two results impose increasing regularity assumptions on one optimal transport map, resulting in improved stability.

Theorem 2.18. *[Gig11] Let γ have a density and let $\rho, \mu \in \mathcal{P}_2(\mathbb{R}^d)$ both lie in a compact set of radius $r > 0$. If $T^{\gamma \to \rho}$ is L-Lipschitz, then*

$$\|T^{\gamma \to \rho} - T^{\gamma \to \mu}\|_{L^2(\gamma)} \leq \sqrt{2rL}\, W_1^{1/2}(\rho, \mu) \,.$$

Theorem 2.19. *[MBNWW24] Let γ, ρ be such that $T^{\gamma \to \rho}$ is bi-Lipschitz. Then for any $\mu \in \mathcal{P}_2(\mathbb{R}^d)$*

$$\|T^{\gamma \to \rho} - T^{\gamma \to \mu}\|_{L^2(\gamma)} \leq \sqrt{L/\ell}\, W_2(\rho, \mu) \,.$$

The next theorem is the latest state-of-the-art result pertaining to the stability of optimal transport maps with minimal regularity assumptions.

Theorem 2.20. *[Mér26] Suppose $\gamma \in \mathcal{P}_{\mathrm{ac}}(\mathbb{R}^d)$ over a convex compact set, with density bounded above. Let ρ, μ be supported on a convex compact subset of $\mathbb{R}^d$. Then*

$$\|T^{\gamma\to\rho} - T^{\gamma\to\mu}\|_{L^2(\gamma)} \lesssim W_2^{1/4}(\rho,\mu)\,,$$

where the constant depends on the dimension, and the radius of the supports of γ, ρ, μ.

2.8 Statistical optimal transport

In many applications, the source and target measures are not directly available to the practitioner. Instead, our goal will be to *estimate* various quantities on the basis of samples: given $X_1, \dots, X_n \sim \rho$ and $Y_1, \dots, Y_m \sim \mu$, how large does n have to be in order to meaningfully estimate $\frac{1}{2}W_2^2(\rho,\mu)$? What about estimating optimal transport maps?

2.8.1 Preliminaries

The proofs in this chapter will rely on basic applications of empirical process theory and chaining. Specifically, we will be interested in characterizing the behaviors of quantities such as

$$\mathbb{E}\sup_{f\in\mathcal{F}} |X_f| := \mathbb{E}\sup_{f\in\mathcal{F}} \left| \int f \,\mathrm{d}(P - P_n) \right|,$$

where P_n is an empirical measure of P (written $P_n = n^{-1}\sum_{i=1}^n \delta_{X_i}$ for $X_1, \dots, X_n \sim P$), and $\mathcal{F}$ is a suitable function class. The term X_f is known as an empirical process, and the class $\mathcal{F}$ can be as simple or complicated as the problem requires.

The simplest example would be if $\mathcal{F}$ consists of a single function, in which case the estimation rate would be parametric. For a larger function class, we recall the basic definition of the (log-)covering number of a set, which describes the complexity of the function class by the number of functions required to cover the class:[2] for $\tau > 0$

$$N(\tau, \mathcal{F}) := \inf\{|\mathcal{F}_N| : \mathcal{F}_N \text{ is a } \tau\text{-covering of } \mathcal{F}\}\,.$$

We will then control the empirical process using the following version of Dudley's (improved) chaining technique.

Proposition 2.21. *Let $\mathcal{F}$ be a class of real-valued functions over $\Omega \subseteq \mathbb{R}^d$, and suppose there exists $D > 0$ such that $\|f\|_{L^\infty(\Omega)} \le D$ for all $f \in \mathcal{F}$. Then*

$$\mathbb{E}\sup_{f\in\mathcal{F}} \left| \int f \,\mathrm{d}(\mu - \mu_n) \right| \lesssim \inf_{0<\delta\le D} \left\{ \delta + n^{-1/2} \int_\delta^D \sqrt{\log\big(2N(\tau,\mathcal{F})\big)}\,\mathrm{d}\tau \right\}.$$

[2] We call $F = \{f_1, \dots, f_N\}$ a τ-covering of a function class $\mathcal{F}$ defined over Ω if for all $f \in \mathcal{F}$, there exists an $i \in [N]$ such that $\|f_i - f\|_{L^\infty(\Omega)} \le \tau$.

2.8.2 Statistical estimation of Wasserstein distances

Recall the setup: given $X_1, \dots, X_n \sim \rho$ and $Y_1, \dots, Y_n \sim \mu$, how large does n have to be in order to meaningfully estimate $\frac{1}{2}W_2^2(\rho, \mu)$? Recall the notion of *empirical measures*

$$\rho_n = \tfrac{1}{n}\sum_{i=1}^n \delta_{X_i} \quad \text{and} \quad \mu_n := \tfrac{1}{n}\sum_{i=1}^n \delta_{Y_i}\,.$$

We will prove the following theorem.

Theorem 2.22. *Take $\rho, \mu \in \mathcal{P}_2(\mathbb{R}^d)$ with compact support in $B(0, R)$. Then*

$$\mathbb{E}|\tfrac{1}{2}W_2^2(\rho_n, \mu_n) - \tfrac{1}{2}W_2^2(\rho, \mu)| \lesssim_d \begin{cases} n^{-1/2} & d < 4\,, \\ n^{-1/2}\log(n) & d = 4\,, \\ n^{-2/d} & d > 4\,, \end{cases} \tag{2.28}$$

where we suppress constants in the dimension d.

For simplicity, let us assume full access to ρ and access to n i.i.d. samples from μ, with both measures supported in $B(0, R)$. We begin by recalling the dual formulation:

$$\tfrac{1}{2}W_2^2(\rho, \mu_n) = \tfrac{1}{2}m_2^2(\rho) + \tfrac{1}{2}m_2^2(\mu_n) - \inf_{\varphi \in \Phi} \mathcal{S}^{\rho\mu_n}(\varphi)\,,$$

and similarly for $\frac{1}{2}W_2^2(\rho, \mu)$, where Φ is the space of nice convex functions. We need some notation: for any P and Q, define

$$\varphi^{PQ} := \operatorname*{argmin}_{\varphi} \mathcal{S}^{PQ}(\varphi)\,.$$

Then it holds that $\mathcal{S}^{\rho\mu_n}(\varphi^{\rho\mu_n}) \le \mathcal{S}^{\rho\mu_n}(\varphi^{\rho\mu})$ for instance.

Then we can upper bound the estimation gap as follows

$$\mathbb{E}|\tfrac{1}{2}W_2^2(\rho, \mu_n) - \tfrac{1}{2}W_2^2(\rho, \mu)| \le \tfrac{1}{2}\mathbb{E}\Big|\int \|\cdot\|^2 \,\mathrm{d}(\mu - \mu_n)\Big| + \mathbb{E}\big|\mathcal{S}^{\rho\mu}(\varphi^{\rho\mu}) - \mathcal{S}^{\rho\mu_n}(\varphi^{\rho\mu_n})\big|\,.$$

It is easy to believe that the first term on the right-hand side scales favorably, with

$$\mathbb{E}\Big|\int \|\cdot\|^2 \,\mathrm{d}(\mu - \mu_n)\Big| \lesssim n^{-1/2}\,.$$

Whereas the first term can be controlled through elementary means, the second term is more involved. We require the following lemma, which relates the difference in semidual values by a suitable empirical process.

Lemma 2.23. *Suppose $\rho, \mu \in \mathcal{P}(B(0,R))$. Then*

$$\mathbb{E}\left|\mathcal{S}^{\rho\mu}(\varphi^{\rho\mu}) - \mathcal{S}^{\rho\mu_n}(\varphi^{\rho\mu_n})\right| \le \mathbb{E}\sup_{\varphi\in\Phi_R}\left|\int \varphi \,\mathrm{d}(\mu-\mu_n)\right|,$$

where Φ_R is the set of convex, R-Lipschitz functions on $B(0,R)$ with $\varphi(0)=0$.

Proof. Repeatedly playing with the optimality conditions to arrive at the following inequalities

$$\mathcal{S}_{\rho,\mu}(\varphi_{\rho,\mu}) - \mathcal{S}_{\rho,\mu_n}(\varphi_{\rho,\mu}) \le \mathcal{S}_{\rho,\mu}(\varphi_{\rho,\mu}) - \mathcal{S}_{\rho,\mu_n}(\varphi_{\rho,\mu_n}) \le \mathcal{S}_{\rho,\mu}(\varphi_{\rho,\mu_n}) - \mathcal{S}_{\rho,\mu_n}(\varphi_{\rho,\mu_n}),$$

which results in the following upper bound

$$\begin{aligned}|\mathcal{S}_{\rho,\mu}(\varphi_{\rho,\mu}) - \mathcal{S}_{\rho,\mu_n}(\varphi_{\rho,\mu_n})| &\le \max\{|\mathcal{S}_{\rho,\mu}(\varphi_{\rho,\mu}) - \mathcal{S}_{\rho,\mu_n}(\varphi_{\rho,\mu})|, |\mathcal{S}_{\rho,\mu}(\varphi_{\rho,\mu_n}) - \mathcal{S}_{\rho,\mu_n}(\varphi_{\rho,\mu_n})|\}\\ &\le \sup_{\varphi\in\Phi}\left|\int \varphi \mathrm{d}(\mu-\mu_n)\right|.\end{aligned}$$

Taking expectations, we have found our empirical process over "nice" convex functions. Let's try to identify more properties about this class: take $\varphi \in \Phi$ and $x \in B(0;R)$, then

$$\varphi(x) = \sup_y\left\{x^\top y - \varphi^*(y)\right\} = x^\top y_0 - \varphi^*(y_0),$$

where y_0 is the optimal choice. However, note that this choice y_0 is a valid *lower bound* for any input $x' \in B(0;R)$. Now, take $x, x' \in B(0;R)$ arbitrary,

$$\begin{aligned}\varphi(x') - \varphi(x) &\ge \left((x')^\top y_0 - \varphi^*(y_0)\right) - \left(x^\top y_0 - \varphi^*(y_0)\right)\\ &= (x'-x)^\top y_0\\ &\ge -R\|x'-x\|,\end{aligned}$$

where we used the fact that $y_0 \in B(0;R)$ in the last inequality (along with Cauchy-Schwarz). Repeating this argument with the signs switched (do this!), we see that $\varphi \in \Phi$ is R-Lipschitz. Thus, the supremum is taken with respect to the class of convex and R-Lipschitz functions, which we now call Φ_R. Moreover, without loss of generality, we can assume $\varphi(0)=0$, leading to $\|\varphi\|_\infty \le R^2$ for any $\varphi \in \Phi_R$ (why?). □

Our bound is now

$$\mathbb{E}|\tfrac{1}{2}W_2^2(\rho,\mu_n) - \tfrac{1}{2}W_2^2(\rho,\mu)| \le \mathbb{E}\sup_{\varphi\in\Phi_R}\left|\int \varphi \,\mathrm{d}(\mu-\mu_n)\right| + \mathcal{O}(n^{-1/2}).$$

We proceed by employing the following covering number bound from the statistics literature.

Theorem 2.24 ([GS12]). *Let Φ_R denote the set of convex, R-Lipschitz functions on $B(0;R)$ with $\varphi(0)=0$. There exist constants $C_1, C_2 > 0$ that depend on the dimension d such that for $\tau/R^2 < C_1$*

$$\log N(\tau, \Phi_R, \|\cdot\|_\infty) \le C_2(\tau/R^2)^{-d/2}.$$

Proof of Theorem 2.22. We now compute the standard Dudley's chaining bound using our covering number bounds on Φ_R.

$$\begin{aligned}
\mathbb{E}\sup_{\varphi\in\Phi_R}\left|\int \varphi \,\mathrm{d}(\mu-\mu_n)\right| &\lesssim \inf_{0<\delta\le R^2}\left\{\delta + \frac{1}{\sqrt{n}}\int_\delta^{R^2}\sqrt{\log\left(2N(\tau,\Phi_R,\|\cdot\|_\infty)\right)}\,\mathrm{d}\tau\right\} \\
&= R^2\inf_{0<\delta\le 1}\left\{\delta + \frac{1}{\sqrt{n}}\int_\delta^{1}\sqrt{\log\left(2N(R^2\tau,\Phi_R,\|\cdot\|_\infty)\right)}\,\mathrm{d}\tau\right\} \\
&\lesssim R^2\inf_{0<\delta\le 1}\left\{\delta + \frac{1}{\sqrt{n}}\int_\delta^{1}\tau^{-d/4}\,\mathrm{d}\tau\right\} \\
&\asymp \begin{cases} R^2 n^{-1/2} & d<4, \\ R^2 n^{-1/2}\log(n) & d=4, \\ R^2 n^{-2/d} & d>4, \end{cases}
\end{aligned}$$

where, in the last step, we optimize with respect to δ to obtain the tightest possible bound on the integral. □

2.8.3 One-sample map estimation

Suppose we have full access to the source measure ρ and only n i.i.d. samples $Y_1,\dots,Y_n \sim \mu$, and our task is to *generate* new samples that are distributed according to μ. By Brenier's theorem, we know that if ρ has a density (e.g. the standard normal distribution), then a unique optimal transport map T_0 exists between ρ and μ. Thus, if we were able to get an estimator $\widehat{T}_0$ of the optimal transport map, we can simply sample points from ρ and push them according to this learned map in order to generate new points.

Under suitable assumptions, [HR21] proved that there exists an estimator that achieves the following estimation error

$$\mathbb{E}\|\widehat{T} - T_0\|^2_{L^2(\rho)} \lesssim n^{-\frac{2\alpha}{2\alpha-2+d}}\log^3(n), \tag{2.29}$$

where $T_0 \in \mathcal{C}^\alpha$, and moreover that this rate is minimax optimal up to logarithmic factors. Since then, multiple works have expanded upon theirs, resulting in simpler estimators, simpler proofs, and less stringent assumptions.

We present a much simpler result that can be easily proved using the stability bounds stated above.

Theorem 2.25. *Suppose ρ and μ are probability measures with densities supported on $B(0,R)$, the optimal transport map T_0 is L-Lipschitz, and let μ_n denote the empirical measure of μ from n i.i.d. samples. Then there exists an estimator $\widehat{T}_n$ such that*

$$\mathbb{E}\|\widehat{T}_n - T_0\|^2_{L^2(\rho)} \lesssim \begin{cases} n^{-1/4} & d < 4\,, \\ n^{-1/4}\sqrt{\log(n)} & d = 4\,, \\ n^{-1/d} & d > 4\,, \end{cases} \tag{2.30}$$

where the suppressed constants depend on L, d, and R.

Proof. Set $\widehat{T}_n := T^{\rho\to\mu_n}$, the optimal transport map from ρ to the empirical measure μ_n. By the stability results above and Jensen's inequality, we have

$$\mathbb{E}\|\widehat{T}_n - T_0\|^2_{L^2(\rho)} \lesssim \mathbb{E}W_2(\mu_n,\mu) \le \left(\mathbb{E}W_2^2(\mu_n,\mu)\right)^{1/2}.$$

We conclude by applying Theorem 2.22. □

2.9 Algorithms

2.9.1 Optimal transport as a linear program

As previously mentioned, the source and target measures are rarely available in closed-form, as we often merely have access to samples from the two measures. In this case, we can write the optimal transport problem (for any reasonable cost function) as a simple linear program, as first proposed by [Kan42].

Proposition 2.26. *For $X_1,\dots,X_n \sim \rho$ and $Y_1,\dots,Y_m \sim \mu$, the finite dimensional Kantorovich problem (2.4) for the empirical measures ρ_n and μ_m is the following linear program*

$$\tfrac{1}{2}W_2^2(\rho_n,\mu_m) = \min_{\Pi\in\mathbb{R}_+^{n\times m}} \langle C,\Pi\rangle \quad s.t. \quad \Pi\mathbf{1}_m = \tfrac{1}{n}\mathbf{1}_n\,, \Pi^\top\mathbf{1}_n = \tfrac{1}{m}\mathbf{1}_m\,, \tag{2.31}$$

where $\mathbf{1}_m$ (resp. $\mathbf{1}_n$) is the vector of all ones of size m (resp. n) and $C \in \mathbb{R}_+^{n\times m}$ with $C_{ij} = \frac{1}{2}\|X_i - Y_j\|^2$ for all $(i,j) \in [n]\times[m]$.

At the time it was first written down, linear programs were not implementable in any reasonable sense. A range of methods have since been proposed, such as the simplex method, the Hungarian algorithm, etc. We encourage those interested to consult [PC19, Chapter 3], though we will not discuss these algorithms in any detail. Instead, we provide the following code snippet which illustrates how to use the `POT` (Python Optimal Transport) package for computing plans and costs [FCGA+21].

Algorithm 1 Discrete optimal transport with `POT`

Input. Arrays $X = (X_1, \dots, X_n)$ and $Y = (Y_1, \dots, Y_m)$ containing the atoms of the empirical measures $\rho_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ and $\mu_m = \frac{1}{m}\sum_{j=1}^m \delta_{Y_j}$, respectively.

```
import numpy as np
import ot

n, m = len(X), len(Y)
a, b = np.ones(n) / n, np.ones(m) / m
C = ot.dist(X, Y, metric="sqeuclidean") / 2

plan = ot.emd(a, b, C)
cost = ot.emd2(a, b, C)
```

To close this section, we give the runtime complexity of the network simplex algorithm, due to [Tar97]:

Theorem 2.27 (Kantorovich linear program runtime). *The discrete Kantorovich linear program (2.31), which has mn variables and $n + m$ equality constraints, can be solved in time $\widetilde{\mathcal{O}}((n+m)nm)$.*

2.9.2 Univariate optimal transport via sorting

In 1D, the discrete Kantorovich problem admits a closed-form solution via sorting. Given n i.i.d. samples $X_1, \dots, X_n \sim \rho$ and $Y_1, \dots, Y_n \sim \mu$ on $\mathbb{R}$, let $X_{(1)} \leq \dots \leq X_{(n)}$ and $Y_{(1)} \leq \dots \leq Y_{(n)}$ denote the order statistics. Note that computing these quantities takes $\mathcal{O}(n \log n)$ time. The following result tells us how to make use of these sorted values.

Proposition 2.28. *Let $\rho, \mu \in \mathcal{P}(\mathbb{R})$ and let $X_1, \dots, X_n \sim \rho$ and $Y_1, \dots, Y_n \sim \mu$, writing their empirical counterparts as ρ_n and μ_n, respectively. Then the optimal transport cost between the two empirical univariate measures is given by*

$$\tfrac{1}{2}W_2^2(\rho_n, \mu_n) = \frac{1}{2n}\sum_{i=1}^n \left(X_{(i)} - Y_{(i)}\right)^2. \tag{2.32}$$

The connection to sorting is related to the following more general result regarding the monotonicity of optimal transport plans (which follows from Kantorovich duality).

Lemma 2.29 (Monotonicity of optimal transport plans). *Let π be an optimal transport*

plan for cost c. Then for any $((x_i, y_i))_{i=1}^n$ in the support of π, it holds that

$$\sum_{i=1}^n c(x_i, y_i) \leq \sum_{i=1}^n c(x_i, y_{i+1}), \quad y_{n+1} = y_1 .$$

Proof of Proposition 2.28. Letting c be the quadratic cost, then Lemma 2.29 states that for any (x, y) and (x', y') in the support of π, it holds that

$$(x - x')(y - y') \geq 0 .$$

So, if $X_{(i)} < X_{(j)}$, then it must be that $Y_{\pi(i)} \leq Y_{\pi(j)}$, else the monotonicity relation does not hold. □

2.9.3 Sliced Wasserstein distances via Monte Carlo

Leveraging the $\mathcal{O}(n \log n)$ runtime for computing univariate optimal transport distances, we can now easily compute Sliced Wasserstein distances via Monte Carlo approximation. First, recall that

$$\tfrac{1}{2}\,\mathrm{SW}_2^2(\rho, \mu) = \int_{\mathbb{S}^{d-1}} \tfrac{1}{2} W_2^2(\rho_\theta, \mu_\theta)\ \mathrm{d}\sigma_{d-1}(\theta) .$$

A natural estimation procedure is the following: Given the fixed samples

1. sample $\theta_1, \ldots, \theta_L \sim \mathrm{Unif}(\mathbb{S}^{d-1})$ by drawing $Z_l \sim \mathcal{N}(0, I_d)$ and setting $\theta_l = Z_l / \|Z_l\|$,
2. for each $l \in [L]$: form projections $\tilde{X}_i^l = \langle X_i, \theta_l \rangle$ and $\tilde{Y}_j^l = \langle Y_j, \theta_l \rangle$, sort each, and compute $\frac{1}{2} W_2^2((\rho_n)_{\theta_l}, (\mu_n)_{\theta_l}) = \frac{1}{2n} \sum_{i=1}^n (\tilde{X}_{(i)}^l - \tilde{Y}_{(i)}^l)^2$,
3. return $\frac{1}{2}\widehat{\mathrm{SW}_2^2} = \frac{1}{L} \sum_{l=1}^L \frac{1}{2} W_2^2((\rho_n)_{\theta_l}, (\mu_n)_{\theta_l})$.

The total runtime is $O(L(dn + n \log n))$. By the law of large numbers, for fixed empirical measures,

$$\tfrac{1}{2}\widehat{\mathrm{SW}_2^2} \to \tfrac{1}{2}\,\mathrm{SW}_2^2(\rho_n, \mu_n) \qquad \text{as } L \to \infty ,$$

with Monte Carlo error of order $O(L^{-1/2})$.

2.9.4 Estimating optimal transport maps

Nearest-neighbor estimator. The most direct approach is to first solve the discrete Kantorovich problem and then use the resulting optimal plan Π_0 to define map values on the observed source samples. Since the ith row of Π_0 has total mass $1/n$, the quantities

$$\widehat{T}_{\Pi_0}(X_i) := n \sum_{j=1}^m \Pi_{0,ij} Y_j \tag{2.33}$$

are well-defined points in the convex hull of the target samples. This construction is called the *barycentric projection* of the optimal plan. Equivalently, $n\Pi_{0,ij}$ can be interpreted as the conditional probability of sending X_i to Y_j, and $\widehat{T}_{\Pi_0}(X_i)$ is the corresponding conditional mean.

Proposition 2.30. *Let Π be any feasible transport plan for (2.31). For every $i \in [n]$,*

$$q_i := \big(n\Pi_{i1}, \dots, n\Pi_{im}\big) \in \Delta_m \,.$$

Thus $\widehat{T}_\Pi(X_i) = \sum_{j=1}^m q_{ij} Y_j$ is the conditional expectation of the target location under the conditional law induced by Π. In the special case $n = m$ and $\Pi = \frac{1}{n} P_\sigma$ for a permutation matrix P_σ, this reduces to

$$\widehat{T}_\Pi(X_i) = Y_{\sigma(i)} \,.$$

The barycentric projection only defines the estimator at the observed points $X_1, \dots, X_n$. To evaluate the map at a new point x, a simple extension is nearest-neighbor interpolation:

$$i_n(x) \in \operatorname*{argmin}_{i \in [n]} \|x - X_i\| \,, \qquad \widehat{T}_{\mathrm{NN}}(x) := \widehat{T}_{\Pi_0}(X_{i_n(x)}) \,. \tag{2.34}$$

Thus, $\widehat{T}_{\mathrm{NN}}$ is piecewise constant on the Voronoi cells generated by the source samples.

Estimators based on neural networks. An alternative is to choose a parametric family of (convex) functions

$$\{\varphi_\theta : \mathbb{R}^d \to \mathbb{R}\}_{\theta \in \Theta}$$

and estimate the Brenier potential by optimizing over θ. The learning problem is then

$$\widehat{\theta} \in \operatorname*{argmin}_{\theta \in \Theta} \Big\{ \tfrac{1}{n} \sum_{i=1}^n \varphi_\theta(X_i) + \tfrac{1}{m} \sum_{j=1}^m \varphi_\theta^*(Y_j) \Big\} \,. \tag{2.35}$$

The parametric Brenier map estimator is then

$$\widehat{T}_{\mathrm{net}}(x) := \nabla \varphi_{\widehat{\theta}}(x) \,. \tag{2.36}$$

A naive parametrization is that of quadratic potentials: for $b \in \mathbb{R}^d$ and $B \in \mathbb{S}_+^d$, write

$$\varphi_{b,B}(x) := \frac{1}{2} \langle x, Bx \rangle + \langle x, b \rangle \,.$$

Here the optimization can be performed in a relatively straightforward manner. On the other end of the spectrum, one can choose a complicated neural network to act as the

parametrization of φ_θ. In this case, the conjugate φ_θ^* may not be available in closed form, so it is often approximated by an inner optimization problem [Amo22]:

$$\varphi_\theta^*(y) = \sup_{z \in \mathbb{R}^d} \{\langle y, z\rangle - \varphi_\theta(z)\}\,.$$

Thus, training a neural optimal transport map typically involves two nested tasks: optimizing the parameters θ of the convex potential and approximately computing the conjugate values at the target samples.

2.9.5 Application: Fréchet Inception Distance (FID)

When generative modeling first started making initial progress, a fundamental question was to devise *quantitative* metrics for discerning the quality of a generated image. To make this more precise, let μ be some distribution of interest and let μ_θ be a generative model (an object that, upon querying, returns samples $Y \sim \mu_\theta$) that is meant to closely approximate μ. How can we determine that μ_θ is indeed close to μ *at scale*?

The answer proposed by the community was the Fréchet Inception Distance (FID), which compares two *collections* of images in a learned feature space. Upon resizing and normalizing the images accordingly, both real and generated images are passed through an Inception-v3 network pretrained on ImageNet (a state-of-the-art classifier many years ago). For each image, we extract the 2048-dimensional activations of the network's final global-average-pooling layer, immediately before the linear classification layer. These features are presumed to encode perceptually and semantically meaningful properties of an image while being less sensitive than raw pixels to small local changes.

Now, let $\mathcal{Z} := \{Z_1, \dots, Z_n\} \subset \mathbb{R}^{2048}$ and $\widehat{\mathcal{Z}} := \{\widehat{Z}_1, \dots, \widehat{Z}_m\} \subset \mathbb{R}^{2048}$ denote the resulting features of the real and generated images, respectively. We then compute the empirical means and covariances of the two feature distributions

$$a = \frac{1}{n}\sum_{i=1}^{n} Z_i\,, \qquad A = \frac{1}{n}\sum_{i=1}^{n}(Z_i - a)(Z_i - a)^\top\,,$$
$$\widehat{a} = \frac{1}{m}\sum_{j=1}^{m} \widehat{Z}_j\,, \qquad \widehat{A} = \frac{1}{m}\sum_{j=1}^{m}(\widehat{Z}_j - \widehat{a})(\widehat{Z}_j - \widehat{a})^\top\,,$$

and compare the empirical distributions via

$$\begin{aligned}\tfrac{1}{2}\,\mathrm{FID}(\mathcal{Z}, \widehat{\mathcal{Z}}) &:= \tfrac{1}{2} W_2^2(\mathcal{N}(a, A), \mathcal{N}(\widehat{a}, \widehat{A})) \\ &= \tfrac{1}{2}\|a - \widehat{a}\|_2^2 + \tfrac{1}{2}\,\mathrm{Tr}\Big(A + \widehat{A} - 2\big(A^{1/2}\widehat{A}A^{1/2}\big)^{1/2}\Big)\,.\end{aligned}$$

Many alternatives exist (see e.g., this new paper for a recent proposal).

3 Entropic optimal transport

Entropic optimal transport forms the cornerstone of modern computational optimal transport. In this chapter, we'll explore the derivation of this formulation, its implementation, and how it connects to the unregularized optimal transport problem from the previous chapter.

3.1 Primal formulation

For two probability measures $P, Q \in \mathcal{P}(\mathbb{R}^d)$, recall that the *Kullback–Leibler divergence* between them is defined as

$$\mathrm{KL}(P \,\|\, Q) := \int \log\Big(\frac{\mathrm{d}P}{\mathrm{d}Q}\Big)\,\mathrm{d}P\,, \tag{3.1}$$

whenever $P \ll Q$ (P is absolutely continuous with respect to Q), and the value diverges to $+\infty$ otherwise. Some fundamental properties of the KL divergence include that it is jointly convex, and strictly convex in its first argument. Importantly, it is also asymmetric and does not satisfy the triangle inequality, which does not allow it to be a metric.

Remark 3.1 (KL divergence between Gaussians). Let $P = \mathcal{N}(a, A)$ and $Q = \mathcal{N}(b, B)$, where $a, b \in \mathbb{R}^d$ and $A, B \succ 0$. Then

$$\mathrm{KL}(P \,\|\, Q) = \frac{1}{2}\Big[\log\frac{\det B}{\det A} - d + \mathrm{tr}(B^{-1}A) + (a-b)^\top B^{-1}(a-b)\Big]\,. \tag{3.2}$$

In particular, taking $Q = \mathcal{N}(0, I)$ gives the useful specialization

$$\mathrm{KL}(\mathcal{N}(a, A) \,\|\, \mathcal{N}(0, I)) = \frac{1}{2}\big[-\log\det A - d + \mathrm{tr}(A) + \|a\|^2\big]\,.$$

Indeed, if $X \sim P$, then expanding the log-density ratio gives

$$\mathrm{KL}(P \,\|\, Q) = \frac{1}{2}\log\frac{\det B}{\det A} - \frac{1}{2}\mathbb{E}\big[(X-a)^\top A^{-1}(X-a)\big] + \frac{1}{2}\mathbb{E}\big[(X-b)^\top B^{-1}(X-b)\big]\,.$$

The middle expectation is d, while

$$\mathbb{E}\big[(X-b)^\top B^{-1}(X-b)\big] = \mathrm{tr}(B^{-1}A) + (a-b)^\top B^{-1}(a-b)\,,$$

which yields (3.2).

We now define the *entropic* optimal transport problem, which takes (2.4) and incorporates the KL divergence as a regularizer.

Definition 3.2 (Entropic optimal transport). For $\varepsilon > 0$, the Kantorovich problem with

entropic regularization is

$$\mathrm{OT}_\varepsilon(\rho,\mu) := \inf_{\pi\in\Gamma(\rho,\mu)} \iint \tfrac{1}{2}\|x-y\|^2 \,\mathrm{d}\pi(x,y) + \varepsilon\,\mathrm{KL}(\pi \,\|\, \rho\otimes\mu)\,. \tag{3.3}$$

Obviously if $\varepsilon = 0$, we recover (2.4), for which we know that uniqueness of solutions is not guaranteed. However, for every positive value $\varepsilon > 0$, we have the following result for EOT.

Proposition 3.3. *Let $\rho,\mu \in \mathcal{P}_2(\mathbb{R}^d)$. For any $\varepsilon > 0$, there exists a unique minimizer to (3.3), written $\pi_\varepsilon \in \Gamma(\rho,\mu)$.*

Next, note that the definition of the KL divergence forces π_ε to have a density with respect to $\rho\otimes\mu$. This is in stark contrast to the optimal transport setting, where the support of π_0 is just a "sliver" on the space of joint probability measures should it uniquely exist.

3.2 Dual formulation and consequences

As before, to obtain the dual for (3.3), we mimic the computations of Proposition 2.8. This leads to

$$\inf_{\pi\geq 0} \sup_{\substack{f\in L^1(\rho)\\ g\in L^1(\mu)}} \int f\,\mathrm{d}\rho + \int g\,\mathrm{d}\mu \\ + \iint \left(\tfrac{1}{2}\|x-y\|^2 - f(x) - g(y)\right) + \varepsilon\left[\log\left(\frac{\mathrm{d}\pi(x,y)}{\mathrm{d}\rho(x)\,\mathrm{d}\mu(y)}\right) - 1\right] \mathrm{d}\pi(x,y) + \varepsilon\,.$$

Writing $a = \mathrm{d}\pi/\,\mathrm{d}(\rho\otimes\mu)$ and $\alpha(x,y) := \frac{1}{2}\|x-y\|^2 - f(x) - g(y)$, and interchanging the inf and sup again, this yields

$$\sup_{\substack{f\in L^1(\rho)\\ g\in L^1(\mu)}} \int f\,\mathrm{d}\rho + \int g\,\mathrm{d}\mu - \varepsilon\Big(\sup_{a\geq 0} \iint (-\alpha(x,y)/\varepsilon)a - a(\log a - 1)\;\mathrm{d}\rho(x)\,\mathrm{d}\mu(y)\Big) + \varepsilon\,.$$

Passing the supremum inside the integral, we see that this is essentially an infinite-dimensional analogue of the convex conjugate operator, acting on the function $h(a) = a(\log(a) - 1)$. One can check that

$$h^*(b) = \sup_a\{ab - h(a)\} = e^b\,, \tag{3.4}$$

with maximizer $a^* = e^b$ as well, resulting in the following unconstrained maximization problem

$$\sup_{\substack{f\in L^1(\rho)\\ g\in L^1(\mu)}} \int f\,\mathrm{d}\rho + \int g\,\mathrm{d}\mu - \varepsilon \iint \exp\big((-\alpha(x,y)/\varepsilon)\big)\,\mathrm{d}\rho(x)\,\mathrm{d}\mu(y) + \varepsilon\,.$$

Expanding our expression of $a(x,y)$, we recover the resulting dual formulation

Proposition 3.4. *For $\varepsilon > 0$, suppose ρ and μ have finite second moments. Then we have the following dual formulation of (3.3)*

$$\mathrm{OT}_\varepsilon(\rho,\mu) := \sup_{\substack{f\in L^1(\rho)\\ g\in L^1(\mu)}} \int f\,\mathrm{d}\rho + \int g\,\mathrm{d}\mu - \varepsilon \iint e^{(f(x)+g(y)-\frac{1}{2}\|x-y\|^2)/\varepsilon}\,\mathrm{d}\rho(x)\,\mathrm{d}\mu(y) + \varepsilon\,, \tag{3.5}$$

The objective function (3.5) should be reminiscent of (2.5), with one glaring exception: the *hard* constraint is now an *unconstrained* optimization problem. Convince yourself that the $\varepsilon \to 0$ limit morally recovers the hard constraint from before.

Definition 3.5 (Optimal entropic Kantorovich potentials). We write the maximizers of (3.5) as $(f_\varepsilon, g_\varepsilon)$, which are called *(optimal) entropic Kantorovich potentials*.

3.2.1 Explicit primal solution recovery

The following proposition explicitly relates the primal and dual solutions for the entropic optimal transport problem; this result is often attributed to [Csi75].

Proposition 3.6. *For $\rho, \mu \in \mathcal{P}_2(\mathbb{R}^d)$, the optimal entropic plan is explicitly written as*

$$\mathrm{d}\pi_\varepsilon(x,y) := \exp\Big(\frac{f_\varepsilon(x) + g_\varepsilon(y) - \frac{1}{2}\|x-y\|^2}{\varepsilon}\Big)\,\mathrm{d}\rho(x)\,\mathrm{d}\mu(y)\,. \tag{3.6}$$

3.2.2 An alternative dual and extending the potentials

We now arrive at an alternative dual formulation. Fix $y \in \mathrm{supp}(\mu)$, $f \in L^1(\rho)$ and maximize (3.5) only in "g"; this yields

$$g(y) = -\varepsilon \log \int e^{(f(x)-\frac{1}{2}\|x-y\|^2)/\varepsilon}\,\mathrm{d}\rho(x) =: f^{(c,\varepsilon)}(y)\,,$$

where the c superscript indicates the quadratic cost (though, in principle, this holds for a wide range of suitable costs). Repeating these manipulations, we arrive at a more compact representation of $\mathrm{OT}_\varepsilon(\rho,\mu)$, written

$$\mathrm{OT}_\varepsilon(\rho,\mu) := \sup_{f\in L^1(\rho)} \int f\,\mathrm{d}\rho + \int f^{(c,\varepsilon)}\,\mathrm{d}\mu\,. \tag{3.7}$$

Given such a representation, it is not surprising that f_ε is optimal if (and only if) it satisfies a fixed-point relationship of the type

$$f_\varepsilon = ((f_\varepsilon)^{(c,\varepsilon)})^{(c,\varepsilon)}\,.$$

Finally, it is worth mentioning that the maximizers f_ε and g_ε are only initially defined on the supports of ρ and μ, respectively. However, there exists an appropriate choice of potentials such that the following relationships hold over all $x \in \mathbb{R}^d$ and $y \in \mathbb{R}^d$:

$$f_\varepsilon(x) = -\varepsilon \log \int e^{(g_\varepsilon(y) - \frac{1}{2}\|x-y\|^2)/\varepsilon} \, \mathrm{d}\mu(y) \qquad (x \in \mathbb{R}^d)\,, \tag{3.8}$$

$$g_\varepsilon(y) = -\varepsilon \log \int e^{(f_\varepsilon(x) - \frac{1}{2}\|x-y\|^2)/\varepsilon} \, \mathrm{d}\rho(x) \qquad (y \in \mathbb{R}^d)\,. \tag{3.9}$$

We will assume these two relationships hold for the remainder of these notes. For more context, see [MNW19; NW21].

3.3 Sinkhorn's algorithm

In this section, we discuss Sinkhorn's algorithm [Cut13; Sin67] and provide two distinct proofs of convergence.

As we saw before, there is a one-to-one correspondence between computing optimal solutions π_ε and $(f_\varepsilon, g_\varepsilon)$, given by the relationship

$$\mathrm{d}\pi_\varepsilon(x,y) = \exp\Big(\frac{f_\varepsilon(x) + g_\varepsilon(y) - \frac{1}{2}\|x-y\|^2}{\varepsilon}\Big) \, \mathrm{d}\rho(x) \, \mathrm{d}\mu(y)\,.$$

What are the f_ε and g_ε that give rise to this expression? How do we compute them algorithmically?

Sinkhorn's algorithm states the following: *iteratively* fit the marginals of π_ε until you arrive at a fixed point, which *must* be the optimal plan.

To make this explicit: initialize $f_\varepsilon^{(0)}(x) = 0, g_\varepsilon^{(0)}(y) = 0$, i.e.,

$$\pi_\varepsilon^{(0)} \propto e^{-\frac{1}{2}\|x-y\|^2/\varepsilon} \, \mathrm{d}\rho(x) \, \mathrm{d}\mu(y)\,.$$

We begin by wanting to enforce the marginal constraint $\int \pi_\varepsilon^{(0)}(x,y) \, \mathrm{d}y = \mathrm{d}\rho(x)$, which means

$$e^{f_\varepsilon^{(1)}(x)/\varepsilon} \, \mathrm{d}\rho(x) \int e^{(g_\varepsilon^{(1)}(y) - \frac{1}{2}\|x-y\|^2)/\varepsilon} \, \mathrm{d}\mu(y) = \mathrm{d}\rho(x)\,.$$

This leads to the update

$$g_\varepsilon^{(1)}(y) := g_\varepsilon^{(0)}(y), \quad f_\varepsilon^{(1)}(x) := -\varepsilon \log \int e^{(g_\varepsilon^{(1)}(y) - \frac{1}{2}\|x-y\|^2)/\varepsilon} \, \mathrm{d}\mu(y)\,. \tag{3.10}$$

Our plan now is

$$\pi_\varepsilon^{(1)} \propto e^{(f_\varepsilon^{(1)}(x) + g_\varepsilon^{(1)}(y) - \frac{1}{2}\|x-y\|^2)/\varepsilon} \, \mathrm{d}\rho(x) \, \mathrm{d}\mu(y)\,,$$

where the first marginal is ρ, but the second marginal is *not* μ. We now want to enforce the other marginal constraint, $\int \pi_\varepsilon^{(1)}(x,y)\,\mathrm{d}x = \mathrm{d}\mu(y)$, resulting in the update

$$f_\varepsilon^{(2)}(x) := f_\varepsilon^{(1)}(x), \quad g_\varepsilon^{(2)}(y) := -\varepsilon \log \int e^{(f_\varepsilon^{(2)}(x) - \frac{1}{2}\|x-y\|^2)/\varepsilon}\,\mathrm{d}\rho(x)\,. \tag{3.11}$$

Now, $\pi_\varepsilon^{(2)}$ has second marginal μ but its first marginal is not equal to ρ.

Sinkhorn's algorithm suggests to repeat this until convergence, as the two marginal constraints have a non-empty intersection, and, eventually, π_ε will have both marginals ρ and μ. Note that this iterative procedure also makes sense from the perspective of (3.8) and (3.9), which are attained by Sinkhorn's algorithm at optimality.

3.3.1 Analysis via coordinate ascent

We now provide a basic runtime analysis of Sinkhorn's algorithm based on the block coordinate-ascent interpretation. To continue, we require Pinsker's inequality, which bounds the total variation distance by the KL-divergence between two measures.

Proposition 3.7 (Pinsker's inequality). *For $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$,*

$$\mathrm{TV}(\mu,\nu) := \tfrac{1}{2}\|\mu - \nu\|_{L^1} \le \sqrt{\tfrac{1}{2}\mathrm{KL}(\mu \,\|\, \nu)}\,. \tag{3.12}$$

We will prove the following theorem.

Theorem 3.8. *Suppose $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$ with compact support, and fix $\varepsilon > 0$. Then the iterates of Sinkhorn's algorithm converge at the following rate*

$$\sup_{f,g} \mathcal{D}_\varepsilon^{\mu\nu}(f,g) - \mathcal{D}_\varepsilon^{\mu\nu}(f_\varepsilon^{(k)}, g_\varepsilon^{(k)}) \lesssim \frac{c_\infty^2}{\varepsilon k}\,,$$

where $c_\infty := \sup_{x,y} c(x,y)$.

Proof. Suppose k is even, so the reverse potential is being updated and the forward potential remains fixed. We can rewrite the iterate update as

$$g_\varepsilon^{(k+1)} = g_\varepsilon^{(k)} - \varepsilon \log\left(\nu^{(k,k)}/\nu\right), \tag{3.13}$$

where $\nu^{(k,k)} := \int \pi^{(k,k)}(x,\cdot)\,\mathrm{d}x$; we symmetrically define $\mu^{(k,k)}$.

Next, note that at any iteration of the algorithm, it holds that $\mathcal{D}_\varepsilon^{\mu\nu}(f^{(k)}, g^{(k)}) = \int f^{(k)}\,\mathrm{d}\mu + \int g^{(k)}\,\mathrm{d}\nu$. Using (3.13) and Pinsker's inequality, we obtain

$$\begin{aligned}
\mathcal{D}_\varepsilon^{\mu\nu}(f^{(k)}, g^{(k)}) - \mathcal{D}_\varepsilon^{\mu\nu}(f^{(k+1)}, g^{(k+1)}) &= \int \varepsilon \log\left(\nu^{(k,k)}/\nu\right)\mathrm{d}\nu \\
&= -\varepsilon \mathrm{KL}(\nu \,\|\, \nu^{(k,k)}) \\
&\le -2\varepsilon \mathrm{TV}^2(\nu, \nu^{(k,k)})\,,
\end{aligned}$$

We next require the following lemma, whose proof we leave as an exercise.

Lemma 3.9. *Under the assumptions of Theorem 3.8, it holds that for any $k \in \mathbb{N}$*

$$\mathcal{D}_\varepsilon^{\mu\nu}(f^{\mu\nu}, g^{\mu\nu}) - \mathcal{D}_\varepsilon^{\mu\nu}(f^{(k)}, g^{(k)}) \le c_\infty\big(\mathrm{TV}(\mu, \mu^{(k,k)}) + \mathrm{TV}(\nu, \nu^{(k,k)})\big)\,.$$

To complete the proof, define the optimality gap via $\Delta_k := \mathcal{D}_\varepsilon^{\mu\nu}(f^{\mu\nu}, g^{\mu\nu}) - \mathcal{D}_\varepsilon^{\mu\nu}(f^{(k)}, g^{(k)})$. Since only one of the two marginal constraints is satisfied at a given iteration, we can use Lemma 3.9 to obtain

$$\Delta_{k+1} - \Delta_k \le \frac{-2\varepsilon\Delta_k^2}{c_\infty^2}\,.$$

Dividing through by $\Delta_k\Delta_{k+1}$, we arrive at

$$\frac{1}{\Delta_k} - \frac{1}{\Delta_{k+1}} \le \frac{-2\varepsilon\Delta_k}{c_\infty^2\Delta_{k+1}} \le \frac{-2\varepsilon}{c_\infty^2}\,,$$

where we used the fact that $\Delta_{k+1} \le \Delta_k$. Taking the sum from $k = 0$ to $k = K-1$ and rearranging the remaining terms from the telescoping series completes the proof. □

3.3.2 Analysis via Bregman gradient descent

We now present a separate analysis based on *Bregman gradient descent*, or *mirror descent*, which offers a third perspective on Sinkhorn's algorithm. First, we rewrite the entropic optimal transport problem (3.3), up to an additive constant, as

$$\mathrm{KL}^\star(\mu, \nu, R) := \inf_{\pi \in \Gamma(\mu,\nu)} \varepsilon\,\mathrm{KL}(\pi \,\|\, R)\,, \tag{3.14}$$

where

$$\mathrm{d}R(x,y) = Z^{-1}e^{-c(x,y)/\varepsilon}\,\mathrm{d}\mu(x)\,\mathrm{d}\nu(y)\,, \qquad Z = \iint e^{-c(x,y)/\varepsilon}\,\mathrm{d}\mu(x)\,\mathrm{d}\nu(y)\,.$$

In other words, for any coupling π we have that

$$\varepsilon\mathrm{KL}(\pi \,\|\, R) = \int c\,\mathrm{d}\pi + \varepsilon\mathrm{KL}(\pi \,\|\, \mu\otimes\nu) + \varepsilon\log Z\,.$$

We will prove the following.

Theorem 3.10. *Suppose $g^{(0)} = 0$ and $\pi^{(0)} = (\mathrm{d}\mu/\,\mathrm{d}R_X)R$, where R_X is the X-marginal of R. Let $\pi^{(k)}$ denote the coupling after k iterations of Sinkhorn's algorithm (so $\pi^{(k)}$ has X-marginal μ and Y-marginal $\nu^{(k)}$). Then, for $k \ge 1$,*

$$\mathrm{KL}(\nu^{(k)} \,\|\, \nu) \le \frac{\mathrm{KL}^\star(\mu, \nu, R)}{\varepsilon k}\,.$$

In particular, if $c(x,y) = \frac{1}{2}\|x-y\|^2$ and $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$, we obtain

$$\mathrm{KL}(\nu^{(k)} \,\|\, \nu) \leq \frac{m_2^2(\mu) + m_2^2(\nu)}{\varepsilon k}\,.$$

The premise of Bregman gradient descent is closely related to the vast literature on *proximal gradient descent*. In this scheme, we want to minimize an unconstrained function $F : \mathbb{R}^d \to \mathbb{R}$ by limiting the movement of the iterates using a *Bregman divergence* B_G, resulting in the iterates

$$x^{(k+1)} = \operatorname*{argmin}_x F(x^{(k)}) + \langle F'(x^{(k)}), x - x^{(k)}\rangle + \mathsf{B}_G(x\|x^{(k)})\,, \tag{3.15}$$

where B_G is defined via a differentiable convex function $G : \mathbb{R}^d \to \mathbb{R}$

$$\mathsf{B}_G(x_2\|x_1) := G(x_2) - G(x_1) - \langle G'(x_1), x_2 - x_1\rangle\,.$$

It is easy to see that the optimality conditions for (3.15) result in

$$G'(x^{(k+1)}) - G'(x^{(k)}) = -F'(x^{(k)})\,. \tag{3.16}$$

Letting $y^\bullet = G'(x^\bullet)$, we can express (3.16) as

$$y^{(k+1)} - y^{(k)} = -F'(x^{(k)})\,.$$

From here, we might hope to draw immediate parallels to (3.13) by clever choices of the functions F and G that act on probability measures. To start, for a fixed measure $\nu \in \mathcal{P}(\mathbb{R}^d)$, define $\rho \mapsto \mathcal{H}_\nu(\rho) := \mathrm{KL}(\rho \,\|\, \nu)$. By computing the first variation of this functional, we see that

$$\rho \mapsto \mathcal{H}_\nu'(\rho) = \log(\rho/\nu)\,,$$

We can write the Sinkhorn update as

$$g^{(k+1)} - g^{(k)} = -\varepsilon \mathcal{H}_\nu'(\nu^{(k)})\,.$$

To draw parallels for the potentials, consider the functional

$$g \mapsto \mathcal{G}(g) := \int \mathcal{F}_\varepsilon^\nu[g]\,\mathrm{d}\mu = \varepsilon \int \log\Big(\int e^{(g(y)-c(x,y))/\varepsilon}\,\mathrm{d}\nu(y)\Big)\,\mathrm{d}\mu(x)\,.$$

We then compute

$$\frac{\mathrm{d}\mathcal{G}'(g)}{\mathrm{d}\nu}(y) = \int \exp\left(\frac{g(y) - \mathcal{F}_\varepsilon^\nu[g](x) - c(x,y)}{\varepsilon}\right)\,\mathrm{d}\mu(x)\,,$$

which is simply the Y-projection of the entropic plan given by potentials $(-\mathcal{F}_\varepsilon^\nu[g], g)$. Indeed,

$$\frac{\mathrm{d}\pi_g}{\mathrm{d}(\mu \otimes \nu)}(x,y) = \exp\left(\frac{g(y) - \mathcal{F}_\varepsilon^\nu[g](x) - c(x,y)}{\varepsilon}\right)$$

has X-marginal μ by construction. Thus,

$$\mathcal{G}'(g^{(k+1)}) = \nu^{(k+1)}\,, \quad \mathcal{G}'(g^{(k)}) = \nu^{(k)}\,.$$

As convex conjugation reverses derivatives, we can finally write the Sinkhorn update as

$$(\mathcal{G}^*)'(\nu^{(k+1)}) - (\mathcal{G}^*)'(\nu^{(k)}) = -\varepsilon \mathcal{H}_\nu'(\nu^{(k)})\,, \tag{3.17}$$

thus matching the Bregman gradient descent scheme. The following result summarizes a well-known optimization rate for Bregman gradient descent.

Theorem 3.11. *Let F, G be convex differentiable functions, and suppose that the Bregman divergence of F is bounded by that of G, i.e., for $x, \tilde{x}$,*

$$\mathsf{B}_F(\tilde{x}\|x) \le \mathsf{B}_G(\tilde{x}\|x)\,. \tag{3.18}$$

Then, for all $k \ge 1$,

$$F(x^{(k)}) \le F(x) + \frac{\mathsf{B}_G(x\|x^{(0)})}{k}\,.$$

In order to use Theorem 3.11 to prove Theorem 3.10, we will need the following lemma.

Lemma 3.12. *Let g_1, g_2 be potentials, and define for $i \in \{1,2\}$*

$$\frac{\mathrm{d}\pi_i}{\mathrm{d}R}(x,y) = Z \exp\big((-\mathcal{F}_\varepsilon^\nu[g_i](x) + g_i(y))/\varepsilon\big)\,.$$

Letting ν_i be the Y-marginal of π_i, the following equality holds:

$$\mathsf{B}_{\mathcal{G}}(g_1\|g_2) = \mathsf{B}_{\mathcal{G}^*}(\nu_2\|\nu_1) = \varepsilon \mathrm{KL}(\pi_2 \,\|\, \pi_1)\,.$$

Proof of Theorem 3.10. Let ν_1, ν_2 be any two probability measures. First, note the following equality holds true

$$\mathsf{B}_{\mathcal{H}_\nu}(\nu_2\|\nu_1) = \mathsf{B}_{\mathcal{H}}(\nu_2\|\nu_1) = \mathrm{KL}(\nu_2 \,\|\, \nu_1)\,.$$

If we now choose ν_1, π_1 and ν_2, π_2 as in Lemma 3.12, by an application of Jensen's inequality, it follows that

$$\varepsilon \mathsf{B}_{\mathcal{H}_\nu}(\nu_2\|\nu_1) = \varepsilon \mathrm{KL}(\nu_2 \,\|\, \nu_1) \le \varepsilon \mathrm{KL}(\pi_2 \,\|\, \pi_1) = \mathsf{B}_{\mathcal{G}^*}(\nu_2\|\nu_1)\,.$$

To invoke Theorem 3.11, we choose $F \equiv \varepsilon\mathcal{H}_\nu$ and $G \equiv \mathcal{G}^*$ which implies

$$\varepsilon\mathcal{H}_\nu(\nu^{(k)}) = \varepsilon\mathrm{KL}(\nu^{(k)} \,\|\, \nu) \leq \frac{\mathsf{B}_{\mathcal{G}^*}(\nu\|\nu^{(0)})}{k} = \frac{\varepsilon\mathrm{KL}(\pi^{\mu\nu} \,\|\, \pi^{(0)})}{k}\,.$$

Since $\mathrm{d}\pi^{(0)} / \mathrm{d}R = \mathrm{d}\mu / \mathrm{d}R_X$, we have that

$$\begin{aligned}\varepsilon\mathrm{KL}(\pi^{\mu\nu} \,\|\, \pi^{(0)}) &= \mathrm{KL}^\star(\mu,\nu,R) - \varepsilon\mathrm{KL}(\mu \,\|\, R_X)\\ &\leq \mathrm{KL}^\star(\mu,\nu,R)\,,\end{aligned}$$

where in the last line, we used that R and (thus) R_X are probability measures in order to bound $\mathrm{KL}(\mu \,\|\, R_X) \geq 0$. For the quadratic cost, the result follows as

$$\begin{aligned}\mathrm{KL}^\star(\mu,\nu,R) &\leq \varepsilon\mathrm{KL}(\mu\otimes\nu \,\|\, R)\\ &= \iint \tfrac{1}{2}\|x-y\|^2 \,\mathrm{d}\mu(x)\,\mathrm{d}\nu(y) + \varepsilon\log Z\\ &\leq m_2^2(\mu) + m_2^2(\nu).\end{aligned}$$

□

3.4 Entropic Brenier maps

In this section, we use ideas from entropic regularization to define an analogue of Brenier's optimal transport map.

In what follows, the following definition will be convenient.

Definition 3.13. For $\varepsilon > 0$, let $(f_\varepsilon, g_\varepsilon)$ be the entropic Kantorovich potentials that maximize (3.5), extended to all of $\mathbb{R}^d$ as in (3.8) and (3.9). The *entropic Brenier potentials* are then defined as

$$(\varphi_\varepsilon, \psi_\varepsilon) = (\tfrac{1}{2}\|\cdot\|^2 - f_\varepsilon, \tfrac{1}{2}\|\cdot\|^2 - g_\varepsilon) \tag{3.19}$$

From a pure pattern-matching perspective, entropic Brenier maps are a natural object to define. A natural question to ask is the following: given that the optimal transport map is given by $\nabla\varphi_0$ (which is not trivial to compute in practice), is there a way to leverage entropic optimal transport to obtain an approximate mapping? This is the basis of [PNW21; SDFC+18], where the *barycentric projection* of the entropic plan was used as a proxy for the true map.

Definition 3.14 (Entropic map). Given the optimal entropic plan π_ε between ρ and μ, we define its barycentric projection to be

$$T_\varepsilon(x) := \int y \,\mathrm{d}\pi_\varepsilon^x(y) = \mathbb{E}_{\pi_\varepsilon}[Y \mid X = x]\,. \tag{3.20}$$

We also call this the *entropic (Brenier) map*.

However, a priori, (3.20) is only defined ρ-almost everywhere. The purpose of the entropic map is to be able to map *any* new input in $\mathbb{R}^d$ into the target space. In order to make this rigorous, we appeal to the optimality conditions (3.8), resulting in the following expression

Proposition 3.15. *For $\varepsilon > 0$, the entropic map is equivalent to*

$$T_\varepsilon = \nabla \varphi_\varepsilon \,. \tag{3.21}$$

Proof. We proceed by direct computation, using the relationship (3.8):

$$\begin{aligned}
\nabla \varphi_\varepsilon(x) = \nabla(\tfrac{1}{2}\|\cdot\|^2 - f_\varepsilon)(x) &= x + \varepsilon \frac{\int \frac{-(x-y)}{\varepsilon} e^{(g_\varepsilon(y) - \frac{1}{2}\|x-y\|^2)/\varepsilon} \,\mathrm{d}\mu(y)}{\int e^{(g_\varepsilon(y) - \frac{1}{2}\|x-y\|^2)/\varepsilon} \,\mathrm{d}\mu(y)} \\
&= \frac{\int y e^{(g_\varepsilon(y) - \frac{1}{2}\|x-y\|^2)/\varepsilon} \,\mathrm{d}\mu(y)}{\int e^{(g_\varepsilon(y) - \frac{1}{2}\|x-y\|^2)/\varepsilon} \,\mathrm{d}\mu(y)} \\
&= \mathbb{E}_{\pi_\varepsilon}[Y|X=x] \,.
\end{aligned}$$

□

Exercise 3.1. Prove that $\nabla^2 \varphi_\varepsilon(x) = \varepsilon^{-1}\mathrm{Cov}_{\pi_\varepsilon}(Y|X=x)$, and conclude that φ_ε is always convex. Similarly, show $\nabla^2 \psi_\varepsilon(y) = \varepsilon^{-1}\mathrm{Cov}_{\pi_\varepsilon}(X|Y=y)$ and is also convex, though φ_ε and ψ_ε are *not* convex conjugates of one another.

3.5 Approximation to unregularized optimal transport

The main selling point for entropic optimal transport in practical applications is that it provides a computationally friendly surrogate for $\frac{1}{2}W_2^2(\rho,\mu)$. Indeed, when the Hungarian algorithm takes $\mathcal{O}(n^3)$, Sinkhorn's algorithm takes $\mathcal{O}(n^2/\varepsilon^2)$; see [AWR17]. If it's faster to compute, just how close is this approximation? How close is the entropic Brenier map to its unregularized counterpart? These are the questions tackled here. (Unfortunately, some of the proofs in this section are unenlightening enough that they will not be included in their entirety, at least not at the moment.)

3.5.1 Small regularization regime of costs: The general case

We first investigate the approximation of the entropic cost to $\frac{1}{2}W_2^2$. Before stating the result, let $\Lambda_\varepsilon := (2\pi\varepsilon)^{d/2}$, and let $I_0(\rho,\mu)$ denote the *integrated relative Fisher information along Wasserstein geodesics* i.e.,

$$I_0(\rho,\mu) = \int_0^1 \mathbb{E}_{\mu_t}\|\nabla \log \mu_t\|^2 \,\mathrm{d}t \,,$$

where μ_t is the optimal transport path from ρ to μ (recall (2.20)). The following fundamental approximation result is due to [CRLV+20; CT21].

Theorem 3.16 (Small-regularization expansion of entropic costs). *Assume ρ and μ have compact support with sufficiently regular positive densities, and suppose $I_0(\rho,\mu) < +\infty$. Then, as $\varepsilon \to 0$,*

$$\mathrm{OT}_\varepsilon(\rho,\mu) = \tfrac{1}{2}W_2^2(\rho,\mu) - \varepsilon \log \Lambda_\varepsilon - \frac{\varepsilon}{2}\big(\mathcal{H}(\rho)+\mathcal{H}(\mu)\big) + O\big(\varepsilon^2 I_0(\rho,\mu)\big)\,. \tag{3.22}$$

A useful quantitative version is

$$\mathrm{OT}_\varepsilon(\rho,\mu) - \tfrac{1}{2}W_2^2(\rho,\mu) + \varepsilon \log \Lambda_\varepsilon + \frac{\varepsilon}{2}\big(\mathcal{H}(\rho)+\mathcal{H}(\mu)\big) \le \frac{\varepsilon^2}{8} I_0(\rho,\mu)\,. \tag{3.23}$$

To summarize the result, it says that the entropic OT cost is approximately the unregularized cost up to an initial factor of $O(\varepsilon \log \varepsilon)$, followed by progressively higher-order terms.

A rigorous proof of this result is outside the scope of these notes; though the upper bound (3.23) is quite easy to prove once one has access to the machinery in Chapter **??**.

3.5.2 Small regularization regime of costs: The Gaussian case

While we did not see the proof of Theorem 3.16, one might still ask whether, under certain (potentially strong) conditions, one can do better.

As the Gaussian-to-Gaussian family is the standard place to start, our conversation continues there. The following result describes the form of the optimal entropic plan in this setting.

Theorem 3.17. *For $\rho = \mathcal{N}(0,A)$ and $\mu = \mathcal{N}(b,B)$, the optimal entropic plan is given by the following joint probability measure*

$$\pi_\varepsilon = \mathcal{N}\left(\begin{pmatrix}0\\ b\end{pmatrix}, \begin{pmatrix}A & \Sigma_\varepsilon\\ \Sigma_\varepsilon^\top & B\end{pmatrix}\right), \tag{3.24}$$

where $\Sigma_\varepsilon = A^{1/2}(A^{1/2}BA^{1/2} + \frac{\varepsilon^2}{4}I)^{1/2}A^{-1/2} - \frac{\varepsilon}{2}I$.

The proof of this result is quite technical; we refer the reader to [JMPC20; MGM22] and take the result at face value.

From here, we have the following corollary pertaining to the expansion of the entropic cost.

Corollary 3.18. *For $\rho = \mathcal{N}(0,A)$ and $\mu = \mathcal{N}(b,B)$ with $A, B \succ 0$, let*

$$M := A^{1/2}BA^{1/2} \quad \text{and} \quad D_\varepsilon := \big(M + \tfrac{\varepsilon^2}{4}I\big)^{1/2}.$$

Then the entropic cost admits the exact expression

$$\mathrm{OT}_\varepsilon(\rho,\mu) = \frac{1}{2}\|b\|^2 + \frac{1}{2}\operatorname{Tr}(A+B) - \operatorname{Tr}(D_\varepsilon) + \frac{d\varepsilon}{2} + \frac{\varepsilon}{2}\log\det\Big(\varepsilon^{-1}\big(D_\varepsilon + \tfrac{\varepsilon}{2}I\big)\Big)\,.$$

Consequently, as $\varepsilon \to 0$,

$$\begin{aligned}\mathrm{OT}_\varepsilon(\rho,\mu) &= \tfrac{1}{2}W_2^2(\rho,\mu) + \frac{d\varepsilon}{2}(1-\log\varepsilon) + \frac{\varepsilon}{4}\log\det M + \frac{\varepsilon^2}{8}\operatorname{Tr}(M^{-1/2}) + O(\varepsilon^4)\\ &= \tfrac{1}{2}W_2^2(\rho,\mu) - \varepsilon\log\Lambda_\varepsilon - \frac{\varepsilon}{2}\big(\mathcal{H}(\rho)+\mathcal{H}(\mu)\big) + \frac{\varepsilon^2}{8}I_0(\rho,\mu) + O(\varepsilon^4)\,,\end{aligned}$$

where $I_0(\rho,\mu) = \operatorname{Tr}(M^{-1/2})$ *in the Gaussian case. Thus the second-order term in* (3.23) *is sharp for Gaussian marginals.*

3.5.3 Small regularization regime of maps: The Gaussian case

We now investigate the closeness of entropic Brenier maps to Brenier maps in the Gaussian case. Following the definition as a barycentric projection, one can readily verify that

$$T_\varepsilon(x) = A^{-1/2}\big[(A^{1/2}BA^{1/2} + \tfrac{\varepsilon^2}{4}I)^{1/2} - \tfrac{\varepsilon}{2}I\big]A^{-1/2}x + b\,. \tag{3.25}$$

For even greater simplicity, let's consider the case where $A = I$, then this amounts to

$$T_\varepsilon(x) = \big[(B + \tfrac{\varepsilon^2}{4}I)^{1/2} - \tfrac{\varepsilon}{2}I\big]x + b\,.$$

We know from previous computations that $T_0(x) = B^{1/2}x + b$ and so to assess closeness, we are interested in obtaining upper bounds on the following quantity

$$\|T_\varepsilon - T_0\|^2_{L^2(\rho)} = \mathbb{E}_{X\sim\mathcal{N}(0,I)}\|MX\|^2 = \operatorname{Tr}(M^\top M)\,,$$

where $M = (B + \varepsilon^2 I/4)^{1/2} - (\varepsilon/2)I - B^{1/2}$. Setting now $D := (B + \frac{\varepsilon^2}{4}I)^{1/2} - B^{1/2}$ we expand under the trace.

$$\begin{aligned}\|T_\varepsilon - T_0\|^2_{L^2(\rho)} &= \operatorname{Tr}\big((D - \tfrac{\varepsilon}{2}I)^2\big)\\ &= \frac{d\varepsilon^2}{4} + \operatorname{Tr}\big(D(D-\varepsilon I)\big) \le \frac{d\varepsilon^2}{4}\,,\end{aligned}$$

where the last line used that $D(D-\varepsilon I) \preceq 0$ which is easy to verify. More generally, for $\rho = \mathcal{N}(0,A)$ and $\mu = \mathcal{N}(b,B)$, it holds that for ε small enough

$$\|T_\varepsilon - T_0\|^2_{L^2(\rho)} \le \frac{\varepsilon^2}{4}I_0(\rho) + O(\varepsilon^4)\,,$$

where recall $I_0(\rho) = \mathbb{E}_\rho\|\nabla\log\rho\|^2$ is the Fisher information of ρ.

3.5.4 Small regularization regime of maps: The general case

We now discuss convergence of the entropic Brenier map to the unregularized Brenier map beyond the Gaussian-to-Gaussian case. This result first appeared in [PNW21] though recent developments have enabled "slower" approximation rates of estimation albeit under significantly weaker assumptions.

Theorem 3.19 (Convergence of the entropic Brenier map). *Suppose ρ and μ have densities with compact support in $\Omega \subseteq \mathbb{R}^d$, $\varphi_0 \in C^2(\Omega)$, $\varphi_0^* \in C^{\alpha+1}(\Omega)$ for some $\alpha > 3$, and*

$$0 \prec \ell I_d \preceq \nabla^2 \varphi_0(x) \preceq L I_d$$

on Ω. Then

$$\|T_\varepsilon - T_0\|^2_{L^2(\rho)} \lesssim \varepsilon^2 I_0(\rho, \mu)\,. \tag{3.26}$$

More generally, under lower smoothness $\alpha > 1$, one obtains the interpolation rate

$$\|T_\varepsilon - T_0\|^2_{L^2(\rho)} \lesssim \varepsilon^{\min\{4,\alpha+1\}/2}\,.$$

We now describe the main ideas of the proof.

Let $x^\star := T_0(x)$ and define the optimality gap

$$D[y \mid x^\star] := \frac{1}{2}\|x - y\|^2 - f_0(x) - g_0(y) = -x^\top y + \varphi_0(x) + \varphi_0^*(y)\,,$$

where, by optimality, it holds that $D[y \mid x^\star] \geq 0$ and $D[x^\star \mid x^\star] = 0$. A second-order Taylor expansion around $x^\star$ gives

$$\begin{aligned} D[y \mid x^\star] &= \frac{1}{2}(y - x^\star)^\top \nabla^2 \varphi_0^*(x^\star)(y - x^\star) + o(\|y - x^\star\|^2) \\ &= \frac{1}{2}(y - x^\star)^\top (\nabla^2 \varphi_0(x))^{-1}(y - x^\star) + o(\|y - x^\star\|^2)\,, \end{aligned}$$

where the second line used a fundamental relationship in convex analysis. Thus the conditional density

$$q_\varepsilon^x(y) := \frac{1}{Z_\varepsilon(x)\Lambda_\varepsilon} \exp\Big(-\frac{1}{\varepsilon} D[y \mid x^\star]\Big) \tag{3.27}$$

behaves like

$$q_\varepsilon^x \approx \mathcal{N}\big(x^\star, \varepsilon \nabla^2 \varphi_0(x)\big)\,.$$

The premise of the proof is to then show that

$$\mathbb{E}_{\pi_\varepsilon}[Y|X = x] \approx \int y \,\mathrm{d}q_\varepsilon^x(y)\,, \quad \text{and} \quad q_\varepsilon^x(y) \approx T_0(x)\,.$$

To make sense of this, the following lemma states that q_ε^x is effectively sub-Gaussian.

Lemma 3.20 (Gaussian localization). *Under the assumptions of Theorem 3.19, let $Y^x \sim q_\varepsilon^x$, $\bar{y}_\varepsilon(x) := \mathbb{E}Y^x$, and $\Delta_\varepsilon(x) := \bar{y}_\varepsilon(x) - T_0(x)$. Then, uniformly for small ε,*

$$\varepsilon^{-1/2}\big(Y^x - \bar{y}_\varepsilon(x)\big) \quad \text{has Gaussian tails}, \qquad \|\Delta_\varepsilon(x)\|^2 \lesssim \varepsilon^2\,,$$

and the normalizing constant satisfies

$$\int \big|\log Z_\varepsilon(x) - \tfrac{1}{2}\log\det\nabla^2\varphi_0(x)\big|\,\mathrm{d}\rho(x) \lesssim \varepsilon\,.$$

Moreover, for $a \gtrsim \varepsilon$ and any measurable $h : \mathbb{R}^d \to \mathbb{R}^d$,

$$\iint \exp\big(h(x)^\top(y - T_0(x)) - a\|h(x)\|^2\big)q_\varepsilon^x(y)\,\mathrm{d}y\,\mathrm{d}\rho(x) \le \int \exp\big(\|\Delta_\varepsilon(x)\|^2/(4\varepsilon)\big)\,\mathrm{d}\rho(x)\,.$$

Another proof ingredient is the enlarged dual representation

$$\mathrm{OT}_\varepsilon(\rho,\mu) = \sup_{\eta\in L^1(\pi_\varepsilon)}\Big\{\iint \eta\,\mathrm{d}\pi_\varepsilon - \varepsilon\iint e^{(\eta(x,y)-\frac{1}{2}\|x-y\|^2)/\varepsilon}\,\mathrm{d}\rho(x)\,\mathrm{d}\mu(y) + \varepsilon\Big\}\,.$$

One then tests this dual with

$$\begin{aligned}\eta_h(x,y) &:= \frac{1}{2}\|x-y\|^2 + \varepsilon\Big[j_h(x,y) + \log\frac{q_\varepsilon^x(y)}{\mu(y)}\Big]\\ &:= \frac{1}{2}\|x-y\|^2 + \varepsilon\Big[h(x)^\top(y-T_0(x)) - a\|h(x)\|^2 + \log\frac{q_\varepsilon^x(y)}{\mu(y)}\Big]\end{aligned}$$

where $a \asymp \varepsilon$ and $h : \mathbb{R}^d \to \mathbb{R}^d$ is arbitrary. Substituting η_h into the new dual representation of OT_ε and dividing through by ε gives

$$\begin{aligned}\iint j_h\,\mathrm{d}\pi_\varepsilon + \iint \log\Big(\frac{q_\varepsilon^x(y)e^{\|x-y\|^2/(2\varepsilon)}}{\mu(y)}\Big)\,\mathrm{d}\pi_\varepsilon(x,y)\\ - \iint e^{j_h(x,y)}q_\varepsilon^x(y)\,\mathrm{d}y\,\mathrm{d}\rho(x) + 1 \le \varepsilon^{-1}\,\mathrm{OT}_\varepsilon(\rho,\mu)\,.\end{aligned}$$

Since $D[y \mid x^\star] = \frac{1}{2}\|x-y\|^2 - f_0(x) - g_0(y)$ and $\pi_\varepsilon \in \Gamma(\rho,\mu)$, the second term on the left-hand side equals

$$\frac{W_2^2(\rho,\mu)}{2\varepsilon} - \log\Lambda_\varepsilon - \int \log Z_\varepsilon\,\mathrm{d}\rho - \mathcal{H}(\mu)\,.$$

Applying a change-of-variables identity for $T_0 = \nabla\varphi_0$,

$$\mathcal{H}(\mu) - \mathcal{H}(\rho) = -\int \log\det\nabla^2\varphi_0\,\mathrm{d}\rho\,,$$

then implies that

$$\iint \log\Big(\frac{q_\varepsilon^x(y)e^{\|x-y\|^2/(2\varepsilon)}}{\mu(y)}\Big)\,\mathrm{d}\pi_\varepsilon(x,y) = \frac{W_2^2(\rho,\mu)}{2\varepsilon} - \log\Lambda_\varepsilon - \frac{1}{2}\big(\mathcal{H}(\rho)+\mathcal{H}(\mu)\big) - r_\varepsilon\,,$$

where we additionally define

$$r_\varepsilon := \int \left(\log Z_\varepsilon(x) - \tfrac{1}{2} \log\det \nabla^2 \varphi_0(x) \right) \mathrm{d}\rho(x) \,,$$

Recall that Lemma 3.20 gives $|r_\varepsilon| \lesssim \varepsilon$. Combining the preceding display with (3.23), and using that q_ε^x integrates to one gives

$$\sup_h \left\{ \iint j_h \,\mathrm{d}\pi_\varepsilon - \iint \left(e^{j_h(x,y)} - 1\right) q_\varepsilon^x(y) \,\mathrm{d}y \,\mathrm{d}\rho(x) \right\} \le \frac{\varepsilon}{8} I_0(\rho,\mu) + r_\varepsilon \,.$$

Finally, we use the last results in Lemma 3.20 to imply that

$$\iint \left(e^{j_h(x,y)} - 1\right) q_\varepsilon^x(y) \,\mathrm{d}y \,\mathrm{d}\rho(x) \lesssim \varepsilon$$

uniformly in h. Hence

$$\sup_h \iint \left\{ h(x)^\top (y - T_0(x)) - a\|h(x)\|^2 \right\} \mathrm{d}\pi_\varepsilon(x,y) \lesssim \varepsilon I_0(\rho,\mu) + \varepsilon \,.$$

The proof concludes by choosing

$$h(x) = \frac{T_\varepsilon(x) - T_0(x)}{2a}$$

where we make use of the choice $a \asymp \varepsilon$ to yield the correct scaling.

3.5.5 Sinkhorn divergence and debiased entropic maps

One main drawback of using $\mathrm{OT}_\varepsilon(\rho,\mu)$ as a discrepancy measure is that it is *biased*: for $\varepsilon > 0$, $\mathrm{OT}_\varepsilon(\rho,\rho) > 0$ even though $\frac{1}{2}W_2^2(\rho,\rho) = 0$. Nevertheless, the null case ($\rho = \mu$) has a particularly nice structural form.

Proposition 3.21 (Entropic self-transport)**.** *For the self-transport problem* $\mathrm{OT}_\varepsilon(\rho,\rho)$*, there is a single self-potential* α_ε *satisfying*

$$\begin{aligned} \mathrm{OT}_\varepsilon(\rho,\rho) &= \sup_{\alpha \in L^1(\rho)} \left\{ 2\int \alpha \,\mathrm{d}\rho - \varepsilon \iint e^{(\alpha(x)+\alpha(y)-\frac{1}{2}\|x-y\|^2)/\varepsilon} \,\mathrm{d}\rho(x)\,\mathrm{d}\rho(y) + \varepsilon \right\} \\ &= 2\int \alpha_\varepsilon \,\mathrm{d}\rho \,. \end{aligned} \tag{3.28}$$

Moreover, when $I_0(\rho) < +\infty$*,*

$$\mathrm{OT}_\varepsilon(\rho,\rho) + \varepsilon \log \Lambda_\varepsilon + \varepsilon \mathcal{H}(\rho) \le \frac{\varepsilon^2}{8} I_0(\rho) \,. \tag{3.29}$$

Instead, it is always preferred to compute the *Sinkhorn divergence*, which removes this bias by subtracting the self-transport costs.

Definition 3.22 (Sinkhorn divergence). For $\rho, \mu \in \mathcal{P}_2(\mathbb{R}^d)$ and $\varepsilon > 0$, the *Sinkhorn divergence* is defined as

$$S_\varepsilon(\rho, \mu) := \mathrm{OT}_\varepsilon(\rho, \mu) - \tfrac{1}{2}\left[\mathrm{OT}_\varepsilon(\rho, \rho) + \mathrm{OT}_\varepsilon(\mu, \mu)\right]. \tag{3.30}$$

Per the name, the following proposition verifies that it is indeed an appropriate divergence (which we present without proof; see [FSVA+19] instead). Notably, it does not satisfy the triangle inequality, preventing it from being a true metric on the space of probability measures.

Proposition 3.23. *The Sinkhorn divergence satisfies:*

1. Non-negativity and identity: $S_\varepsilon(\rho, \mu) \geq 0$, *with equality iff* $\rho = \mu$.
2. Symmetry: $S_\varepsilon(\rho, \mu) = S_\varepsilon(\mu, \rho)$.
3. Convergence to OT: $S_\varepsilon(\rho, \mu) \to \frac{1}{2} W_2^2(\rho, \mu)$ *as* $\varepsilon \to 0^+$.

From this discussion, we can also infer that $T_\varepsilon^{\rho\rho}(x) \neq x$. This suggests to analogously define a *debiased* entropic map as follows:

Definition 3.24 (Debiased entropic map). Let $\varphi_\varepsilon^{\rho\rho}$ be the entropic Brenier potential for the self-transport problem $\mathrm{OT}_\varepsilon(\rho, \rho)$. The *debiased entropic map* from ρ to μ is

$$\overline{T}_\varepsilon^{\rho\mu}(x) := T_\varepsilon^{\rho\mu}(x) - T_\varepsilon^{\rho\rho}(x) + x. \tag{3.31}$$

3.6 Algorithms

3.6.1 Implementation of Sinkhorn's algorithm

Algorithm 2 provides self-contained pseudocode for implementing Sinkhorn's algorithm for empirical measures; the changes are minimal for generic discrete measures.

As the per-iteration cost is $\mathcal{O}(nm)$, the following corollary follows directly from Theorem 3.8.

Corollary 3.25. *Let* $\delta > 0$ *be a tolerance parameter, and let* $\mu_n := n^{-1} \sum_{i=1}^n \delta_{X_i}$ *and* $\nu_m := m^{-1} \sum_{j=1}^m \delta_{Y_j}$ *be empirical measures, with* $X_1, \ldots, X_n \sim \mu$ *and* $Y_1, \ldots, Y_m \sim \nu$. *Then Sinkhorn's algorithm returns a* δ*-accurate approximation of* $\mathrm{OT}_\varepsilon(\mu_n, \nu_m)$ *in* $\mathcal{O}(nm\, c_\infty^2 / (\delta\varepsilon))$.

3.6.2 Implementing entropic Brenier maps

Given the converged empirical potentials $\widehat{g}_\varepsilon \in \mathbb{R}^m$ from Algorithm 2, the optimality condition (3.8) and Proposition 3.15 yield the following out-of-sample formula for the en-

Algorithm 2 Implementation of Sinkhorn's algorithm

Inputs: Source samples $(X_i)_{i=1}^n \sim \mu$, target samples $(Y_j)_{j=1}^m \sim \nu$, parameters $\varepsilon, \delta > 0$, and cost function $(x, y) \mapsto c(x, y)$.
Initialize: $g^{(0)} = \mathbf{0}_m \in \mathbb{R}^m$, $r^{(0)} = +\infty$, and $k = 0$.
while $r^{(k)} > \delta$ **do**
For all $i \in [n]$: $f_i^{(k+1)} = -\varepsilon \log\Big(\frac{1}{m}\sum_{j=1}^m \exp((g_j^{(k)} - c(X_i, Y_j))/\varepsilon\Big)$
For all $j \in [m]$: $g_j^{(k+1)} = -\varepsilon \log\Big(\frac{1}{n}\sum_{i=1}^n \exp((f_i^{(k+1)} - c(X_i, Y_j))/\varepsilon\Big)$
$\Pi_{ij}^{(k+1)} \leftarrow \frac{1}{nm}\exp((f_i^{(k+1)} + g_j^{(k+1)} - c(X_i, Y_j))/\varepsilon)$ for all $(i, j) \in [n] \times [m]$.
$r^{(k+1)} \leftarrow \|\Pi^{(k+1)}\mathbf{1}_m - \frac{1}{n}\mathbf{1}_n\|_1 + \|(\Pi^{(k+1)})^\top \mathbf{1}_n - \frac{1}{m}\mathbf{1}_m\|_1$.
$k \leftarrow k + 1$
end while
Return: $(f^{(k)}, g^{(k)}) \in \mathbb{R}^n \times \mathbb{R}^m$

tropic Brenier map at any new input $x \in \mathbb{R}^d$:

$$T_\varepsilon(x) = \frac{\displaystyle\sum_{j=1}^m \exp\big((g_{\varepsilon,j} - c(x, Y_j))/\varepsilon\big)\, Y_j}{\displaystyle\sum_{j=1}^m \exp\big((g_{\varepsilon,j} - c(x, Y_j))/\varepsilon\big)}. \tag{3.32}$$

This is a weighted average of the target samples, with weights proportional to $\exp((g_{\varepsilon,j} - c(x, Y_j))/\varepsilon)$. At observed source points $x = X_i$, (3.32) reduces to the barycentric projection (2.33).

Algorithm 3 Entropic Brenier map (out-of-sample evaluation)

Inputs: Target samples $(Y_j)_{j=1}^m$, converged potentials $g_\varepsilon \in \mathbb{R}^m$, new point $x \in \mathbb{R}^d$, parameter $\varepsilon > 0$, cost c.
Compute: $w_j \leftarrow \exp((g_{\varepsilon,j} - c(x, Y_j))/\varepsilon)$ for all $j \in [m]$.
Return: $T_\varepsilon(x) = \big(\sum_{j=1}^m w_j Y_j\big) / \big(\sum_{j=1}^m w_j\big)$.

The per-query cost is $\mathcal{O}(md)$ (one pass over target samples to evaluate $c(x, Y_j)$ and accumulate the weighted sum). In practice, the log-sum-exp trick should be applied for numerical stability, especially for small ε.

3.6.3 Numerical implementation with `OTT-JAX`

The following implementation is based off the `OTT-JAX` package [CMPTB+22].

Algorithm 4 Sinkhorn and the entropic Brenier map with `OTT-JAX`

Input. Arrays `X` and `Y`, regularization `eps`, and query points `X_new`.

```
import jax.numpy as jnp
from ott.geometry import costs, pointcloud
from ott.problems.linear import linear_problem
from ott.solvers.linear import sinkhorn

X, Y = jnp.asarray(X), jnp.asarray(Y)
n, m = len(X), len(Y)
a, b = jnp.ones(n) / n, jnp.ones(m) / m

cost_fn = costs.PNormP(p=2)
geom = pointcloud.PointCloud(
    X, Y, cost_fn=cost_fn, epsilon=eps
)
problem = linear_problem.LinearProblem(geom, a=a, b=b)
out = sinkhorn.Sinkhorn(threshold=1e-4)(problem)

plan = out.matrix
cost = out.kl_reg_cost

dual_potentials = out.to_dual_potentials()
T_eps_X = dual_potentials.transport(X)
T_eps_new = dual_potentials.transport(jnp.asarray(X_new))
```

References

[Amo22] Brandon Amos. "On amortizing convex conjugates for optimal transport". In: *arXiv preprint arXiv:2210.12153* (2022) (cit. on p. 24).

[AWR17] Jason Altschuler, Jonathan Weed, and Philippe Rigollet. "Near-linear time approximation algorithms for optimal transport via Sinkhorn iteration". In: *Advances in Neural Information Processing Systems 30*. 2017 (cit. on p. 34).

[Bre91] Yann Brenier. "Polar factorization and monotone rearrangement of vector-valued functions". In: *Comm. Pure Appl. Math.* 44.4 (1991), pp. 375–417 (cit. on p. 9).

[CMPTB+22] Marco Cuturi, Laetitia Meng-Papaxanthos, Yingtao Tian, Charlotte Bunne, Geoff Davis, and Olivier Teboul. "Optimal transport tools (ott): A jax toolbox for all things wasserstein". In: *arXiv preprint arXiv:2201.12324* (2022) (cit. on p. 41).

[CRLV+20] Lenaic Chizat, Pierre Roussillon, Flavien Léger, François-Xavier Vialard, and Gabriel Peyré. "Faster Wasserstein distance estimation with the Sinkhorn divergence". In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 2257–2269 (cit. on p. 35).

[Csi75] I. Csiszár. "I-divergence geometry of probability distributions and minimization problems". In: *Ann. Probability* 3 (1975), pp. 146–158 (cit. on p. 27).

[CT21] Giovanni Conforti and Luca Tamanini. "A formula for the time derivative of the entropic cost and applications". In: *Journal of Functional Analysis* 280.11 (2021), p. 108964 (cit. on p. 35).

[Cut13] Marco Cuturi. "Sinkhorn distances: Lightspeed computation of optimal transport". In: *Advances in Neural Information Processing Systems* 26 (2013) (cit. on p. 28).

[FCGA+21] Rémi Flamary, Nicolas Courty, Alexandre Gramfort, Mokhtar Z. Alaya, Aurélie Boisbunon, Stanislas Chambon, Laetitia Chapel, Adrien Corenflos, Kilian Fatras, Nemo Fournier, Léo Gautheron, Nathalie T.H. Gayraud, Hicham Janati, Alain Rakotomamonjy, Ievgen Redko, Antoine Rolet, Antony Schutz, Vivien Seguy, Danica J. Sutherland, Romain Tavenard, Alexander Tong, and Titouan Vayer. "POT: Python Optimal Transport". In: *Journal of Machine Learning Research* 22.78 (2021), pp. 1–8 (cit. on p. 20).

[FSVA+19] Jean Feydy, Thibault Séjourné, François-Xavier Vialard, Shun-ichi Amari, Alain Trouvé, and Gabriel Peyré. "Interpolating between optimal transport and MMD using Sinkhorn divergences". In: *The 22nd International Conference on Artificial Intelligence and Statistics*. PMLR. 2019, pp. 2681–2690 (cit. on p. 40).

[Gig11] Nicola Gigli. “On Hölder continuity-in-time of the optimal transport map towards measures along a curve”. In: *Proceedings of the Edinburgh Mathematical Society* 54.2 (2011), pp. 401–409 (cit. on p. 15).

[GM96] Wilfrid Gangbo and Robert J McCann. “The geometry of optimal transportation”. In: (1996) (cit. on p. 11).

[GS12] Adityanand Guntuboyina and Bodhisattva Sen. “Covering numbers for convex functions”. In: *IEEE Transactions on Information Theory* 59.4 (2012), pp. 1957–1965 (cit. on p. 19).

[HR21] Jan-Christian Hütter and Philippe Rigollet. “Minimax estimation of smooth optimal transport maps”. In: *The Annals of Statistics* 49.2 (2021), pp. 1166–1194 (cit. on p. 19).

[JMPC20] Hicham Janati, Boris Muzellec, Gabriel Peyré, and Marco Cuturi. “Entropic optimal transport between unbalanced Gaussian measures has a closed form”. In: *Advances in Neural Information Processing Systems* 33 (2020) (cit. on p. 35).

[Kan42] L. Kantorovitch. “On the translocation of masses”. In: *C. R. (Doklady) Acad. Sci. URSS (N.S.)* 37 (1942), pp. 199–201 (cit. on pp. 6, 20).

[MBNWW24] Tudor Manole, Sivaraman Balakrishnan, Jonathan Niles-Weed, and Larry Wasserman. “Plugin estimation of smooth optimal transport maps”. In: *The Annals of Statistics* 52.3 (2024), pp. 966–998 (cit. on p. 15).

[MDC20] Quentin Mérigot, Alex Delalande, and Frédéric Chazal. “Quantitative stability of optimal transport maps and linearization of the 2-Wasserstein space”. In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2020, pp. 3186–3196 (cit. on p. 15).

[Mér26] Quentin Mérigot. “Sharp stability of Brenier maps via quantitative regularity of potentials”. In: (2026) (cit. on p. 16).

[MGM22] Anton Mallasto, Augusto Gerolin, and Hà Quang Minh. “Entropy-regularized 2-Wasserstein distance between Gaussian measures”. In: *Information Geometry* 5.1 (2022), pp. 289–323 (cit. on p. 35).

[MNW19] Gonzalo Mena and Jonathan Niles-Weed. “Statistical bounds for entropic optimal transport: sample complexity and the central limit theorem”. In: *Advances in Neural Information Processing Systems* 32 (2019) (cit. on p. 28).

[NW21] Marcel Nutz and Johannes Wiesel. “Entropic optimal transport: Convergence of potentials”. In: *Probability Theory and Related Fields* (2021), pp. 1–24 (cit. on p. 28).

[PC19] Gabriel Peyré and Marco Cuturi. “Computational optimal transport”. In: *Foundations and Trends® in Machine Learning* 11.5-6 (2019), pp. 355–607 (cit. on p. 20).

[PNW21] Aram-Alexandre Pooladian and Jonathan Niles-Weed. "Entropic estimation of optimal transport maps". In: *arXiv preprint arXiv:2109.12004* (2021) (cit. on pp. 33, 37).

[SDFC+18] Vivien Seguy, Bharath Bhushan Damodaran, Rémi Flamary, Nicolas Courty, Antoine Rolet, and Mathieu Blondel. "Large-scale optimal transport and mapping estimation". In: *International Conference on Learning Representations*. 2018 (cit. on p. 33).

[Sin67] Richard Sinkhorn. "A relationship between arbitrary positive matrices and doubly stochastic matrices". In: *The Annals of Mathematical Statistics* 35.2 (1967), pp. 876–879 (cit. on p. 28).

[Tar97] Robert E Tarjan. "Dynamic trees as search trees via euler tours, applied to the network simplex algorithm". In: *Mathematical Programming* 78.2 (1997), pp. 169–177 (cit. on p. 21).