

Topics in flow-based generative modeling and optimal transport

(work in progress)

Aram-Alexandre Pooladian*

September 27, 2026

Contents

1	Primer on optimal transport	4
1.1	Transport maps	4
1.2	The Monge problem	4
1.3	Couplings	5
1.4	Kantorovich relaxation and dual formulation	6
1.4.1	The 2-Wasserstein distance and general costs	6
1.4.2	The Wasserstein space is a metric space	7
1.4.3	Duality	7
1.4.4	Semi-dual	8
1.5	Brenier's Theorem and generalizations	9
1.5.1	Gaussian to Gaussian case	10
1.5.2	Optimal transport and Brenier's theorem for general costs	11
1.5.3	Paths defined by optimal transport maps	12
1.6	The univariate case	13
1.6.1	Closed-form expression of maps	13
1.6.2	Isometry property	15
1.6.3	Sliced Wasserstein distances	15
1.7	On the stability of optimal transport maps	15
1.7.1	A first instability result	16
1.7.2	Some positive stability results	16
1.8	Statistical optimal transport	17
1.8.1	Preliminaries	17

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

1.8.2	Statistical estimation of Wasserstein distances	18
1.8.3	One-sample map estimation	20
1.9	Algorithms	21
1.9.1	Optimal transport as a linear program	21
1.9.2	Univariate optimal transport via sorting	22
1.9.3	Sliced Wasserstein distances via Monte Carlo	23
1.9.4	Estimating optimal transport maps	23
1.9.5	Application: Fréchet Inception Distance (FID)	25
2	Entropic optimal transport	27
2.1	Primal formulation	27
2.2	Dual formulation and consequences	29
2.2.1	Explicit primal solution recovery	30
2.2.2	An alternative dual and extending the potentials	30
2.3	Sinkhorn’s algorithm	31
2.3.1	Analysis via coordinate ascent	32
2.3.2	Analysis via Bregman gradient descent	34
2.4	Entropic Brenier maps	37
2.5	Approximation to unregularized optimal transport	38
2.5.1	Small regularization regime of costs: The general case	38
2.5.2	Small regularization regime of costs: The Gaussian case	39
2.5.3	Small regularization regime of maps: The Gaussian case	40
2.5.4	Small regularization regime of maps: The general case	41
2.5.5	Sinkhorn divergence and debiased entropic maps	43
2.6	Lipschitz properties of (entropic) transport maps	44
2.6.1	Application: Transferring functional inequalities	45
2.6.2	A proof via entropic optimal transport	46
2.7	Algorithms	47
2.7.1	Implementation of Sinkhorn’s algorithm	47
2.7.2	Implementing entropic Brenier maps	48
2.7.3	Numerical implementation with OTT-JAX	48
3	Flow-based transport	50
3.1	What is a flow?	50
3.2	Deterministic flows and the continuity equation	50
3.2.1	Discretizations	51
3.3	Stochastic flows and the Fokker–Planck equation	52
3.3.1	Brownian motion, Itô processes, Itô’s lemma and isometry	53
3.3.2	Derivation of the Fokker–Planck equation	58
3.4	Properties of SDEs	59
3.4.1	ODEs preserve marginal laws	59

3.4.2	Forward and reverse SDEs	59
3.4.3	Girsanov’s theorem	62
4	Dynamic formulation of optimal transport	65
4.1	Optimal vector fields	65
4.1.1	Benamou–Brenier formulation of optimal transport	65
4.1.2	Proof via dynamic dual formulation	65
4.2	Geometry of optimal transport	67
4.2.1	Primer on Riemannian geometry	67
4.2.2	Riemannian view of the Wasserstein space	67
4.2.3	Otto calculus	70
5	Diffusion models	71
5.1	Derivation via forward and reverse OU	71
5.2	Denoising score matching	72
5.3	Sampling with DDPM	73
5.4	The probability flow ODE and its properties	74
6	Dynamic transport via neural ODEs	76
6.1	Flow matching	76
6.1.1	Algorithm derivation and pseudocode	76
6.2	Alternative derivation via stochastic interpolants	78
6.2.1	Mini-batch sampling	79
6.3	Flow matching beyond Euclidean spaces	80
6.3.1	Riemannian flow matching	80
6.3.2	Wasserstein flow matching	81
7	Schrödinger bridge problem on path space	84
7.1	A minimization problem on path space	84
7.1.1	Equivalence to entropic optimal transport	84
7.1.2	Computing entropic interpolations	85
7.2	Formulation as a drift minimization problem	85
7.3	Algorithms	88
7.3.1	Plug-in estimation of Schrödinger bridges	88
8	Few- and one-step generative algorithms	89
8.1	Flow map matching	89
8.1.1	Distillation	89
8.1.2	Flow map matching loss	90
8.1.3	Eulerian estimation (aka consistency models)	90
8.2	An optimal transport perspective on generative drifting models	90

1 Primer on optimal transport

1.1 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 1.1 (Transport maps). For $\rho, \mu \in \mathcal{P}(\mathbb{R}^d)$, the set of *transport maps* from ρ to μ is

$$\mathcal{T}(\rho, \mu) := \{T : \mathbb{R}^d \rightarrow \mathbb{R}^d \mid X \sim \rho, T(X) \sim \mu\}. \quad (1.1)$$

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

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

Exercise 1.1. 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.

1.2 The Monge problem

The minus sign in Example 1.1 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 1.2. 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 1.2 (Optimal transport map). An *optimal transport map* between $\rho, \mu \in \mathcal{P}_2(\mathbb{R}^d)$ is the solution to

$$T_0 \in \operatorname{argmin}_{T \in \mathcal{T}(\rho, \mu)} \int \frac{1}{2} \|x - T(x)\|^2 d\rho(x), \quad (1.2)$$

where (1.2) is called the *Monge problem*.

1.3 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 (1.2) 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. Figure 1 illustrates this phenomenon with three points in the plane.

Definition 1.3 (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

$$\Gamma(\rho, \mu) := \{\pi \in \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d) \mid \pi(A \times \mathbb{R}^d) = \rho(A), \pi(\mathbb{R}^d \times A) = \mu(A)\}. \quad (1.3)$$

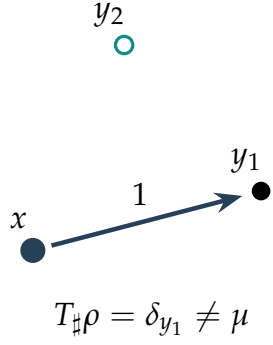
Notice that the set of couplings always includes the independent coupling

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

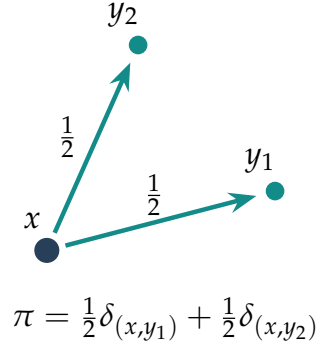
Of course, *if* there exists a deterministic function $T : \mathbb{R}^d \rightarrow \mathbb{R}^d$ such that $T_{\#}\rho = \mu$, then a candidate coupling is

$$\pi_T = (\text{id}, T)_{\#}\rho.$$

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.



(a) A map chooses one destination.



(b) A coupling splits the mass.

Figure 1: A map cannot split an atom, whereas a coupling can.

Lemma 1.4. 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. □

1.4 Kantorovich relaxation and dual formulation

1.4.1 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 1.5 (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

$$\frac{1}{2}W_2^2(\rho, \mu) := \inf_{\pi \in \Gamma(\rho, \mu)} \iint \frac{1}{2} \|x - y\|^2 d\pi(x, y) \quad (1.4)$$

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

$$\pi \mapsto \int \|\cdot\|^2 d\pi$$

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 (1.4) as the definition of the 2-Wasserstein distance.

Theorem 1.6. *If $\rho, \mu \in \mathcal{P}_2(\mathbb{R}^d)$, then (1.4) 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

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

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

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

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

1.4.2 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 1.7. *The Wasserstein space, denoted $\mathbb{W} := (\mathcal{P}_2(\mathbb{R}^d), W_2)$, defines a metric space.*

Exercise 1.3. Use Hölder's inequality to prove that for $1 \leq p \leq q$ it holds that $W_p(\mu, \nu) \leq W_q(\mu, \nu)$.

1.4.3 Duality

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

Proposition 1.8 (Kantorovich dual). *For $\rho, \mu \in \mathcal{P}_2(\mathbb{R}^d)$, the dual Kantorovich problem is the following constrained maximization problem*

$$\frac{1}{2}W_2^2(\rho, \mu) = \sup_{\substack{f \in L^1(\rho) \\ g \in L^1(\mu)}} \int f d\rho + \int g d\mu \quad \text{s.t.} \quad f(x) + g(y) \leq \frac{1}{2}\|x - y\|^2 \forall x, y. \quad (1.5)$$

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

$$\iint f(x) d\pi(x, y) - \int f(x) d\rho(x) = 0 \quad (1.6)$$

over all $f \in L^1(\rho)$, the space of ρ -integrable measurable functions, and analogously with $g \in L^1(\mu)$ for the second marginal. From these definitions, using (f, g) as the two Lagrange multipliers, we have that (1.4) is equivalent to

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

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

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

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

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

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

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

which must hold for all (x, y) . Under this constraint the inner infimum vanishes, resulting in the given dual formulation. □

We conclude with a definition for the optimal maximizers of (1.5).

Definition 1.9 (Optimal Kantorovich potentials). We write the maximizers to (1.5) as (f_0, g_0) , which are called *(optimal) Kantorovich potentials*.

1.4.4 Semi-dual

Looking again at the dual Kantorovich problem, we can make the following adjustments to the objective function: let $m_2^2(\rho) := \int \|\cdot\|^2 d\rho$ and similarly for μ , then (1.5) is

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

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

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

then (1.5) is equivalently written as (up to additive constants)

$$\inf_{\varphi, \psi} \int \varphi \, d\rho + \int \psi \, d\mu \quad \text{s.t.} \quad \varphi(x) + \psi(y) \geq \langle x, y \rangle. \quad (1.7)$$

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

$$\tfrac{1}{2}W_2^2(\rho, \mu) = \tfrac{1}{2}m_2^2(\rho) + \tfrac{1}{2}m_2^2(\mu) - \inf_{\varphi} \mathcal{S}^{\rho\mu}(\varphi), \quad (1.8)$$

where we write

$$\mathcal{S}^{\rho\mu}(\varphi) := \int \varphi \, d\rho + \int \varphi^* \, d\mu, \quad (1.9)$$

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

1.5 Brenier's Theorem and generalizations

The semi-dual formulation (1.9) 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.

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

Theorem 1.11 (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

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

where φ_0 is the minimizer to (1.8), called a Brenier potential. Moreover, the optimal plan π_0 is given by

$$\pi_0 = (\text{id}, T_0)_\# \rho. \quad (1.10)$$

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

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

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

$$\int (\varphi_0(x) + \varphi_0^*(y) - \langle x, y \rangle) d\pi_0(x, y) = 0. \quad (1.11)$$

Notice that, by the Fenchel–Young inequality, the integrand vanishes π_0 -almost everywhere. For a pair $(x_0, y_0) \in \text{supp}(\pi_0)$, it then holds that

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

By definition of convex conjugacy, we have that for every x

$$\langle x, y_0 \rangle - \varphi_0(x) \leq \langle x_0, y_0 \rangle - \varphi_0(x_0) \implies \varphi_0(x) \geq \varphi_0(x_0) + \langle y_0, x - x_0 \rangle,$$

which implies that $y_0 \in \partial \varphi_0(x_0)$, i.e., y_0 lies in the subgradient of φ_0 at x_0 . As φ_0 is convex and ρ has a density with respect to Lebesgue measure, we have that φ_0 is differentiable ρ -almost everywhere on its domain and, moreover, by the marginal constraint, we have that for π_0 -almost every (x, y)

$$y = \nabla \varphi_0(x).$$

This implies that π_0 is concentrated on the graph of $\nabla \varphi_0$, i.e., $\pi_0 = (\text{id}, \nabla \varphi_0)_\# \rho$, which completes the claim. $\square$

1.5.1 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 1.12 (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*

$$\frac{1}{2}W_2^2(\rho, \mu) = \frac{1}{2}\|m_\mu - m_\rho\|^2 + \frac{1}{2}\text{tr}(\Sigma_\rho + \Sigma_\mu - 2(\Sigma_\rho^{1/2}\Sigma_\mu\Sigma_\rho^{1/2})^{1/2}), \quad (1.12)$$

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 semidefinite 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

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

Since this is a positive semidefinite matrix, the proof is complete. $\square$

1.5.2 Optimal transport and Brenier's theorem for general costs

Let $c : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}_+$ 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

$$W_c(\rho, \mu) = \inf_{\pi \in \Gamma(\rho, \mu)} \iint c(x, y) d\pi(x, y). \quad (1.13)$$

We also have a dual representation

$$W_c(\rho, \mu) = \sup_{\substack{f \in L^1(\rho) \\ g \in L^1(\mu)}} \mathcal{D}^{\rho\mu}(f, g) \quad (1.14)$$

with

$$\mathcal{D}^{\rho\mu}(f, g) := \int f d\rho + \int g d\mu \quad \text{s.t.} \quad f(x) + g(y) \leq c(x, y), \quad (1.15)$$

where again the inequality holds almost everywhere.

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

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

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

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

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

$$W_c(\rho, \mu) = \sup_{f \in L^1(\rho)} \int f \, d\rho + \int f^c \, d\mu, \quad (1.16)$$

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

Exercise 1.4. 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 1.13 (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 (ρ -a.e. unique) optimal transport map T_c between ρ and μ exists and is given by*

$$T_c(x) := \nabla_x c(x, \cdot)^{-1} \circ \nabla f_c(x) \quad (\rho\text{-a.e.}) \quad (1.17)$$

where f_c is the maximizer to (1.16).

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 1.14 (Translation-invariant convex costs). *Under the assumptions of Theorem 1.13, for $c(x, y) = h(x - y)$ with h strictly convex, the optimal transport map is given by*

$$T_h(x) = x - (\nabla h^*) \circ \nabla f_h(x), \quad (1.18)$$

where h^* is the convex conjugate of h , and f_h is the maximizer to (1.16) under the cost function above.

1.5.3 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 $t \in [0, 1]$

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

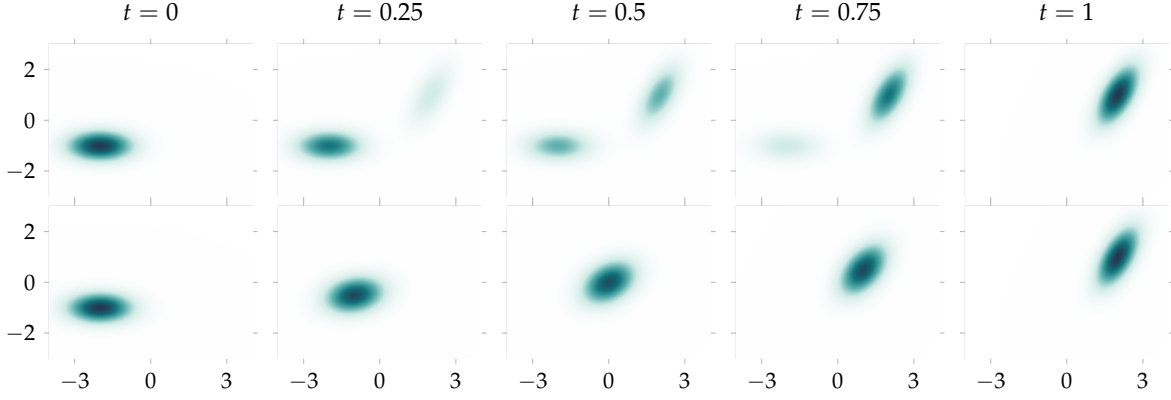


Figure 2: Mixture interpolation (**top**) versus displacement interpolation (**bottom**).

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 $\mu, \nu \in \mathcal{P}_{2,\text{ac}}(\mathbb{R}^d)$, one such interpolation is

$$\tilde{\mu}_t = (1-t)\mu + t\nu. \quad (1.19)$$

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

$$\mu_t = (T_t^{\mu\nu})_{\#}\mu := ((1-t)\text{id} + tT^{\mu\nu})_{\#}\mu. \quad (1.20)$$

For example, if μ, ν are both Gaussian measures, then one can check that (1.20) and (1.19) 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. These two interpolations are visualized in Figure 2.

1.6 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.

1.6.1 Closed-form expression of maps

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

$$F_{\rho}(t) = \int_{-\infty}^t d\rho(x) = \mathbb{P}_{\rho}(X \leq t), \quad (1.21)$$

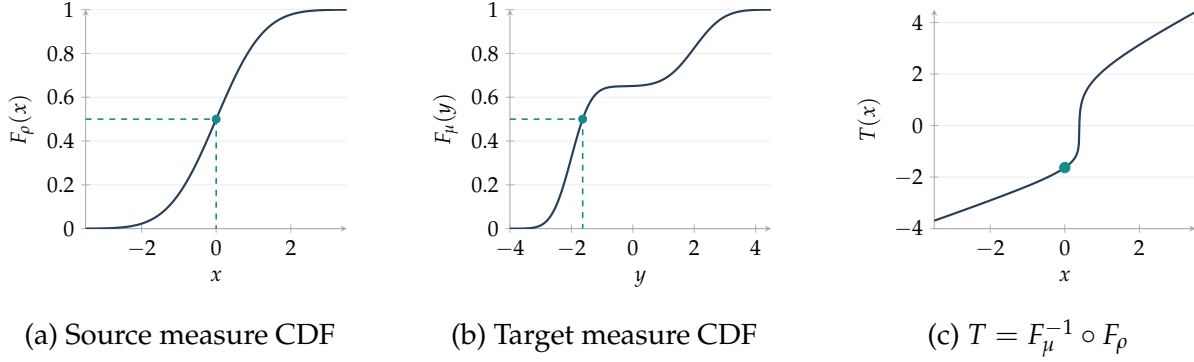


Figure 3: Illustration of the univariate OT map between a Gaussian and Gaussian mixture.

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

$$F_\rho^{-1}(u) := \inf\{x \in \mathbb{R} : F_\rho(x) \geq u\}, \quad (1.22)$$

where $F_\rho^{-1} : [0, 1] \rightarrow \mathbb{R} \cup \{+\infty\}$. 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; see Figure 3 for an illustration.

Theorem 1.15 (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*

$$T^{\rho \rightarrow \mu} = F_\mu^{-1} \circ F_\rho. \quad (1.23)$$

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 $X \sim \rho$, $F_\rho(X) \sim \text{Unif}([0, 1])$,
- for $U \sim \text{Unif}([0, 1])$, 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. $\square$

Exercise 1.5. Fill in the remaining details of Theorem 1.15.

1.6.2 Isometry property

From Theorem 1.15, we know that the cost of transport is

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

where the second equality is nothing but a change-of-variables. This is just to say that, in 1D, the Wasserstein space with distance W_2 embeds isometrically into $L^2([0, 1])$. 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.

1.6.3 Sliced Wasserstein distances

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

Definition 1.16 (Sliced Wasserstein distance). For $\rho, \mu \in \mathcal{P}_2(\mathbb{R}^d)$, the *sliced Wasserstein distance* is

$$\frac{1}{2} \text{SW}_2^2(\rho, \mu) := \int_{\mathbb{S}^{d-1}} \frac{1}{2} W_2^2(\rho_\theta, \mu_\theta) \, d\sigma_{d-1}(\theta), \quad (1.24)$$

where σ_{d-1} is the uniform probability measure on $\mathbb{S}^{d-1}$. Equivalently, by the 1D isometry (Section 1.6.2),

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

Proposition 1.17. SW_2 is a metric on $\mathcal{P}_2(\mathbb{R}^d)$.

1.7 On the stability of optimal transport maps

Note that a trivial coupling argument always implies that

$$\frac{1}{2}W_2^2(\rho, \mu) \leq \frac{1}{2}\|T^{\gamma \rightarrow \rho} - T^{\gamma \rightarrow \mu}\|_{L^2(\gamma)}^2,$$

for any probability measure 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

$$\|T^{\gamma \rightarrow \rho} - T^{\gamma \rightarrow \mu}\|_{L^2(\gamma)} \leq C W_2^\alpha(\rho, \mu), \quad (1.25)$$

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, (1.25) is an equality with $C = \alpha = 1$.

1.7.1 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 $\rho = \text{Unif}(B(0, 1))$ 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

$$T_\theta(x) := T^{\rho \rightarrow \mu_\theta}(x) := e_\theta \text{sign}(\langle x, e_\theta \rangle) \quad (1.26)$$

The first step is to verify that

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

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:

$$\|T_\theta - T_0\|_{L^2(\rho)}^2 \geq \int_{D_\theta} \|e_\theta + e_0\|^2 d\rho(x) \geq 2\rho(D_\theta) = \frac{\theta}{\pi}, \quad (1.27)$$

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.

1.7.2 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 1.18. [Gig11] *Let γ have a density and compact support, and let $\rho, \mu \in \mathcal{P}_2(\mathbb{R}^d)$ both have compact support. If $T^{\gamma \rightarrow \rho}$ is L -Lipschitz, then*

$$\|T^{\gamma \rightarrow \rho} - T^{\gamma \rightarrow \mu}\|_{L^2(\gamma)} \leq \sqrt{2L \text{diam}(\text{supp}(\gamma))} W_1^{1/2}(\rho, \mu).$$

Theorem 1.19. [MBNWW24] *Let γ, ρ be such that $T^{\gamma \rightarrow \rho}$ is bi-Lipschitz i.e.,*

$$\ell I \preceq \nabla T^{\gamma \rightarrow \rho} \preceq LI.$$

Then for any $\mu \in \mathcal{P}_2(\mathbb{R}^d)$

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

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

Theorem 1.20. [Mér26] Suppose $\gamma \in \mathcal{P}_{\text{ac}}(\mathbb{R}^d)$ over a convex compact set, with density bounded above. Let ρ, μ be supported on a convex compact subset of $\mathbb{R}^d$. Then

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

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

1.8 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?

1.8.1 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

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

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 1.21. Let $\mathcal{F}$ be a class of real-valued functions over $\Omega \subseteq \mathbb{R}^d$, and suppose there

²We 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 $\|f_i - f\|_{L^\infty(\Omega)} \leq \tau$.

exists $D > 0$ such that $\|f\|_{L^\infty(\Omega)} \leq D$ for all $f \in \mathcal{F}$. Then

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

1.8.2 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*

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

We will prove the following theorem.

Theorem 1.22. Take $\rho, \mu \in \mathcal{P}_2(\mathbb{R}^d)$ with compact support in $B(0, R)$. Then

$$\mathbb{E} \left| \frac{1}{2}W_2^2(\rho_n, \mu_n) - \frac{1}{2}W_2^2(\rho, \mu) \right| \lesssim_d R^2 \begin{cases} n^{-1/2} & d < 4, \\ n^{-1/2} \log(n) & d = 4, \\ n^{-2/d} & d > 4. \end{cases} \quad (1.28)$$

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:

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

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 $\mathcal{S}^{\rho\mu_n}(\varphi^{\rho\mu_n}) \leq \mathcal{S}^{\rho\mu_n}(\varphi^{\rho\mu})$ for instance.

Then we can upper bound the estimation gap as follows

$$\mathbb{E} \left| \frac{1}{2}W_2^2(\rho, \mu_n) - \frac{1}{2}W_2^2(\rho, \mu) \right| \leq \frac{1}{2} \mathbb{E} \left| \int \|\cdot\|^2 d(\mu - \mu_n) \right| + \mathbb{E} \left| \mathcal{S}^{\rho\mu}(\varphi^{\rho\mu}) - \mathcal{S}^{\rho\mu_n}(\varphi^{\rho\mu_n}) \right|.$$

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

$$\mathbb{E} \left| \int \|\cdot\|^2 d(\mu - \mu_n) \right| \leq R^2 n^{-1/2},$$

which follows from Cauchy–Schwarz. (Indeed, we can bound this expectation by $\sqrt{\operatorname{Var}(\|Y\|^2)/n} \leq R^2/\sqrt{n}$.) For the second term, we require the following lemma, which relates the difference in semidual values by a suitable empirical process.

Lemma 1.23. Suppose $\rho, \mu \in \mathcal{P}(B(0, R))$. Then

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

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

Proof. Writing $a = \varphi^{\rho\mu}$ and $b = \varphi^{\rho\mu_n}$, define the gap

$$\Delta = \mathcal{S}^{\rho\mu}(a) - \mathcal{S}^{\rho\mu_n}(b).$$

By optimality of the minimizers a and b , it is easy to derive the “sandwich” bound

$$\int a^* d(\mu - \mu_n) \leq \Delta \leq \int b^* d(\mu - \mu_n),$$

which implies that

$$|\Delta| \leq \max \left\{ \left| \int a^* d(\mu - \mu_n) \right|, \left| \int b^* d(\mu - \mu_n) \right| \right\}.$$

Notice that on compact supports, we may choose conjugate pairs by taking successive conjugates restricted to $B(0, R)$; in particular, we can write

$$\psi(y) = \sup_{x \in B(0, R)} \{ \langle x, y \rangle - \varphi(x) \},$$

which implies that ψ is convex and R -Lipschitz, and, without loss of generality, may be centered. These properties define the function class Φ_R . This completes the proof, as

$$|\Delta| \leq \max \left\{ \left| \int a^* d(\mu - \mu_n) \right|, \left| \int b^* d(\mu - \mu_n) \right| \right\} \leq \sup_{\psi \in \Phi_R} \left| \int \psi d(\mu - \mu_n) \right|.$$

□

Our bound is now

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

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

Theorem 1.24 ([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 $\tau/R^2 < C_1$

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

Proof of Theorem 1.22. Rescaling the function class, we can apply Dudley's bound with $D = R^2$, resulting in

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

□

1.8.3 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 $\hat{T}_0$ 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

$$\mathbb{E} \|\hat{T} - T_0\|_{L^2(\rho)}^2 \lesssim n^{-\frac{2\alpha}{2\alpha-2+d}} \log^3(n) \vee n^{-1}, \quad (1.29)$$

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 1.25. *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 $\hat{T}_n$ such that*

$$\mathbb{E} \|\hat{T}_n - T_0\|_{L^2(\rho)}^2 \lesssim \begin{cases} n^{-1/4} & d < 4, \\ n^{-1/4} \sqrt{\log(n)} & d = 4, \\ n^{-1/d} & d > 4, \end{cases} \quad (1.30)$$

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

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

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

We conclude by applying Theorem 1.22. □

1.9 Algorithms

1.9.1 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 1.26. *For $X_1, \dots, X_n \sim \rho$ and $Y_1, \dots, Y_m \sim \mu$, the finite dimensional Kantorovich problem (1.4) for the empirical measures ρ_n and μ_m is the following linear program*

$$\frac{1}{2} W_2^2(\rho_n, \mu_m) = \min_{\Pi \in \mathbb{R}_+^{n \times m}} \langle C, \Pi \rangle \quad \text{s.t.} \quad \Pi \mathbf{1}_m = \frac{1}{n} \mathbf{1}_n, \Pi^\top \mathbf{1}_n = \frac{1}{m} \mathbf{1}_m, \quad (1.31)$$

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 1.27 (Kantorovich linear program runtime). *The discrete Kantorovich linear program (1.31), which has mn variables and $n + m$ equality constraints, can be solved in time $\tilde{O}((n + m)nm)$.*

1.9.2 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 1.28. *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*

$$\frac{1}{2} W_2^2(\rho_n, \mu_n) = \frac{1}{2n} \sum_{i=1}^n (X_{(i)} - Y_{(i)})^2. \quad (1.32)$$

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 1.29 (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 1.28. Letting c be the quadratic cost, then Lemma 1.29 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. $\square$

1.9.3 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

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

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

1. sample $\theta_1, \dots, \theta_L \sim \text{Unif}(\mathbb{S}^{d-1})$ 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 $\frac{1}{2} \widehat{\text{SW}}_2^2 = \frac{1}{L} \sum_{l=1}^L \frac{1}{2} W_2^2((\rho_n)_{\theta_l}, (\mu_n)_{\theta_l})$.

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

$$\frac{1}{2} \widehat{\text{SW}}_2^2 \rightarrow \frac{1}{2} \text{SW}_2^2(\rho_n, \mu_n) \quad \text{as } L \rightarrow \infty,$$

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

1.9.4 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 i th row of Π_0 has total mass $1/n$, the quantities

$$\hat{T}_{\Pi_0}(X_i) := n \sum_{j=1}^m \Pi_{0,ij} Y_j \tag{1.33}$$

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 $\hat{T}_{\Pi_0}(X_i)$ is the corresponding conditional mean.

Proposition 1.30. *Let Π be any feasible transport plan for (1.31). For every $i \in [n]$,*

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

Thus $\hat{T}_{\Pi}(X_i) = \sum_{j=1}^m q_{ij}Y_j$ 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

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

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:

$$i_n(x) \in \operatorname{argmin}_{i \in [n]} \|x - X_i\|, \quad \hat{T}_{\text{NN}}(x) := \hat{T}_{\Pi_0}(X_{i_n(x)}). \quad (1.34)$$

Thus, $\hat{T}_{\text{NN}}$ 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

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

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

$$\hat{\theta} \in \operatorname{argmin}_{\theta \in \Theta} \hat{L}(\theta) := \operatorname{argmin}_{\theta \in \Theta} \left\{ \frac{1}{n} \sum_{i=1}^n \varphi_{\theta}(X_i) + \frac{1}{m} \sum_{j=1}^m \varphi_{\theta}^*(Y_j) \right\}. \quad (1.35)$$

If one could solve this optimization problem, the resulting estimator would be the parametric Brenier map estimator

$$\hat{T}_{\text{net}}(x) := \nabla \varphi_{\hat{\theta}}(x). \quad (1.36)$$

One can choose an expressive neural network to parametrize φ_{θ} , with the goal of updating the parameters θ via simple (stochastic) gradient descent methods. Here, computing $\nabla_{\theta} \hat{L}$ involves differentiating through the conjugate φ_{θ}^* , which is itself defined as a maximization problem. To deal with this, we will make use of the following theorem; see [Amo22].

Theorem 1.31 (Envelope theorem). *Let $U \subset \mathbb{R}^p$ be open and let $F : U \times \mathbb{R}^d \rightarrow \mathbb{R}$ be twice continuously differentiable. Suppose that $F(\theta, \cdot)$ is uniformly strongly concave, and define*

$$V(\theta) := \max_{z \in \mathbb{R}^d} F(\theta, z), \quad z_\theta := \arg \max_{z \in \mathbb{R}^d} F(\theta, z).$$

Then z_θ is unique, V is continuously differentiable, and

$$\nabla_\theta V(\theta) = \nabla_\theta F(\theta, z_\theta).$$

We apply the theorem assuming that φ_θ satisfies the required regularity and strong convexity conditions. For example, given an input-convex network [AXK17], one can add $\frac{\lambda}{2} \|x\|^2$ (for $\lambda > 0$) to the output to enforce strong convexity. The first-order condition is then

$$z_\theta(y) := \arg \max_{z \in \mathbb{R}^d} \{ \langle y, z \rangle - \varphi_\theta(z) \}, \quad \nabla \varphi_\theta(z_\theta(y)) = y.$$

Now if we differentiate the loss, we obtain

$$\nabla_\theta \hat{L}(\theta) = \frac{1}{n} \sum_{i=1}^n \nabla_\theta \varphi_\theta(X_i) - \frac{1}{m} \sum_{j=1}^m (\nabla_\theta \varphi_\theta)(z_\theta(Y_j)),$$

which allows us to perform gradient descent on the parameters

$$\theta_{k+1} = \theta_k - \eta_k \nabla_\theta \hat{L}(\theta_k), \quad \eta_k > 0.$$

Computing conjugates for high-dimensional functions can be costly and undesirable. The main idea of [Amo22] is to accelerate these steps via *amortization*: train a separate neural network to predict $z_\theta(y)$ from y , and use the prediction to warm-start the numerical solver that performs the conjugate operation.

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

$$\begin{aligned} a &= \frac{1}{n} \sum_{i=1}^n Z_i, & A &= \frac{1}{n} \sum_{i=1}^n (Z_i - a)(Z_i - a)^\top, \\ \hat{a} &= \frac{1}{m} \sum_{j=1}^m \hat{Z}_j, & \hat{A} &= \frac{1}{m} \sum_{j=1}^m (\hat{Z}_j - \hat{a})(\hat{Z}_j - \hat{a})^\top, \end{aligned}$$

and compare the empirical distributions via

$$\begin{aligned} \frac{1}{2} \text{FID}(\mathcal{Z}, \hat{\mathcal{Z}}) &:= \frac{1}{2} W_2^2(\mathcal{N}(a, A), \mathcal{N}(\hat{a}, \hat{A})) \\ &= \frac{1}{2} \|a - \hat{a}\|_2^2 + \frac{1}{2} \text{Tr}\left(A + \hat{A} - 2(A^{1/2} \hat{A} A^{1/2})^{1/2}\right). \end{aligned}$$

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

2 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.

2.1 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

$$\text{KL}(P \parallel Q) := \int \log\left(\frac{dP}{dQ}\right) dP, \quad (2.1)$$

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 2.1 (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

$$\text{KL}(P \parallel Q) = \frac{1}{2} \left[\log \frac{\det B}{\det A} - d + \text{tr}(B^{-1}A) + (a - b)^\top B^{-1}(a - b) \right]. \quad (2.2)$$

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

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

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

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

The middle expectation is d , while

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

which yields (2.2).

We now define the *entropic* optimal transport problem, which takes (1.4) and incorporates the KL divergence as a regularizer.

Definition 2.2 (Entropic optimal transport). For $\varepsilon > 0$, the Kantorovich problem with

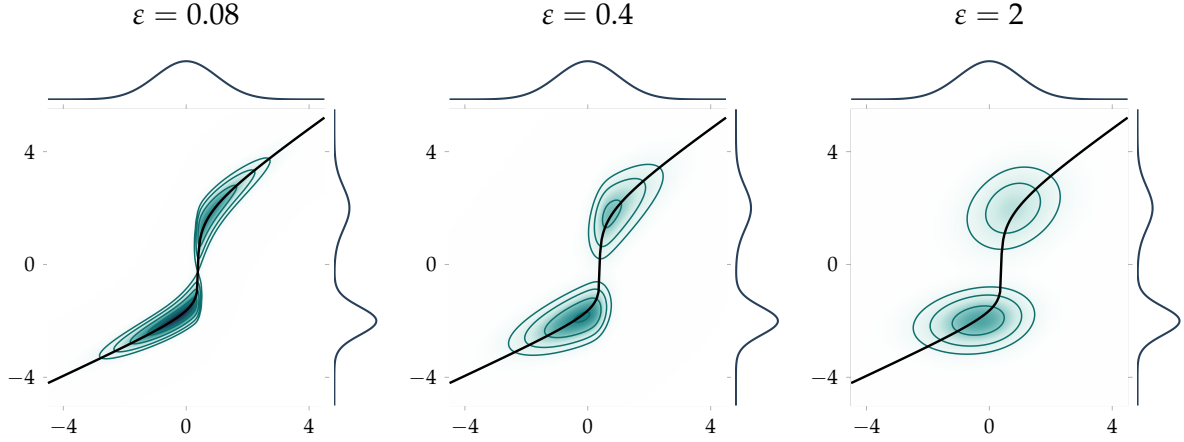


Figure 4: Entropic couplings between a Gaussian and Gaussian mixture with various regularization strengths, and the solid black curve is the optimal transport map.

entropic regularization is

$$\text{OT}_\varepsilon(\rho, \mu) := \inf_{\pi \in \Gamma(\rho, \mu)} \iint \frac{1}{2} \|x - y\|^2 d\pi(x, y) + \varepsilon \text{KL}(\pi \| \rho \otimes \mu). \quad (2.3)$$

Obviously if $\varepsilon = 0$, we recover (1.4), 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 2.3. *Let $\rho, \mu \in \mathcal{P}_2(\mathbb{R}^d)$. For any $\varepsilon > 0$, there exists a unique minimizer to (2.3), 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. As $\varepsilon \rightarrow \infty$, notice that the cost plays less of a role in the objective, and the optimal plan will be closer to the product coupling. This phenomenon is illustrated in Figure 4 for various regularization strengths.

2.2 Dual formulation and consequences

As before, to obtain the dual for (2.3), we mimic the computations of Proposition 1.8. This leads to

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

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

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

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

$$h^*(b) = \sup_a \{ab - h(a)\} = e^b, \quad (2.4)$$

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

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

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

Proposition 2.4. *For $\varepsilon > 0$, suppose ρ and μ have finite second moments. Then we have the following dual formulation of (2.3)*

$$\text{OT}_\varepsilon(\rho, \mu) := \sup_{\substack{f \in L^1(\rho) \\ g \in L^1(\mu)}} \int f \, d\rho + \int g \, d\mu - \varepsilon \iint \left(e^{(f(x) + g(y) - \frac{1}{2} \|x - y\|^2) / \varepsilon} - 1 \right) d\rho(x) d\mu(y). \quad (2.5)$$

The objective function (2.5) should be reminiscent of (1.5), with one glaring exception: the *hard* constraint is now an *unconstrained* optimization problem. Convince yourself that the $\varepsilon \rightarrow 0$ limit morally recovers the hard constraint from before.

Definition 2.5 (Optimal entropic Kantorovich potentials). We write the maximizers of (2.5) as $(f_\varepsilon, g_\varepsilon)$, which are called (optimal) entropic Kantorovich potentials.

2.2.1 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]. The proof is a result of strong duality, and so it is omitted.

Proposition 2.6. For $\rho, \mu \in \mathcal{P}_2(\mathbb{R}^d)$, the optimal entropic plan is explicitly written as

$$d\pi_\varepsilon(x, y) := \exp\left(\frac{f_\varepsilon(x) + g_\varepsilon(y) - \frac{1}{2}\|x - y\|^2}{\varepsilon}\right) d\rho(x) d\mu(y). \quad (2.6)$$

2.2.2 An alternative dual and extending the potentials

We now arrive at an alternative dual formulation. Fix $y \in \text{supp}(\mu)$, $f \in L^1(\rho)$ and maximize (2.5) only in “ g ”; this yields

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

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 $\text{OT}_\varepsilon(\rho, \mu)$, written

$$\text{OT}_\varepsilon(\rho, \mu) := \sup_{f \in L^1(\rho)} \int f d\rho + \int f^{(c, \varepsilon)} d\mu. \quad (2.7)$$

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 $y \in \mathbb{R}^d$:

$$f_\varepsilon(x) = -\varepsilon \log \int e^{(g_\varepsilon(y) - \frac{1}{2}\|x - y\|^2)/\varepsilon} d\mu(y) \quad (x \in \mathbb{R}^d), \quad (2.8)$$

$$g_\varepsilon(y) = -\varepsilon \log \int e^{(f_\varepsilon(x) - \frac{1}{2}\|x - y\|^2)/\varepsilon} d\rho(x) \quad (y \in \mathbb{R}^d). \quad (2.9)$$

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

2.3 Sinkhorn's algorithm

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

Throughout this section, the source and target measures are denoted by ρ and μ , respectively.

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

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

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

Sinkhorn's algorithm proposes to *iteratively* fit the marginals of π_ε until you arrive at a fixed point, which *must* be the optimal plan. We will see later this will precisely correspond to either coordinate ascent or mirror descent updates, depending on the point of view.

To this end, we use two indices to indicate the updated plan and potentials, written

$$d\pi^{(i,j)}(x, y) := \exp\left(\frac{f^{(i)}(x) + g^{(j)}(y) - c(x, y)}{\varepsilon}\right) d\rho(x) d\mu(y),$$

for a suitable arbitrary cost. The first index records updates for the f potential, and the second belongs to the g potential. We initialize $f^{(0)} = g^{(0)} = 0$. Then

$$d\pi^{(0,0)}(x, y) = e^{-c(x,y)/\varepsilon} d\rho(x) d\mu(y).$$

While this initial joint measure is not a probability measure, every subsequent iteration will be.

At iteration $k \geq 0$, we first keep $g^{(k)}$ fixed and choose $f^{(k+1)}$ such that $\pi^{(k+1,k)}$ has left-marginal ρ i.e.,

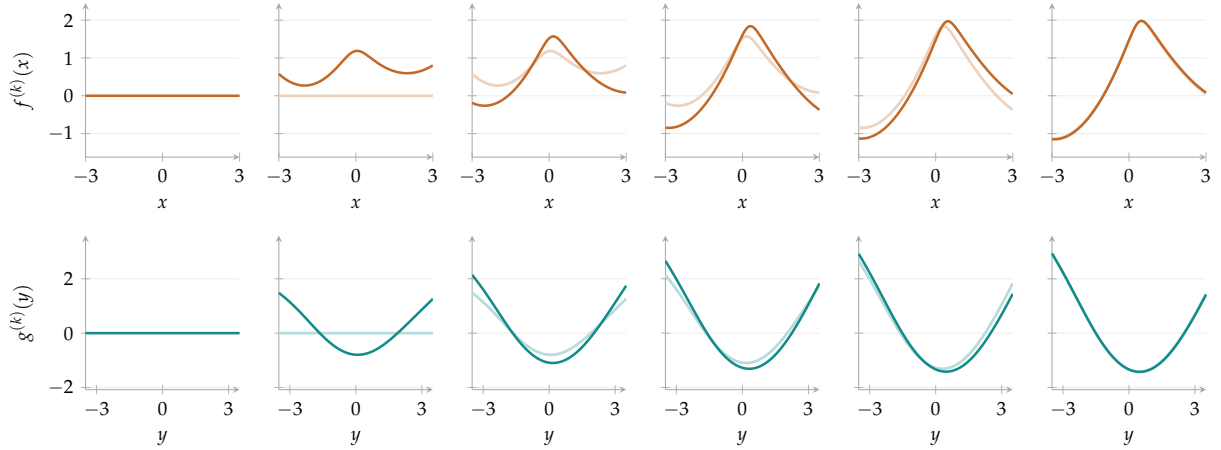
$$e^{f^{(k+1)}(x)/\varepsilon} \int e^{(g^{(k)}(y) - c(x,y))/\varepsilon} d\mu(y) = 1.$$

Rearranging implies

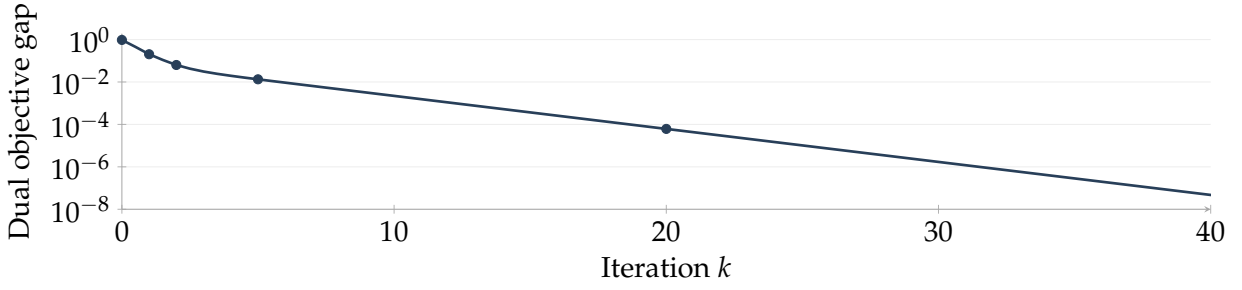
$$f^{(k+1)}(x) := -\varepsilon \log \int e^{(g^{(k)}(y) - c(x,y))/\varepsilon} d\mu(y). \quad (2.10)$$

It is easy to verify that $\pi^{(k+1,k)}$ does not satisfy the right-marginal constraint. We can try to fix this by now keeping $f^{(k+1)}$ fixed and have the coupling now satisfy the other marginal constraint, resulting in the update

$$g^{(k+1)}(y) := -\varepsilon \log \int e^{(f^{(k+1)}(x) - c(x,y))/\varepsilon} d\rho(x). \quad (2.11)$$



(a) Dual iterates $f^{(k)}$ (**top**) and $g^{(k)}$ (**bottom**) of Sinkhorn's algorithm at $k = 0, 1, 2, 5, 20, 121$.



(b) Dual objective gap $\mathcal{D}_\varepsilon^{(k)} - \mathcal{D}_\varepsilon^{\rho\mu}$ of Sinkhorn's algorithm.

Figure 5: Depiction of Sinkhorn's algorithm on the example from Figure 4 with $\varepsilon = 0.4$.

Now $\pi_\varepsilon^{(k+1,k+1)}$ has second marginal μ , while its first marginal need not equal ρ . The iterates then read

$$\pi^{(k+1,k)} \in \Gamma(\rho, *), \quad \pi^{(k+1,k+1)} \in \Gamma(*, \mu), \quad \pi^{(k+2,k+1)} \in \Gamma(\rho, *), \dots$$

Figure 5 illustrates the updates at the level of the potentials between a Gaussian and a Gaussian mixture.

Sinkhorn's algorithm repeats these updates until convergence, at which point the coupling matches both marginal constraints. This fixed-point perspective of the dual variables can also be seen from (2.8) and (2.9).

2.3.1 Analysis via coordinate ascent

We now provide a basic runtime analysis of Sinkhorn's algorithm based on the block coordinate-ascent interpretation. We omit the subscript ε , and define

$$\rho^{(i,j)}(x) := \int \pi^{(i,j)}(x, y) dy, \quad \mu^{(i,j)}(y) := \int \pi^{(i,j)}(x, y) dx.$$

In particular, the updates imply that $\rho^{(k+1,k)} = \rho$, $\mu^{(k+1,k+1)} = \mu$. We additionally write the dual objective as

$$\mathcal{D}_\varepsilon^{\rho\mu}(f, g) := \int f \, d\rho + \int g \, d\mu - \varepsilon \iint (e^{(f(x)+g(y)-\frac{1}{2}\|x-y\|^2)/\varepsilon} - 1) \, d\rho(x) \, d\mu(y). \quad (2.12)$$

Finally, we require Pinsker's inequality, which bounds the total variation distance by the KL-divergence between two measures.

Proposition 2.7 (Pinsker's inequality). *For any $P, Q \in \mathcal{P}(\mathbb{R}^d)$,*

$$\text{TV}(P, Q) := \frac{1}{2} \|P - Q\|_{L^1} \leq \sqrt{\frac{1}{2} \text{KL}(P \| Q)}. \quad (2.13)$$

We will prove the following theorem.

Theorem 2.8. *Suppose $\rho, \mu \in \mathcal{P}(\mathbb{R}^d)$ have compact support, and fix $\varepsilon > 0$. After $k \geq 2$ complete iteration cycles, the iterates of Sinkhorn's algorithm satisfy*

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

where $c_\infty := \sup_{x \in \text{supp}(\rho), y \in \text{supp}(\mu)} c(x, y)$.

Proof. At the end of cycle $k \geq 1$, $\pi^{(k,k)}$ has second marginal μ . The next update starts the next cycle, which first updates f , then g . From (2.10) and (2.11), we can write the updates as

$$\begin{aligned} f^{(k+1)} &= f^{(k)} - \varepsilon \log \left(\frac{d\rho^{(k,k)}}{d\rho} \right), \\ g^{(k+1)} &= g^{(k)} - \varepsilon \log \left(\frac{d\mu^{(k+1,k)}}{d\mu} \right). \end{aligned} \quad (2.14)$$

By (2.14) and Pinsker's inequality,

$$\begin{aligned} \mathcal{D}_\varepsilon^{\rho\mu}(f^{(k+1)}, g^{(k)}) - \mathcal{D}_\varepsilon^{\rho\mu}(f^{(k)}, g^{(k)}) &= \int (f^{(k+1)} - f^{(k)}) \, d\rho \\ &= \varepsilon \text{KL}(\rho \| \rho^{(k,k)}) \\ &\geq 2\varepsilon \text{TV}^2(\rho, \rho^{(k,k)}). \end{aligned}$$

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

Lemma 2.9. *Under the assumptions of Theorem 2.8, for any $k \geq 1$,*

$$\Delta_k := \sup_{f, g} \mathcal{D}_\varepsilon^{\rho\mu}(f, g) - \mathcal{D}_\varepsilon^{\rho\mu}(f^{(k)}, g^{(k)}) \leq 2c_\infty (\text{TV}(\rho, \rho^{(k,k)}) + \text{TV}(\mu, \mu^{(k,k)})).$$

We are now ready to complete the proof. Since $\mu^{(k,k)} = \mu$, Lemma 2.9 and the improvement from the first marginal update yield

$$\Delta_k - \Delta_{k+1} \geq 2\varepsilon \text{TV}^2(\rho, \rho^{(k,k)}) \geq \frac{\varepsilon \Delta_k^2}{2c_\infty^2}.$$

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

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

We conclude by summing from $k = 1$ to $K - 1$, which gives $\Delta_K \leq 2c_\infty^2 / (\varepsilon(K - 1))$. $\square$

2.3.2 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 (2.3), up to an additive constant, as

$$\text{KL}^*(\rho, \mu, R) := \inf_{\pi \in \Gamma(\rho, \mu)} \varepsilon \text{KL}(\pi \| R), \quad (2.15)$$

where

$$\text{d}R(x, y) = Z^{-1} e^{-c(x, y)/\varepsilon} \text{d}\rho(x) \text{d}\mu(y), \quad Z = \iint e^{-c(x, y)/\varepsilon} \text{d}\rho(x) \text{d}\mu(y).$$

In other words, for any coupling π we have that

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

We will prove the following. For the Bregman analysis, we will be interested in the Sinkhorn iterates just after fitting the first marginal, and so we abbreviate

$$\pi^{(k)} := \pi_\varepsilon^{(k+1, k)}, \quad \mu^{(k)} := \mu^{(k+1, k)}.$$

Thus $\pi^{(0)}$ is the first normalized iterate $\pi_\varepsilon^{(1, 0)}$. A step from $\pi^{(k)}$ to $\pi^{(k+1)}$ fits the second marginal and then the first.

Theorem 2.10. Suppose $g^{(0)} = 0$ and $\pi^{(0)} = (\text{d}\rho / \text{d}R_X)R$, 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 ρ and Y -marginal $\mu^{(k)}$). Then, for $k \geq 1$,

$$\text{KL}(\mu^{(k)} \| \mu) \leq \frac{\text{KL}^*(\rho, \mu, R)}{\varepsilon k}.$$

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

$$\text{KL}(\mu^{(k)} \parallel \mu) \leq \frac{m_2^2(\rho) + m_2^2(\mu)}{\varepsilon k}.$$

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 $F : \mathbb{R}^d \rightarrow \mathbb{R}$ by limiting the movement of the iterates using a *Bregman divergence* B_G , resulting in the iterates

$$x^{(k+1)} = \underset{x}{\operatorname{argmin}} F(x^{(k)}) + \langle F'(x^{(k)}), x - x^{(k)} \rangle + B_G(x \parallel x^{(k)}), \quad (2.16)$$

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

$$B_G(x_2 \parallel x_1) := G(x_2) - G(x_1) - \langle G'(x_1), x_2 - x_1 \rangle.$$

It is easy to see that the optimality conditions for (2.16) result in

$$G'(x^{(k+1)}) - G'(x^{(k)}) = -F'(x^{(k)}). \quad (2.17)$$

Letting $y^\bullet = G'(x^\bullet)$, we can express (2.17) as

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

From here, we might hope to draw immediate parallels to (2.14) by clever choices of the functions F and G that act on probability measures. To start, for a fixed measure $\mu \in \mathcal{P}(\mathbb{R}^d)$, define $\tilde{\mu} \mapsto \mathcal{H}_\mu(\tilde{\mu}) := \text{KL}(\tilde{\mu} \parallel \mu)$. By computing the first variation of this functional, we see that

$$\tilde{\mu} \mapsto \mathcal{H}'_\mu(\tilde{\mu}) = \log(\tilde{\mu}/\mu),$$

We can write the Sinkhorn update as

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

In order to draw parallels to Bregman descent, consider the functional

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

We then compute

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

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

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

has X -marginal ρ by construction. Thus,

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

As convex conjugation reverses derivatives,³ we can finally write the Sinkhorn update as

$$(\mathcal{G}^*)'(\mu^{(k+1)}) - (\mathcal{G}^*)'(\mu^{(k)}) = -\varepsilon \mathcal{H}'_\mu(\mu^{(k)}), \quad (2.18)$$

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

Theorem 2.11. *Let F, G be convex differentiable functions, and suppose that the Bregman divergence of F is bounded by that of G , i.e., for $x, \tilde{x}$,*

$$B_F(\tilde{x} \| x) \leq B_G(\tilde{x} \| x). \quad (2.19)$$

Then, for all $k \geq 1$,

$$F(x^{(k)}) \leq F(x) + \frac{B_G(x \| x^{(0)})}{k}.$$

In order to use Theorem 2.11 to prove Theorem 2.10, we will need the following lemma.

Lemma 2.12. *Let g_1, g_2 be potentials, and define for $i \in \{1, 2\}$*

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

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

$$B_{\mathcal{G}}(g_1 \| g_2) = B_{\mathcal{G}^*}(\nu_2 \| \nu_1) = \varepsilon \text{KL}(\pi_2 \| \pi_1).$$

Proof of Theorem 2.10. Let ν_1, ν_2 be any two probability measures. First, note the following equality holds true

$$B_{\mathcal{H}_\mu}(\nu_2 \| \nu_1) = B_{\mathcal{H}}(\nu_2 \| \nu_1) = \text{KL}(\nu_2 \| \nu_1).$$

If we now choose ν_1, π_1 and ν_2, π_2 as in Lemma 2.12, by an application of Jensen's inequality, it follows that

$$\varepsilon B_{\mathcal{H}_\mu}(\nu_2 \| \nu_1) = \varepsilon \text{KL}(\nu_2 \| \nu_1) \leq \varepsilon \text{KL}(\pi_2 \| \pi_1) = B_{\mathcal{G}^*}(\nu_2 \| \nu_1).$$

³Verify that this functional is indeed convex!

To invoke Theorem 2.11, we choose $F \equiv \varepsilon \mathcal{H}_\mu$ and $G \equiv \mathcal{G}^*$ which implies

$$\varepsilon \mathcal{H}_\mu(\mu^{(k)}) = \varepsilon \text{KL}(\mu^{(k)} \parallel \mu) \leq \frac{\text{B}_{\mathcal{G}^*}(\mu \parallel \mu^{(0)})}{k} = \frac{\varepsilon \text{KL}(\pi^{\rho\mu} \parallel \pi^{(0)})}{k}.$$

Since $d\pi^{(0)} / dR = d\rho / dR_X$, we have that

$$\varepsilon \text{KL}(\pi^{\rho\mu} \parallel \pi^{(0)}) = \text{KL}^*(\rho, \mu, R) - \varepsilon \text{KL}(\rho \parallel R_X) \leq \text{KL}^*(\rho, \mu, R),$$

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

$$\begin{aligned} \text{KL}^*(\rho, \mu, R) &\leq \varepsilon \text{KL}(\rho \otimes \mu \parallel R) \\ &= \iint \frac{1}{2} \|x - y\|^2 d\rho(x) d\mu(y) + \varepsilon \log Z \\ &\leq m_2^2(\rho) + m_2^2(\mu). \end{aligned}$$

□

2.4 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 2.13. For $\varepsilon > 0$, let $(f_\varepsilon, g_\varepsilon)$ be the entropic Kantorovich potentials that maximize (2.5), extended to all of $\mathbb{R}^d$ as in (2.8) and (2.9). The *entropic Brenier potentials* are then defined as

$$(\varphi_\varepsilon, \psi_\varepsilon) = (\tfrac{1}{2} \|\cdot\|^2 - f_\varepsilon, \tfrac{1}{2} \|\cdot\|^2 - g_\varepsilon) \quad (2.20)$$

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

$$T_\varepsilon(x) := \int y d\pi_\varepsilon^x(y) = \mathbb{E}_{\pi_\varepsilon}[Y \mid X = x]. \quad (2.21)$$

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

However, a priori, (2.21) 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 (2.8), resulting in the following expression

Proposition 2.15. *For $\varepsilon > 0$, the entropic map is equivalent to*

$$T_\varepsilon = \nabla \varphi_\varepsilon. \quad (2.22)$$

Proof. We proceed by direct computation, using the relationship (2.8):

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

□

Exercise 2.1. Prove that $\nabla^2 \varphi_\varepsilon(x) = \varepsilon^{-1} \text{Cov}_{\pi_\varepsilon}(Y|X = x)$, and conclude that φ_ε is always convex. Similarly, show $\nabla^2 \psi_\varepsilon(y) = \varepsilon^{-1} \text{Cov}_{\pi_\varepsilon}(X|Y = y)$ and conclude that ψ_ε is also convex, though φ_ε and ψ_ε are *not* convex conjugates of one another.

2.5 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.)

2.5.1 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 $\Lambda_\varepsilon := (2\pi\varepsilon)^{d/2}$, $\mathcal{H}(\rho)$ denote the *entropy* of ρ , and let $I_0(\rho, \mu)$ denote the *integrated relative Fisher information along Wasserstein geodesics* i.e.,

$$\mathcal{H}(\rho) := \int \log \rho d\rho \quad \text{and} \quad I_0(\rho, \mu) = \int_0^1 \mathbb{E}_{\mu_t} \|\nabla \log \mu_t\|^2 dt,$$

where μ_t is the optimal transport path from ρ to μ (recall (1.20)). The following fundamental approximation result is due to [CRLV+20; CT21].

Theorem 2.16 (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 $\varepsilon \rightarrow 0$,*

$$\text{OT}_\varepsilon(\rho, \mu) = \frac{1}{2}W_2^2(\rho, \mu) - \varepsilon \log \Lambda_\varepsilon - \frac{\varepsilon}{2}(\mathcal{H}(\rho) + \mathcal{H}(\mu)) + O(\varepsilon^2 I_0(\rho, \mu)). \quad (2.23)$$

A useful quantitative version is

$$\text{OT}_\varepsilon(\rho, \mu) - \frac{1}{2}W_2^2(\rho, \mu) + \varepsilon \log \Lambda_\varepsilon + \frac{\varepsilon}{2}(\mathcal{H}(\rho) + \mathcal{H}(\mu)) \leq \frac{\varepsilon^2}{8}I_0(\rho, \mu). \quad (2.24)$$

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 (2.24) is quite easy to prove once one has access to the machinery from Section 7.

2.5.2 Small regularization regime of costs: The Gaussian case

While we did not see the proof of Theorem 2.16, 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 2.17. *For $\rho = \mathcal{N}(0, A)$ and $\mu = \mathcal{N}(b, B)$, the optimal entropic plan is given by the following joint probability measure*

$$\pi_\varepsilon = \mathcal{N}\left(\begin{pmatrix} 0 \\ b \end{pmatrix}, \begin{pmatrix} A & \Sigma_\varepsilon \\ \Sigma_\varepsilon^\top & B \end{pmatrix}\right), \quad (2.25)$$

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

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

Then the entropic cost admits the exact expression

$$\text{OT}_\varepsilon(\rho, \mu) = \frac{1}{2}\|b\|^2 + \frac{1}{2}\text{Tr}(A + B) - \text{Tr}(D_\varepsilon) + \frac{d\varepsilon}{2} + \frac{\varepsilon}{2}\log\det\left(\varepsilon^{-1}(D_\varepsilon + \frac{\varepsilon}{2}I)\right).$$

Consequently, as $\varepsilon \rightarrow 0$,

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

where $I_0(\rho, \mu) = \text{Tr}(M^{-1/2})$ in the Gaussian case. Thus the second-order term in (2.24) is sharp for Gaussian marginals.

2.5.3 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

$$T_\varepsilon(x) = A^{-1/2}[(A^{1/2}BA^{1/2} + \frac{\varepsilon^2}{4}I)^{1/2} - \frac{\varepsilon}{2}I]A^{-1/2}x + b. \quad (2.26)$$

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

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

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

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

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.

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

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

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

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

2.5.4 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 2.19 (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

$$\|T_\varepsilon - T_0\|_{L^2(\rho)}^2 \lesssim \varepsilon^2 I_0(\rho, \mu). \quad (2.27)$$

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

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

We now describe the main ideas of the proof.

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

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

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

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

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

$$q_\varepsilon^x(y) := \frac{1}{Z_\varepsilon(x) \Lambda_\varepsilon} \exp\left(-\frac{1}{\varepsilon} D[y \mid x^*]\right) \quad (2.28)$$

behaves like

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

The premise of the proof is to then show that

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

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

Lemma 2.20 (Gaussian localization). *Under the assumptions of Theorem 2.19, 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 ,*

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

and the normalizing constant satisfies

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

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

$$\iint \exp(h(x)^\top (y - T_0(x)) - a\|h(x)\|^2) q_\varepsilon^x(y) \, dy \, d\rho(x) \leq \int \exp(\|\Delta_\varepsilon(x)\|^2 / (4\varepsilon)) \, d\rho(x).$$

Another proof ingredient is the enlarged dual representation

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

One then tests this dual with

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

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

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

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

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

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

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

then implies that

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

where we additionally define

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

Recall that Lemma 2.20 gives $|r_\varepsilon| \lesssim \varepsilon$. Combining the preceding display with (2.24), and using that q_ε^x integrates to one gives

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

Finally, we use the last results in Lemma 2.20 to imply that

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

uniformly in h . Hence

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

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.

2.5.5 Sinkhorn divergence and debiased entropic maps

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

Proposition 2.21 (Entropic self-transport). *For the self-transport problem $\text{OT}_\varepsilon(\rho, \rho)$, there is a single self-potential α_ε satisfying*

$$\begin{aligned} \text{OT}_\varepsilon(\rho, \rho) &= \sup_{\alpha \in L^1(\rho)} \left\{ 2 \int \alpha d\rho - \varepsilon \iint e^{(\alpha(x) + \alpha(y) - \frac{1}{2} \|x-y\|^2)/\varepsilon} d\rho(x) d\rho(y) + \varepsilon \right\} \\ &= 2 \int \alpha_\varepsilon d\rho. \end{aligned} \tag{2.29}$$

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

$$\text{OT}_\varepsilon(\rho, \rho) + \varepsilon \log \Lambda_\varepsilon + \varepsilon \mathcal{H}(\rho) \leq \frac{\varepsilon^2}{8} I_0(\rho). \tag{2.30}$$

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

Definition 2.22 (Sinkhorn divergence). For $\rho, \mu \in \mathcal{P}_2(\mathbb{R}^d)$ and $\varepsilon > 0$, the *Sinkhorn divergence* is defined as

$$S_\varepsilon(\rho, \mu) := \text{OT}_\varepsilon(\rho, \mu) - \frac{1}{2}[\text{OT}_\varepsilon(\rho, \rho) + \text{OT}_\varepsilon(\mu, \mu)]. \quad (2.31)$$

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 2.23. *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: $S_\varepsilon(\rho, \mu) \rightarrow \frac{1}{2}W_2^2(\rho, \mu)$ as $\varepsilon \rightarrow 0^+$.

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

$$\bar{T}_\varepsilon^{\rho\mu}(x) := T_\varepsilon^{\rho\mu}(x) - T_\varepsilon^{\rho\rho}(x) + x. \quad (2.32)$$

2.6 Lipschitz properties of (entropic) transport maps

For many theoretical results in, say, statistical optimal transport, uniform upper and lower bounds on the spectrum of $\nabla^2 \varphi^{\rho\mu}$, the Hessian of the Brenier potential, are crucial. In other words, we often require the existence of two constants $\ell, L > 0$ such that

$$\ell I \preceq \nabla^2 \varphi^{\rho\mu}(x) \preceq LI,$$

for all $x \in \mathbb{R}^d$. Though, apart from assuming these bounds hold, a natural question is to know when this kind of assumption is known with explicit constants. This next result provides sufficient conditions for this behavior.

Theorem 2.25 (Caffarelli's contraction theorem). *Let $\rho \propto e^{-V}$ and $\mu \propto e^{-W}$ with full support on $\mathbb{R}^d$, with V and W smooth. Further suppose that $\nabla^2 V(x) \preceq \beta_V I$ and $\nabla^2 W(x) \succeq \alpha_W I$ with $\beta_V, \alpha_W > 0$. Then*

$$0 \preceq \nabla^2 \varphi_0^{\rho\mu} \preceq \sqrt{\beta_V / \alpha_W} I.$$

Similarly, if V is α_V -strongly convex and W is β_W -smooth, with $\alpha_V, \beta_W > 0$, then

$$\nabla^2 \phi_0^{\rho\mu} \succeq \sqrt{\alpha_V / \beta_W} I.$$

The Hessian bounds are understood almost everywhere.

In this part of the chapter, we will provide a proof of Caffarelli's contraction theorem based on entropic optimal transport, and show how this entropic framework can lead to substantial generalizations. We will first provide an application of Caffarelli's contraction theorem to the transferring of functional inequalities.

2.6.1 Application: Transferring functional inequalities

A common application of Caffarelli's contraction theorem is that it allows the transferring of *functional inequalities* from one measure to another. Two such functional inequalities are the Poincaré and log-Sobolev inequalities.

Definition 2.26 (Poincaré inequality). A probability measure ρ satisfies a Poincaré inequality with constant $C_P(\rho)$ if for all smooth test functions f ,

$$\text{Var}_\rho(f) \leq C_P(\rho) \int \|\nabla f\|^2 d\rho.$$

Definition 2.27 (Log-Sobolev inequality). A probability measure ρ satisfies a log-Sobolev inequality with constant $C_{\text{LSI}}(\rho)$ if for all smooth test functions f ,

$$\text{Ent}_\rho(f^2) \leq C_{\text{LSI}}(\rho) \int \|\nabla f\|^2 d\rho,$$

where, for $h \geq 0$,

$$\text{Ent}_\rho(h) := \int h \log h d\rho - \left(\int h d\rho \right) \log \left(\int h d\rho \right).$$

For example, suppose ρ satisfies a Poincaré inequality with constant $C_P(\rho)$. What is the Poincaré constant of μ given that $T_0 = \nabla \phi_0^{\rho\mu}$ is the optimal transport map from ρ to μ ? With the contraction theorem, this becomes a simple computation: writing $L = \sup_x \|\nabla^2 \phi_0^{\rho\mu}(x)\|_{\text{op}}$, for f smooth,

$$\begin{aligned} \text{Var}_\mu(f) &= \text{Var}_\rho(f \circ T_0) \\ &\leq C_P(\rho) \int \|\nabla(f \circ T_0)(x)\|^2 d\rho(x) \\ &\leq C_P(\rho) L^2 \int \|\nabla f(T_0(x))\|^2 d\rho(x) \\ &= C_P(\rho) L^2 \int \|\nabla f\|^2 d\mu. \end{aligned}$$

Thus, $C_P(\mu) \leq C_P(\rho)L^2$, which can be further controlled if a bound on the spectrum of $\nabla^2 \varphi_0^{\rho\mu}$ is known. We leave the log-Sobolev case as an exercise.

2.6.2 A proof via entropic optimal transport

We will prove the following theorem, which acts as an entropic analogue to Caffarelli's contraction theorem.

Theorem 2.28 (Entropic analogue of Caffarelli's contraction theorem). *Under the upper-bound assumptions of Theorem 2.25, for every $\varepsilon > 0$,*

$$\nabla^2 \varphi_\varepsilon^{\rho\mu}(x) \preceq \frac{1}{2} \left(\sqrt{4\beta_V / \alpha_W + \varepsilon^2 \beta_V^2} - \varepsilon \beta_V \right) I.$$

Under its additional lower-bound assumptions,

$$\nabla^2 \varphi_\varepsilon^{\rho\mu}(x) \succeq \frac{1}{2} \left(\sqrt{4\alpha_V / \beta_W + \varepsilon^2 \alpha_V^2} - \varepsilon \alpha_V \right) I.$$

Taking $\varepsilon \searrow 0$, we formally recover Theorem 2.25. This needs a bit of work to make completely rigorous, but we omit this discussion.

In order to prove Theorem 2.28, we require two complementary key covariance inequalities.

Lemma 2.29. *Let $P \propto e^{-U}$ on $\mathbb{R}^d$ and assume U is twice continuously differentiable on its domain. Then the following hold*

1. (Cramér–Rao inequality) $\text{Cov}_P(X) \succeq (\mathbb{E}_P[\nabla^2 U(X)])^{-1}$.
2. (Brascamp–Lieb inequality) If $\nabla^2 U \succ 0$, then $\text{Cov}_P(X) \preceq \mathbb{E}_P[(\nabla^2 U(X))^{-1}]$.

To continue, we omit the superscripts and instead make use of the following abuse of notation by defining $\pi_\varepsilon^x := \pi_\varepsilon^{Y|X=x}$ and $\pi_\varepsilon^y := \pi_\varepsilon^{X|Y=y}$, identified with their Lebesgue densities, which leads into the following simple computation.

Proposition 2.30. *Let $\rho \propto e^{-V}$ and $\mu \propto e^{-W}$, with V and W smooth. Then*

$$\begin{aligned} \nabla_y^2 (-\log \pi_\varepsilon^x(y)) &= \varepsilon^{-1} \nabla^2 \psi_\varepsilon(y) + \nabla^2 W(y), \\ \nabla_x^2 (-\log \pi_\varepsilon^y(x)) &= \varepsilon^{-1} \nabla^2 \varphi_\varepsilon(x) + \nabla^2 V(x). \end{aligned}$$

Proof of Theorem 2.28. We prove the upper bound; the lower bound follows by interchanging the potentials and applying Cramér–Rao. Write $A(x) = \nabla^2 \varphi_\varepsilon(x)$ and $B(y) = \nabla^2 \psi_\varepsilon(y)$. Fix $x \in \mathbb{R}^d$. An application of the Brascamp–Lieb inequality and Proposition 2.30, to-

gether with the covariance representation of the Hessian, yields

$$A(x) = \varepsilon^{-1} \text{Cov}_{\pi_\varepsilon^x}(Y) \preceq \mathbb{E}_{\pi_\varepsilon^x}[(B(Y) + \varepsilon \nabla^2 W(Y))^{-1}].$$

We repeat the same calculation with the potentials swapped and instead using the Cramér–Rao inequality. This gives

$$B(y) = \varepsilon^{-1} \text{Cov}_{\pi_\varepsilon^y}(X) \succeq \left(\mathbb{E}_{\pi_\varepsilon^y}[A(X) + \varepsilon \nabla^2 V(X)] \right)^{-1}.$$

Note that matrix inversion reverses the positive-definite order on matrices. Defining the quantity

$$L_\varepsilon := \sup_{x \in \mathbb{R}^d} \lambda_{\max}(A(x)) \leq (\varepsilon \alpha_W)^{-1} < \infty.$$

where finiteness follows from $B \succeq 0$ and $\nabla^2 W \succeq \alpha_W I$ in the first bound, we have

$$B(y) \succeq (L_\varepsilon + \varepsilon \beta_V)^{-1} I.$$

where we also used our assumptions on V . Nesting the terms appropriately and using our assumptions on W , we obtain

$$L_\varepsilon \leq ((L_\varepsilon + \varepsilon \beta_V)^{-1} + \varepsilon \alpha_W)^{-1}.$$

As L_ε is finite, we can solve this inequality to obtain

$$L_\varepsilon \leq \frac{1}{2} \left(\sqrt{4\beta_V / \alpha_W + \varepsilon^2 \beta_V^2} - \varepsilon \beta_V \right).$$

□

2.7 Algorithms

2.7.1 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 2.8.

Corollary 2.31. *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 $X_1, \dots, X_n \sim \mu$ and $Y_1, \dots, Y_m \sim \nu$. Then Sinkhorn’s algorithm returns a δ -accurate approximation of $\text{OT}_\varepsilon(\mu_n, \nu_m)$ in $\mathcal{O}(nm c_\infty^2 / (\delta \varepsilon))$.*

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 $i \in [n]$: $f_i^{(k+1)} = -\varepsilon \log\left(\frac{1}{m} \sum_{j=1}^m \exp((g_j^{(k)} - c(X_i, Y_j))/\varepsilon)\right)$

For all $j \in [m]$: $g_j^{(k+1)} = -\varepsilon \log\left(\frac{1}{n} \sum_{i=1}^n \exp((f_i^{(k+1)} - c(X_i, Y_j))/\varepsilon)\right)$

$\Pi_{ij}^{(k+1)} \leftarrow \frac{1}{nm} \exp((f_i^{(k+1)} + g_j^{(k+1)} - c(X_i, Y_j))/\varepsilon)$ 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: $(f^{(k)}, g^{(k)}) \in \mathbb{R}^n \times \mathbb{R}^m$

2.7.2 Implementing entropic Brenier maps

Given the converged empirical potentials $\hat{g}_\varepsilon \in \mathbb{R}^m$ from Algorithm 2, the optimality condition (2.8) and Proposition 2.15 yield the following out-of-sample formula for the entropic Brenier map at any new input $x \in \mathbb{R}^d$:

$$T_\varepsilon(x) = \frac{\sum_{j=1}^m Y_j \exp((g_{\varepsilon,j} - c(x, Y_j))/\varepsilon)}{\sum_{j=1}^m \exp((g_{\varepsilon,j} - c(x, Y_j))/\varepsilon)}. \quad (2.33)$$

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 $x = X_i$, (2.33) reduces to the barycentric projection (1.33).

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: $T_\varepsilon(x) = (\sum_{j=1}^m Y_j w_j) / (\sum_{j=1}^m w_j)$.

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 .

2.7.3 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))
```

3 Flow-based transport

3.1 What is a flow?

What does it mean to “flow” or “interpolate” between probability distributions? If one starts from the definition of a transport map or of a coupling, it is not entirely obvious how to understand notions of *paths* between the starting measure ρ and ending measure μ . For us, interest in these paths is motivated largely by recent computational advances in machine learning (e.g., flow-based diffusion models).

To think of paths, a natural place is via ordinary differential equations (ODEs), where displacement is governed by a family of vector fields $(w_t)_{t \in [0,1]}$ with $w_t : \mathbb{R}^d \rightarrow \mathbb{R}^d$. Indeed, for such a family, the ODE

$$\dot{x}_t = w_t(x_t), \quad x_0 = x \in \mathbb{R}^d$$

integrated from time $t = 0$ to $t = 1$ results in a flow map

$$\Phi_t(x) = x + \int_0^t w_\tau(\Phi_\tau(x)) \, d\tau, \quad (3.1)$$

where Φ_t represents the position of the initial particle $x_0 = x \in \mathbb{R}^d$ over time. In fact, one can write

$$\Phi_{s,t}(x) = x + \int_s^t w_\tau(\Phi_{s,\tau}(x)) \, d\tau,$$

as the flow map with $x_s = x$ (i.e., at time s the particle is located at x). We would then define $\gamma_t = (\Phi_t)_\# \rho$ as the law of the particles at time t , where ρ is the initial law of the particles.

3.2 Deterministic flows and the continuity equation

The evolution of probability measures can be viewed from two equivalent perspectives: the evolution of a single particle, or the evolution of the density as a whole (the aggregate behavior of the particles).

The single-particle perspective is called the *Lagrangian* perspective. The evolution of the density, or the law of a single random variable, is called the Eulerian perspective, and is often what interests us at the end of the day.

Letting $\gamma_t := \text{Law}(X_t)$, we want to make sense of the evolution of the path between $t = 0$ and $t = 1$, which asks, as a prerequisite, how to make sense of “ $\partial_t \gamma_t$ ”. To do so, we integrate this quantity against an arbitrary test function. Moving the derivative outside the integral gives

$$\int h \partial_t \gamma_t = \partial_t \mathbb{E}_{\gamma_t} h(X_t) = \mathbb{E}_{\gamma_t} \langle \nabla h(X_t), \dot{X}_t \rangle,$$

where $X_t \sim \gamma_t$. But we know the evolution of the particles: $\dot{X}_t = w_t(X_t)$! Replacing this and using integration-by-parts results in

$$\int h \partial_t \gamma_t = \int \langle \nabla h, w_t \rangle \gamma_t = - \int h \nabla \cdot (w_t \gamma_t).$$

We arrive at the following result:

Theorem 3.1. *Suppose $(w_t)_{t \in [0,1]}$ is a family of vector fields and the random variables $t \mapsto X_t$ evolve according to $\dot{X}_t = w_t(X_t)$. Then the law of the particles γ_t evolves according to*

$$\partial_t \gamma_t + \nabla \cdot (w_t \gamma_t) = 0,$$

called the continuity equation.

The continuity equation allows us to describe a velocity field that acts on each particle distributed according to a curve of probability measures $t \mapsto \gamma_t$. Viewed from the perspective of transport, the continuity equation says that the total mass of γ_t is preserved under the velocity field w_t that acts on each of the particles distributed according to γ_t . Appending an initial and final measure to the continuity equation results in

$$\partial_t \gamma_t + \nabla \cdot (w_t \gamma_t) = 0, \quad \gamma_0 = \rho, \gamma_1 = \mu. \quad (3.2)$$

Note that (3.2) is directly analogous to the definition of a transport map $T \in \mathcal{T}(\rho, \mu)$. The difference here is that we are asking for a *pair* $(\gamma_t, w_t)_{t \in [0,1]}$ to satisfy the continuity equation with the appropriate endpoints as opposed to just a single function.

3.2.1 Discretizations

In practice, given $(w_t)_{t \in [0,1]}$ and a (discretization of a) source measure ρ , we have enough information to compute flows on a computer.

Perhaps the most obvious scheme is known as the *Euler* discretization. For $h \ll 1$, we approximate the time derivative via finite differences, and rearranging gives the update

$$\dot{X}_t \approx \frac{X_{t+h} - X_t}{h} = w_t(X_t) \implies X_{t+h} = X_t + h w_t(X_t).$$

Choosing $N = 1/h$ steps, the iterates can also be read as

$$X_{(k+1)h} = X_{kh} + h w_{kh}(X_{kh}), \quad X_0 = x_0. \quad (3.3)$$

Note that under suitable assumptions on the velocity field, it is easy to show that Euler discretizations incur an error of $O(h)$ over a fixed time interval.

A more accurate discretization is known as the Runge–Kutta method, or RK4, as the accuracy improves to $O(h^4)$. Here, at each time $t = kh$, we evaluate the velocity field at

four distinct points

$$\begin{aligned} K_1 &= w_t(X_t), \\ K_2 &= w_{t+h/2}(X_t + \frac{h}{2}K_1), \\ K_3 &= w_{t+h/2}(X_t + \frac{h}{2}K_2), \\ K_4 &= w_{t+h}(X_t + hK_3), \end{aligned}$$

with the final update given by

$$X_{t+h} = X_t + \frac{h}{6}(K_1 + 2K_2 + 2K_3 + K_4).$$

Thus, at the cost of four evaluations of the velocity field, we improve the accuracy to fourth-order, allowing for potentially larger step-sizes.

3.3 Stochastic flows and the Fokker–Planck equation

Up until now, we focused on deterministic dynamics of the form

$$\dot{X}_t = w_t(X_t), \quad X_0 = x,$$

where $w_t(\cdot)$ is a time-varying vector field. How would we go about introducing *stochasticity* into the dynamics? Perhaps it is easiest to start from a discretization. Letting $(\eta_{kh})_{k \geq 0}$ be auxiliary random variables, consider the following discretization

$$X_{(k+1)h} = X_{kh} + h w_{kh}(X_{kh}) + \eta_{kh}, \quad X_0 = x_0.$$

Note that this is “very random” in that the law of the sequence η_{kh} has yet to be specified. Let’s first assume that $\eta_{kh} \sim \mathcal{N}(0, I)$ for all $k \geq 0$, independently across iterations. What happens to the dynamics?

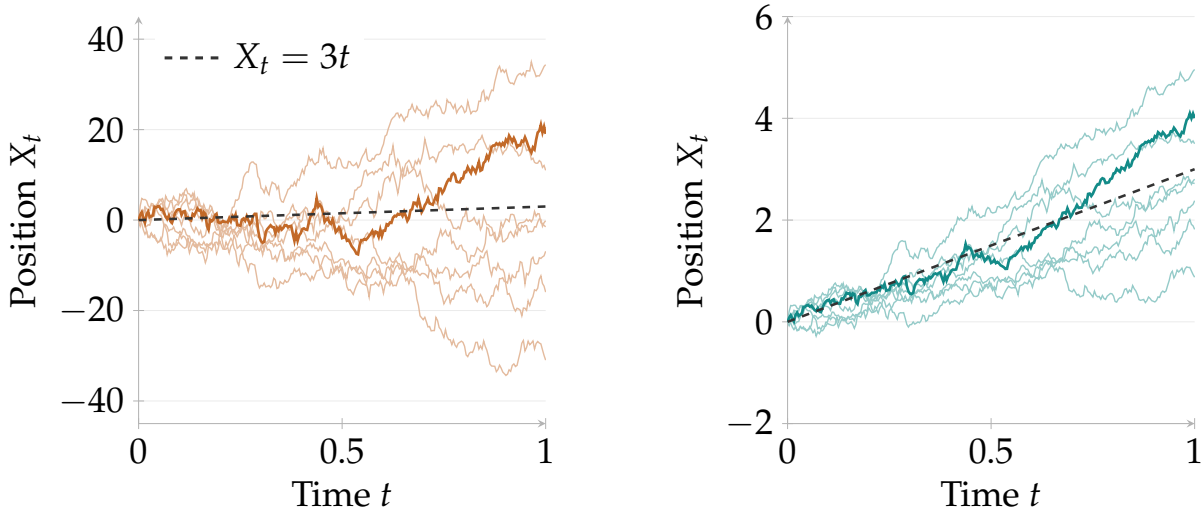
Consider the dynamics given by a constant drift $w_t(x) = 3$ with initialization $X_0 = 0$; the deterministic trajectory is obviously $x_t = 3t$. If we include Gaussian increments in the Euler discretization with stepsize $h = 1/N$ and multiplying each Gaussian increment by an amplitude $a_h > 0$, we obtain at time $t = 1$

$$X_1 = X_0 + 3 \sum_{k=0}^{N-1} h + a_h \sum_{k=0}^{N-1} \eta_{kh} = 3 + a_h \sum_{k=0}^{N-1} \eta_{kh}.$$

We can readily compute the variance as

$$\text{Var}(X_1) = \text{Var}\left(a_h \sum_{k=0}^{N-1} \eta_{kh}\right) = a_h^2 \sum_{k=0}^{N-1} \text{Var}(\eta_{kh}) = a_h^2 N.$$

This calculation shows that if $a_h = 1$, then the variance at time $t = 1$ explodes proportionally to the number of steps; see Figure 6(a). Choosing $a_h = h$ instead gives variance h ,



(a) Unit additive noise yields large variance.

(b) Rescaling noise controls the variance.

Figure 6: Impact of additive Gaussian noise on deterministic dynamics.

so the randomness disappears as $h \searrow 0$. To retain fluctuations at a finite scale, we choose $a_h = \sqrt{\varepsilon h}$ for a fixed $\varepsilon > 0$. In this example, we then have at time $t = nh$

$$X_t = 3t + \sqrt{\varepsilon h} \sum_{k=0}^{n-1} \eta_{kh} \sim \mathcal{N}(3t, \varepsilon t).$$

In particular, for $\varepsilon = 1$, the variance is unity at time $t = 1$; see Figure 6(b). Here the limit $h \searrow 0$ keeps the elapsed time fixed while increasing the number of steps. Each added noise increment becomes smaller, but their variances add to $n\varepsilon h = \varepsilon t$, independently of the discretization. The same scaling applies when the drift depends on positions, and we will henceforth use the discrete dynamics

$$X_{(k+1)h} = X_{kh} + h w_{kh}(X_{kh}) + \sqrt{\varepsilon h} \eta_{kh},$$

Under suitable regularity assumptions on the drift, the limit as $h \searrow 0$ is described by a *stochastic differential equation* (SDE):

$$dX_t = w_t(X_t) dt + \sqrt{\varepsilon} dB_t, \quad X_0 = x, \quad (3.4)$$

where B_t denotes *Brownian motion*.

3.3.1 Brownian motion, Itô processes, Itô's lemma and isometry

Solutions to SDEs have the following “flow map” representation (analogous to (3.1))

$$X_t = X_0 + \int_0^t w_s(X_s) ds + \sqrt{\varepsilon} \int_0^t dB_s. \quad (3.5)$$

The one term that requires some formalization is $\int_0^t dB_s$, because “integrating with respect to Brownian motion” is not a clear concept, as opposed to integrating along the real line. Indeed, making sense of this term requires a bit of care. First, let’s define once and for all what we mean by Brownian motion.

Definition 3.2 (Standard Brownian motion). A stochastic process $(B_t)_{t \geq 0}$ is called (standard) Brownian motion if it satisfies the following properties:

1. $B_0 = 0$,
2. $B_t - B_s \sim \mathcal{N}(0, (t-s)I_d)$ for $0 \leq s < t < +\infty$,
3. for any $k \in \mathbb{N}$, the random variables $(B_{t_1}, B_{t_2} - B_{t_1}, \dots, B_{t_k} - B_{t_{k-1}})$ are mutually independent for $0 < t_1 < \dots < t_k$,
4. and $t \mapsto B_t$ is almost surely continuous.

From the definition, we understand that Brownian increments satisfy

$$B_{t+h} - B_t \sim \mathcal{N}(0, hI_d).$$

As these increments are independent over disjoint time intervals, we can construct the following independent noise variables from a single Brownian motion as

$$\eta_{kh} = \frac{B_{(k+1)h} - B_{kh}}{\sqrt{h}}.$$

Then when discretizing an SDE with a position-dependent drift, writing $X^{(h)}$ for the discretization, summing the updates to time $t = Nh$ yields

$$X_t^{(h)} = x_0 + \sum_{k=0}^{N-1} h w_{kh}(X_{kh}^{(h)}) + \sqrt{\varepsilon} B_t.$$

If, for example, $w_t(x)$ is continuous in time and uniformly Lipschitz in space, the Euler approximations converge as $h \searrow 0$, and the sum above converges to an integral, resulting in the limiting process

$$X_t = x_0 + \int_0^t w_s(X_s) ds + \sqrt{\varepsilon} B_t.$$

The Itô isometry. We now provide a (very brief) introduction to stochastic integrals and the Itô isometry. We will specifically be interested in understanding how to evaluate integrals of the form

$$\mathbb{E} \int_0^T H_t dB_t, \quad \text{and} \quad \mathbb{E} \left| \int_0^T H_t dB_t \right|^2. \quad (3.6)$$

The simplest case is when $H_t = 1$ for all times. Then we can readily compute, by the definition of Brownian motion,

$$\mathbb{E} \int_0^T H_t dB_t = \mathbb{E} \int_0^T 1 dB_t = \mathbb{E}[B_T - B_0] = \mathbb{E}[B_T] = 0,$$

and

$$\mathbb{E} \left| \int_0^T H_t dB_t \right|^2 = \mathbb{E} \left| \int_0^T 1 dB_t \right|^2 = \mathbb{E}[|B_T - B_0|^2] = \text{Var}(B_T) = T.$$

It turns out that this phenomenon can be substantially generalized, where we replace expectations of stochastic integrals by those which are entirely deterministic. This is known as the Itô isometry, which holds for more general stochastic processes $(H_t)_{t \geq 0}$ such that H_t is determined by the information available at time t denoted $\mathcal{F}_t$, which is the adaptedness condition.

Theorem 3.3 (Itô isometry). *Let B be a (one-dimensional) Brownian motion and let $H = (H_t)_{t \geq 0}$ be progressively measurable with $\mathbb{E} \int_0^T |H_t|^2 dt < \infty$. Then $M_t := \int_0^t H_s dB_s$ is a square-integrable martingale, and for every $t \in [0, T]$,*

$$\mathbb{E}[M_t] = 0, \quad \mathbb{E} \left| \int_0^t H_s dB_s \right|^2 = \mathbb{E} \int_0^t |H_s|^2 ds. \quad (3.7)$$

We now provide an informal treatment of the proof of Theorem 3.3.

The first step is to consider a piecewise constant *step process*. Fixing a partition $0 = t_0 < \dots < t_n = T$, and letting $H_i = H_{t_i}$ be square-integrable random variables which are measurable with respect to $\mathcal{F}_{t_i}$, we define our step process with $H_t^{(n)} = H_i$ on $(t_i, t_{i+1}]$, and also write

$$I_n := \int_0^T H_t^{(n)} dB_t = \sum_{i=0}^{n-1} H_i \Delta B_i, \quad \Delta B_i := B_{t_{i+1}} - B_{t_i}.$$

Our goal is to show Itô's isometry for this step process, namely

$$\mathbb{E}[I_n] = 0, \quad \mathbb{E}[I_n^2] = \mathbb{E} \int_0^T (H_t^{(n)})^2 dt. \quad (3.8)$$

Since H_i is $\mathcal{F}_{t_i}$ -measurable and ΔB_i is independent of $\mathcal{F}_{t_i}$ with mean zero (by definition), we have that

$$\mathbb{E}[I_n] = \mathbb{E} \left[\sum_{i=0}^{n-1} H_i \Delta B_i \right] = \sum_{i=0}^{n-1} \mathbb{E}[H_i \Delta B_i] = 0.$$

For the second identity, squaring I_n and expanding gives

$$\mathbb{E}[I_n^2] = \sum_{i=0}^{n-1} \mathbb{E}[H_i^2 (\Delta B_i)^2] + 2 \sum_{0 \leq i < j \leq n-1} \mathbb{E}[H_i \Delta B_i H_j \Delta B_j],$$

The first term above can be simplified to

$$\mathbb{E}[H_i^2(\Delta B_i)^2 \mid \mathcal{F}_{t_i}] = H_i^2 \Delta t_i,$$

The second term vanishes using a conditioning argument: for $i < j$, conditioning just before the later increment gives

$$\mathbb{E}[H_i \Delta B_i H_j \Delta B_j] = \mathbb{E}[H_i \Delta B_i H_j \mathbb{E}[\Delta B_j \mid \mathcal{F}_{t_j}]] = 0.$$

and hence

$$\mathbb{E}[I_n^2] = \sum_{i=0}^{n-1} \mathbb{E}[H_i^2] \Delta t_i = \mathbb{E} \int_0^T (H_t^{(n)})^2 dt.$$

This proves (3.8). Taking suitable limits of step processes then gives the general Itô isometry (3.7).

The Itô isometry is part of the conventional wisdom that says that “ $(dB_t)^2 = dt$ ” in expectation. As a basic application, let $h \in L^2([0, T])$ be a deterministic scalar function and B_t be a d -dimensional Brownian motion. If we apply the isometry coordinatewise, we obtain

$$\text{Cov} \left(\int_0^T h(s) dB_s \right) = \left(\int_0^T h(s)^2 ds \right) I_d.$$

Itô’s lemma. Our final tool from this section will be the following foundational lemma.

Theorem 3.4 (Itô’s lemma). *Let $(X_t)_{t \in [0,1]}$ be an Itô process satisfying $dX_t = w_t(X_t) dt + \sigma_t dB_t$, and let $h \in C^{1,2}([0, 1] \times \mathbb{R}^d)$. Then*

$$\begin{aligned} dh_t(X_t) = & \left(\partial_t h_t + \langle \nabla_x h_t(X_t), w_t(X_t) \rangle + \frac{1}{2} \text{tr}(\sigma_t \sigma_t^\top \nabla_x^2 h_t(X_t)) \right) dt \\ & + \langle \nabla_x h_t(X_t), \sigma_t dB_t \rangle. \end{aligned} \quad (3.9)$$

Proof. Write $h_t(x) = h(t, x)$ and let $0 < \delta \ll 1$. Along a short interval, the expression for the SDE can be written as

$$\Delta X := X_{t+\delta} - X_t \approx w_t(X_t) \delta + \sigma_t \Delta B, \quad \Delta B := B_{t+\delta} - B_t \sim \mathcal{N}(0, \delta I).$$

A key observation here is the size of ΔB , where the *variance* (in other words, its square) has size δ , and so ΔB can be thought of as having size $\sqrt{\delta}$. We now expand $h_t(X_t)$:

$$\begin{aligned} h_{t+\delta}(X_{t+\delta}) - h_t(X_t) & \approx \partial_t h_t(X_t) \delta + \langle \nabla_x h_t(X_t), \Delta X \rangle + \frac{1}{2} \Delta X^\top \nabla_x^2 h_t(X_t) \Delta X \\ & \approx \left(\partial_t h_t(X_t) + \langle \nabla_x h_t(X_t), w_t(X_t) \rangle \right) \delta \\ & \quad + \langle \nabla_x h_t(X_t), \sigma_t \Delta B \rangle + \frac{1}{2} \Delta B^\top K_t \Delta B, \end{aligned} \quad (3.10)$$

where $K_t := \sigma_t^\top \nabla_x^2 h_t(X_t) \sigma_t$, and we exclude lower-order terms in $\delta \approx dt$.

We now want to write the above in the form of (3.5). To do so, we sum the increments across a partition $0 = t_0 < t_1 < \dots < t_N = T$ with $t_i = i\delta$ and $N\delta = T$, and then take $\delta \searrow 0$. The left-hand side of (3.10) would simply give $h_T(X_T) - h_0(X_0)$. The first term on the right-hand side of (3.10) becomes an integral in a standard way

$$\begin{aligned} & \sum_{i=0}^{N-1} \left(\partial_t h(t_i, X_{t_i}) + \langle \nabla_x h_{t_i}(X_{t_i}), w_{t_i}(X_{t_i}) \rangle \right) \delta \\ & \longrightarrow \int_0^T \left(\partial_t h(t, X_t) + \langle \nabla_x h_t(X_t), w_t(X_t) \rangle \right) dt, \end{aligned}$$

whereas the second term becomes an Itô integral:

$$\sum_{i=0}^{N-1} \langle \nabla_x h_{t_i}(X_{t_i}), \sigma_{t_i} \Delta B_i \rangle \longrightarrow \int_0^T \langle \nabla_x h_t(X_t), \sigma_t dB_t \rangle,$$

where $\Delta B_i := B_{t_{i+1}} - B_{t_i}$. For the final term, we then write $\Delta B_i = \sqrt{\delta} \xi_i$ with independent $\xi_i \sim \mathcal{N}(0, I)$, with each ξ_i independent of the information at time t_i and $\mathbb{E}[\xi_i \xi_i^\top] = I$. We write the sum as

$$\frac{1}{2} \sum_{i=0}^{N-1} \Delta B_i^\top K_{t_i} \Delta B_i = \frac{\delta}{2} \sum_{i=0}^{N-1} \text{tr}(K_{t_i} \xi_i \xi_i^\top).$$

As $\delta \rightarrow 0$, the fluctuations of these terms end up vanishing, and we can instead replace the term above by its expectation. Under this heuristic justification (which can be rigorously justified), the resulting Riemann sum converges

$$\begin{aligned} \frac{1}{2} \sum_{i=0}^{N-1} \Delta B_i^\top K_{t_i} \Delta B_i & \approx \frac{1}{2} \sum_{i=0}^{N-1} \Delta \mathbb{E} \text{Tr}(K_{t_i} \xi_i \xi_i^\top) \\ & = \frac{\delta}{2} \sum_{i=0}^{N-1} \text{tr}(K_{t_i}) \longrightarrow \frac{1}{2} \int_0^T \text{tr}(K_t) dt = \frac{1}{2} \int_0^T \text{tr}(\sigma_t \sigma_t^\top \nabla_x^2 h_t(X_t)) dt. \end{aligned}$$

□

Theorem 3.4 is essentially a chain rule for stochastic differential equations. The difference from the usual chain rule is the additional second-order *Itô correction* given by

$$\frac{1}{2} \text{tr}(\sigma_t \sigma_t^\top \nabla_x^2 h_t(X_t)) dt$$

This term precisely appears because Brownian motion has nontrivial “quadratic variation”. More heuristically, this is because $(dB_t)^2 = dt$. Note that setting $\sigma_t = 0$ recovers the ordinary chain rule.

3.3.2 Derivation of the Fokker–Planck equation

The goal of this section is to prove a “diffused” analogue of the continuity equation, known as the *Fokker–Planck equation*. Specifically, we want to understand the law of the particles X_t that undergo the Itô process

$$dX_t = w_t(X_t) dt + \sqrt{\varepsilon} dB_t.$$

Recall that the case $\varepsilon = 0$ gives precisely the dynamics that gave rise to the continuity equation.

Theorem 3.5 (Fokker–Planck equation). *Suppose $(w_t)_{t \in [0,1]}$ is a family of vector fields and the random variables $t \mapsto X_t$ evolve according to $dX_t = w_t(X_t) dt + \sqrt{\varepsilon} dB_t$. Then the law of the particles, denoted ρ_t , evolves according to*

$$\partial_t \rho_t + \nabla \cdot (w_t \rho_t) = \frac{\varepsilon}{2} \Delta \rho_t,$$

which is called the Fokker–Planck equation.

Proof. The proof follows that of the continuity equation: let h be a suitably nice test function; then we have

$$\int h \partial_t \rho_t = \partial_t \mathbb{E}[h(X_t)]. \quad (3.11)$$

On the other hand, we can use Itô’s lemma instead of the chain rule on the test function, which leads to

$$\begin{aligned} d\mathbb{E}[h(X_t)] &= \mathbb{E}[dh(X_t)] \\ &= \mathbb{E}[(\langle \nabla h(X_t), w_t(X_t) \rangle + \frac{\varepsilon}{2} \Delta h(X_t)) dt + \sqrt{\varepsilon} \langle \nabla h(X_t), dB_t \rangle] \\ &= \mathbb{E}[(\langle \nabla h(X_t), w_t(X_t) \rangle + \frac{\varepsilon}{2} \Delta h(X_t)) dt], \end{aligned}$$

where the last equality follows by the martingale property of the stochastic integral. Altogether, we have

$$\partial_t \mathbb{E}[h(X_t)] = \mathbb{E}[(\langle \nabla h(X_t), w_t(X_t) \rangle + \frac{\varepsilon}{2} \Delta h(X_t))].$$

Equating this with (3.11) gives

$$\int h \partial_t \rho_t = \int (\langle \nabla h, w_t \rangle + \frac{\varepsilon}{2} \Delta h) \rho_t.$$

Two applications of integration-by-parts yield the final result. □

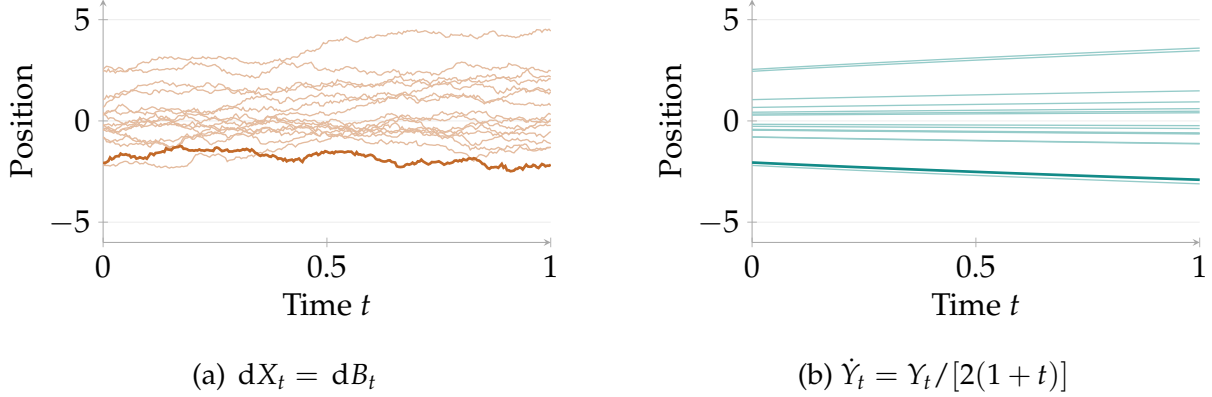


Figure 7: Different trajectories (SDEs and ODEs) with the same t -marginals $\mathcal{N}(0, 1+t)$.

3.4 Properties of SDEs

3.4.1 ODEs preserve marginal laws

Recall that the SDE dynamics

$$dX_t = w_t(X_t) dt + \sigma_t dB_t,$$

result in the Fokker–Planck equation for $\rho_t = \text{Law}(X_t)$

$$\partial_t \rho_t + \nabla \cdot (\rho_t w_t) = \frac{1}{2} \nabla \cdot (a_t \nabla \rho_t),$$

where $a_t = \sigma_t \sigma_t^\top$ is a diffusion matrix. Writing the Laplacian as

$$\Delta \rho_t = \nabla \cdot (\rho_t \nabla \log \rho_t),$$

we can rewrite the Fokker–Planck equation in terms of the continuity equation

$$\partial_t \rho_t + \nabla \cdot (\rho_t (w_t - \frac{1}{2} a_t \nabla \log \rho_t)) = 0.$$

As the definition of ρ_t has not changed between the two formulations, we can conclude that solutions to the ODE

$$\dot{X}_t = (w_t - \frac{1}{2} a_t \nabla \log \rho_t)(X_t),$$

result in the same time- t marginals. Figure 7 illustrates how it's possible for two probability paths to have the same marginal laws.

3.4.2 Forward and reverse SDEs

Consider the “forward” dynamics for $t \in [0, 1]$ which induce transport from ρ to μ

$$dX_t^\rightarrow = b_t^\rightarrow(X_t^\rightarrow) dt + \sqrt{2} dB_t,$$

where we write $\gamma_t^\rightarrow := \text{Law}(X_t^\rightarrow)$. Then for any fixed $T \in [0, 1]$, define for $t \in [0, T]$

$$X_t^\leftarrow := X_{T-t}^\rightarrow.$$

What SDE does the process $(X_t^\leftarrow)_{t \in [0, T]}$ then obey?

Derivation via the Fokker–Planck equation Writing $\gamma_t^\leftarrow := \gamma_{T-t}^\rightarrow$, we have by the Fokker–Planck equation for the forward dynamics

$$\partial_t \gamma_t^\leftarrow = \partial_t (\gamma_{T-t}^\rightarrow) = -(\partial_s \gamma_s^\rightarrow)|_{s=T-t} = \nabla \cdot (\gamma_{T-t}^\rightarrow b_{T-t}^\rightarrow) - \Delta \gamma_{T-t}^\rightarrow.$$

To obtain a similar Fokker–Planck equation, we just need to introduce the right Laplacian scaling:

$$\begin{aligned} \partial_t \gamma_t^\leftarrow &= \nabla \cdot (\gamma_{T-t}^\rightarrow b_{T-t}^\rightarrow) - \Delta \gamma_{T-t}^\rightarrow \\ &= \nabla \cdot (\gamma_{T-t}^\rightarrow b_{T-t}^\rightarrow) - \Delta \gamma_{T-t}^\rightarrow + \Delta \gamma_t^\leftarrow - \Delta \gamma_t^\leftarrow \\ &= \nabla \cdot (\gamma_{T-t}^\rightarrow b_{T-t}^\rightarrow) - 2\Delta \gamma_{T-t}^\rightarrow + \Delta \gamma_t^\leftarrow \\ &= -\nabla \cdot (\gamma_t^\leftarrow (-b_{T-t}^\rightarrow + 2\nabla \ln \gamma_{T-t}^\rightarrow)) + \Delta \gamma_t^\leftarrow. \end{aligned}$$

Thus, the reverse SDE is written as

$$dX_t^\leftarrow = b_t^\leftarrow(X_t^\leftarrow) dt + \sqrt{2} d\bar{B}_t,$$

where $b_t^\leftarrow$ satisfies

$$b_t^\leftarrow = -b_{T-t}^\rightarrow + 2\nabla \ln \gamma_{T-t}^\rightarrow,$$

and $\bar{B}_t$ is Brownian motion with respect to the reverse-time filtration.

Derivation via transition densities. We now provide an alternative heuristic derivation of the reverse SDE using transition densities.

Write

$$X_{t+\Delta t} = X_t + \int_t^{t+\Delta t} f_s(X_s) ds + \int_t^{t+\Delta t} \sigma dB_s \approx X_t + \Delta t f_t(X_t) + \sigma(B_{t+\Delta t} - B_t).$$

If $X_t = x$ is fixed, then

$$X_{t+\Delta t} | X_t = x \approx x + \Delta t f_t(x) + \sqrt{\sigma^2 \Delta t} \xi \sim \mathcal{N}(x + \Delta t f_t(x), \sigma^2 \Delta t I),$$

and thus writing x' as a future draw of the process

$$p(x' | x) \approx \mathcal{N}(x + \Delta t f_t(x), \sigma^2 \Delta t I).$$

Our goal is to compute the reverse, namely $p(x|x')$. By the basic rules of conditional probability, we have

$$p(x|x') = p(x) \cdot p(x'|x) / p(x'),$$

where, by an abuse of notation, $X_t = x$ is the current iterate. Taking logs and using our approximation from before, we have that

$$\begin{aligned} \log p(x|x') &= \log p(x) + \log p(x'|x) - \log p(x') \\ &\approx \log p(x) - \frac{1}{2\sigma^2\Delta t} \|x' - x - \Delta t f_t(x)\|^2 - \log p(x') + C_{\Delta t}. \end{aligned}$$

We will now expand around $x = x' + h$ for $h = \mathcal{O}(\sqrt{\Delta t})$. (Why? Recall that $x' - x = \Delta t f_t(x) + \sqrt{\sigma^2\Delta t}\xi \approx \mathcal{O}(\sqrt{\Delta t})$ for $\Delta t \ll 1$.) This yields

$$\begin{aligned} \log p(x) &\approx \log p(x') + \langle \nabla \log p(x'), h \rangle + \mathcal{O}(h^2) \\ &\approx \log p(x') + \langle \nabla \log p(x'), h \rangle + \mathcal{O}(\Delta t). \end{aligned}$$

Here it is useful to make the time indices of the marginal densities explicit. In particular, $p(x) = p_t(x)$ and $p(x') = p_{t+\Delta t}(x')$. The difference between $\log p_t(x')$ and $\log p_{t+\Delta t}(x')$ is of order $\mathcal{O}(\Delta t)$ and, since it does not depend on h , can be absorbed into the normalizing constant of the conditional density.

We also have

$$\begin{aligned} \|x' - x - \Delta t f_t(x)\|^2 &= \|-h - \Delta t f_t(x' + h)\|^2 \\ &= \|h\|^2 + 2\Delta t \langle h, f_t(x') \rangle + \mathcal{O}(\Delta t^2), \end{aligned}$$

where we have used $h = \mathcal{O}(\sqrt{\Delta t})$ and the smoothness of f_t . Consequently, up to terms of order $\mathcal{O}(\Delta t)$,

$$\begin{aligned} \log p(x | x') &= C(x', \Delta t) - \frac{\|h\|^2}{2\sigma^2\Delta t} + \langle \nabla \log p_t(x') - \frac{1}{\sigma^2} f_t(x'), h \rangle + \mathcal{O}(\Delta t) \\ &= C'(x', \Delta t) - \frac{1}{2\sigma^2\Delta t} \left\| h - \Delta t (\sigma^2 \nabla \log p_t(x') - f_t(x')) \right\|^2 + \mathcal{O}(\Delta t), \end{aligned}$$

where $C(x', \Delta t)$ and $C'(x', \Delta t)$ denote terms that do not depend on h . We therefore obtain the approximate reverse transition

$$X_t | X_{t+\Delta t} = x' \approx \mathcal{N}\left(x' + \Delta t (\sigma^2 \nabla \log p_t(x') - f_t(x')), \sigma^2 \Delta t I\right).$$

Equivalently, a single reverse step can be written as

$$X_t \approx X_{t+\Delta t} + \Delta t (-f_t(X_{t+\Delta t}) + \sigma^2 \nabla \log p_t(X_{t+\Delta t})) + \sigma \sqrt{\Delta t} \xi, \quad \xi \sim \mathcal{N}(0, I).$$

Now let $X_s^{\leftarrow} = X_{T-s}^{\rightarrow}$ and set $r = T - s$. A forward step in reverse time corresponds to a backward step in the original time:

$$\begin{aligned} X_{s+\Delta t}^{\leftarrow} \mid X_s^{\leftarrow} = x' &= X_{r-\Delta t}^{\rightarrow} \mid X_r^{\rightarrow} = x' \\ &\approx \mathcal{N}\left(x' + \Delta t(-f_r(x') + \sigma^2 \nabla \log p_r(x')), \sigma^2 \Delta t I\right). \end{aligned}$$

Replacing quantities evaluated at $r - \Delta t$ by quantities evaluated at r only changes the mean by a higher-order term. Passing formally to the limit $\Delta t \rightarrow 0$, the reverse-time process therefore satisfies

$$dX_s^{\leftarrow} = (-f_{T-s}(X_s^{\leftarrow}) + \sigma^2 \nabla \log p_{T-s}(X_s^{\leftarrow})) ds + \sigma d\bar{B}_s,$$

where $(\bar{B}_s)_{s \in [0, T]}$ is a Brownian motion with respect to the reverse-time filtration.

In the setting of the forward SDE above, $\sigma = \sqrt{2}$, $f_t = b_t^{\rightarrow}$, and $p_t = \gamma_t^{\rightarrow}$. Hence

$$dX_s^{\leftarrow} = (-b_{T-s}^{\rightarrow}(X_s^{\leftarrow}) + 2 \nabla \log \gamma_{T-s}^{\rightarrow}(X_s^{\leftarrow})) ds + \sqrt{2} d\bar{B}_s.$$

Thus

$$b_s^{\leftarrow}(x) = -b_{T-s}^{\rightarrow}(x) + 2 \nabla \log \gamma_{T-s}^{\rightarrow}(x),$$

or, equivalently,

$$b_s^{\leftarrow} + b_{T-s}^{\rightarrow} = 2 \nabla \log \gamma_{T-s}^{\rightarrow}.$$

The reverse process starts from $X_0^{\leftarrow} \sim \gamma_T^{\rightarrow}$ and ends at $X_T^{\leftarrow} \sim \gamma_0^{\rightarrow}$.

3.4.3 Girsanov's theorem

Consider two probability paths with different diffusion coefficients: one is Brownian motion and the other is a standard SDE:

$$\begin{aligned} \mathbb{P} : dY_t &= \sqrt{\alpha} dB_t, \\ \mathbb{Q} : dX_t &= b_t(X_t) dt + \sqrt{\beta} dB_t. \end{aligned}$$

Girsanov's theorem is essentially a statement regarding densities of probability path measures, making sense of statements regarding

$$\frac{d\mathbb{Q}}{d\mathbb{P}}.$$

(To see why this is important, think about what it means to compute the KL divergence on path space.) We will sketch out a heuristic argument for why certain conditions are required before stating the full result.

Instead of writing out the argument with the full SDE, we will consider a discretization and give a heuristic argument. Let $0 < h \ll 1$ be a small time-step and write

$$\delta X = X_{t+h} - X_t \approx \mathcal{N}(b_t(X_t)h, \beta h), \quad \delta Y = Y_{t+h} - Y_t \approx \mathcal{N}(0, \alpha h).$$

We can compute the likelihood ratio of the densities as

$$\frac{\beta^{-d/2} \exp\left(-\|\delta X - b_t(X_t)h\|^2 / (2\beta h)\right)}{\alpha^{-d/2} \exp\left(-\|\delta Y\|^2 / (2\alpha h)\right)}.$$

Since we want to understand the ratio of Q relative to P, we can replace δX by δY (i.e., evaluate both conditional transition densities along the same path) and simplify the log-ratio. This gives

$$\begin{aligned} \ell_{\text{ratio}} &= \frac{d}{2} \log\left(\frac{\alpha}{\beta}\right) - \frac{\|\delta Y - b_t(Y_t)h\|^2}{2\beta h} + \frac{\|\delta Y\|^2}{2\alpha h} \\ &= \frac{d}{2} \log\left(\frac{\alpha}{\beta}\right) + \frac{1}{2}(\alpha^{-1} - \beta^{-1}) \frac{\|\delta Y\|^2}{h} + \frac{\langle b_t(Y_t), \delta Y \rangle}{\beta} - \frac{\|b_t(Y_t)\|^2}{2\beta} h. \end{aligned}$$

If we want to approximate the whole process until time T , we can take a sum of the log-likelihood ratios (i.e. product of the likelihood ratios) to arrive at

$$\sum_{k=0}^{N-1} \ell_{\text{ratio}}^{(k)} = \frac{Nd}{2} \log\left(\frac{\alpha}{\beta}\right) + \frac{1}{2}(\alpha^{-1} - \beta^{-1}) \sum_{k=0}^{N-1} \frac{\|\delta Y\|^2}{h} + \sum_{k=0}^{N-1} \left[\frac{\langle b_t(Y_t), \delta Y \rangle}{\beta} - \frac{\|b_t(Y_t)\|^2}{2\beta} h \right].$$

As the stochastic processes go until time T (potentially with $T \rightarrow \infty$), two of the terms above pose difficulties. The first term is problematic on its own

$$\frac{Nd}{2} \log\left(\frac{\alpha}{\beta}\right),$$

as the $N \rightarrow \infty$ limit diverges, where we choose $N = T/h$. However, the second term is also problematic, as we have the approximations

$$\sum_{k=0}^{N-1} \frac{\|\delta Y\|^2}{h} \approx \sum_{k=0}^{N-1} \frac{\alpha \|\delta B\|^2}{h} \approx \sum_{k=0}^{N-1} \frac{\alpha dh}{h} = Nd\alpha,$$

which also diverges as $N \rightarrow \infty$.

If we instead restrict ourselves to $\alpha = \beta$, then both of these terms no longer impede us from taking the $N \rightarrow \infty$ limit (indeed, the prefactor for the second term is zero) and we obtain the log-ratio

$$\ell_{\text{ratio}}^{\infty} = \frac{1}{\sqrt{\alpha}} \int_0^T \langle b_t(Y_t), dB_t \rangle - \frac{1}{2\alpha} \int_0^T \|b_t(Y_t)\|^2 dt.$$

Theorem 3.6 (Girsanov's theorem, informal). *Let P be the path measure of*

$$dY_t = \sqrt{\alpha} dB_t, \quad Y_0 \sim p_0,$$

and let Q be the path measure of

$$dX_t = b_t(X_t) dt + \sqrt{\alpha} dB_t, \quad X_0 \sim p_0,$$

on the same time interval $[0, T]$ for $T < +\infty$ and $\alpha > 0$. Under sufficient integrability conditions (e.g., Novikov's condition), Q is absolutely continuous with respect to P and

$$\frac{dQ}{dP}(Y_{[0,T]}) = \exp\left(\frac{1}{\sqrt{\alpha}} \int_0^T \langle b_t(Y_t), dB_t \rangle - \frac{1}{2\alpha} \int_0^T \|b_t(Y_t)\|^2 dt\right).$$

In particular, the KL divergence between the two measures on path space is given by

$$\text{KL}(Q \parallel P) = \frac{1}{2\alpha} \mathbb{E}_Q \int_0^T \|b_t(X_t)\|^2 dt.$$

4 Dynamic formulation of optimal transport

4.1 Optimal vector fields

As with pushforward maps, we can hope to constrain our velocity fields to be as “lazy” as possible, where we want to estimate two quantities: $(\gamma_t, w_t)_{t \in [0,1]}$. Since the velocity field w_t acts on particles $X_t \sim \gamma_t$, and the pair satisfies the continuity equation, we arrive at the following natural definition of an optimal vector field, and its associated density, defined in terms of *minimal kinetic energy*.

Definition 4.1 (Optimal kinetic vector field). A pair $(\rho_t^*, v_t^*)_{t \in [0,1]}$ is called optimal w.r.t. the continuity equation with $\rho_0 = \rho$ and $\rho_1 = \mu$ if it solves

$$(\rho_t^*, v_t^*)_{t \in [0,1]} := \operatorname{argmin}_{\gamma_t, w_t} \int_0^1 \int \frac{1}{2} \|w_t(x)\|^2 d\gamma_t(x) dt \quad \text{s.t.} \quad \begin{cases} \partial_t \gamma_t + \nabla \cdot (w_t \gamma_t) = 0, \\ \gamma_0 = \rho, \gamma_1 = \mu. \end{cases}$$

4.1.1 Benamou–Brenier formulation of optimal transport

Our goal will be to prove the following result, which unifies the static formulations we saw in Section 1.4 and the dynamic version of transport via vector fields.

Corollary 4.2 (Benamou–Brenier formulation). *It holds that*

$$\frac{1}{2} W_2^2(\rho, \mu) = \min_{\gamma_t, w_t} \int_0^1 \int \frac{1}{2} \|w_t(z)\|^2 d\gamma_t(z) dt,$$

over all pairs (γ_t, w_t) that satisfy the continuity equation with endpoint constraints $\gamma_0 = \rho$ and $\gamma_1 = \mu$.

4.1.2 Proof via dynamic dual formulation

As with all previous formulations of optimal transport, the dynamic formulation also comes with a dual interpretation, reminiscent of objects from the literature on optimal control. We follow the discussions in [CGP21].

We will compute the dynamic formulation via Lagrange multipliers, where our constraint is the continuity equation. Letting $g_t : \mathbb{R}^d \rightarrow \mathbb{R}$ be a family of smooth functions indexed by $t \in [0, 1]$, we have

$$\inf_{(\gamma_t, w_t)} \int_0^1 \int \frac{1}{2} \|w_t(x)\|^2 \gamma_t(x) + g_t(\partial_t \gamma_t + \nabla \cdot (w_t \gamma_t)) dx dt.$$

An application of integration-by-parts gives

$$\int \int_0^1 g_t \partial_t \gamma_t dt dx = \int (g_1 \gamma_1 - g_0 \gamma_0) dx - \int_0^1 \int \partial_t g_t \gamma_t dx dt,$$

where the boundary terms here do not depend on the intermediate path or velocity field, and so we omit this term. Another application of integration-by-parts gives

$$\int \int_0^1 g_t \nabla \cdot (\gamma_t w_t) \, dt \, dx = - \int_0^1 \int \langle \nabla g_t, w_t \rangle \gamma_t \, dx \, dt,$$

where, here, we assume the boundary terms vanish. Together, we arrive at

$$\inf_{(\gamma_t, w_t)} \int_0^1 \int \left(\frac{1}{2} \|w_t(x)\|^2 - \partial_t g_t - \langle \nabla g_t, w_t \rangle \right) \gamma_t(x) \, dx \, dt.$$

For fixed γ_t , optimizing over w_t yields an optimal vector field $v_t := w_t^*$ as a function of the multiplier

$$v_t(x) = \nabla g_t(x), \tag{4.1}$$

with the resulting objective being

$$\inf_{\gamma_t} \int_0^1 \int \left(-\partial_t g_t - \frac{1}{2} \|\nabla g_t\|^2 \right) \gamma_t(x) \, dx \, dt + \int g_1 \, d\mu - \int g_0 \, d\rho.$$

The minimization over nonnegative γ_t is finite only when

$$\partial_t g_t + \frac{1}{2} \|\nabla g_t\|^2 \leq 0.$$

Thus the dynamic dual is naturally posed over Hamilton–Jacobi *subsolutions*, but if the constraint is saturated, then we can safely write

$$\partial_t g_t + \frac{1}{2} \|\nabla g_t\|^2 = 0. \tag{HJ}$$

Subject to a terminal condition $g_1 = g$ at time $t = 1$, the dual objective is

$$\sup_g \left\{ \int g \, d\mu - \int g_0 \, d\rho \right\},$$

where $(g_t)_{t \in [0,1]}$ is the backward Hamilton–Jacobi evolution. Keeping the time indices explicit, we define the backward Hopf–Lax operator for $s > 0$ by

$$Q_s g(x) := \sup_y \left\{ g(y) - \frac{\|x - y\|^2}{2s} \right\},$$

whence $g_t = Q_{1-t} g$ is the viscosity solution of (HJ) with terminal value $g_1 = g$. At time $t = 0$ this gives

$$g_0(x) = Q_1 g(x) = \sup_y \left\{ g(y) - \frac{1}{2} \|x - y\|^2 \right\} = -g^c(x),$$

where we recall the c -transform from Section 1.5.2. Choosing the terminal condition $g = g^{\rho\mu}$ therefore gives

$$\begin{aligned} \int g \, d\mu - \int g_0 \, d\rho &= \int g^{\rho\mu} \, d\mu - \int (-(g^{\rho\mu})^c) \, d\rho \\ &= \int f^{\rho\mu} \, d\rho + \int g^{\rho\mu} \, d\mu \\ &= \frac{1}{2} W_2^2(\rho, \mu). \end{aligned}$$

4.2 Geometry of optimal transport

We previously saw that the Wasserstein space $(\mathcal{P}_2(\mathbb{R}^d), W_2)$ is a metric space, allowing us to meaningfully distinguish probability measures, and metrize weak convergence. Here, we will see that the Wasserstein space effectively mimics a Riemannian manifold and enjoys a notion of calculus.

4.2.1 Primer on Riemannian geometry

A set $\mathcal{M}$ is called a manifold if it is locally homeomorphic to a Euclidean space. (Note that this already poses a challenge with the Wasserstein space, as Euclidean spaces are always finite-dimensional whereas the space of probability measures is infinite-dimensional.) For $x \in \mathcal{M}$, we write $T_x\mathcal{M}$ for the tangent space at x , which consists of all tangent vectors emanating from x . For each x and corresponding $T_x\mathcal{M}$, we can define an inner product g_x , called a Riemannian metric. Altogether, these elements define a Riemannian manifold.

A *geodesic* is defined as the shortest path between two points $x_0, x_1 \in \mathcal{M}$, written $(x_t)_{t \in [0,1]} \in \mathcal{M}$, and given by the solution to the following variational problem

$$(x_t)_{t \in [0,1]} \in \operatorname{argmin}_{(c_t)_{t \in [0,1]}} \int \|\dot{c}_t\|_{g(c_t)}^2 \, dt, \quad (4.2)$$

where the minimization occurs over all admissible curves between x_0 and x_1 .

For $x_0 \in \mathcal{M}$ with velocity $v \in T_{x_0}\mathcal{M}$, recall that the *exponential map* $\exp_{x_0}(v)$ gives the terminal location after a unit of time along v . The inverse mapping is known as the *logarithmic map* $\log_{x_0}(x_1) \in T_{x_0}\mathcal{M}$, which yields the initial tangent vector for the geodesic from x_0 to x_1 . Finally, for $t \in [0,1]$, we can write geodesics over $\mathcal{M}$ as

$$x_t = \exp_{x_0}(t \log_{x_0}(x_1)).$$

4.2.2 Riemannian view of the Wasserstein space

Our goal now is to provide a Riemannian interpretation of the Wasserstein space, the space of probability measures with finite second moment endowed with the 2-Wasserstein

metric. Actually, we will work over the space of probability *densities* with finite second moment, for reasons that will become apparent momentarily.

If $\rho, \mu \in \mathcal{P}_{2,\text{ac}}(\mathbb{R}^d)$, then their geodesic would be defined as the solution to the variational problem

$$(\rho_t)_{t \in [0,1]} \in \operatorname{argmin}_{(\gamma_t)_{t \in [0,1]}} \int_0^1 \|\dot{\gamma}_t\|_{g(\gamma_t)}^2 dt,$$

where the curves have endpoints ρ and μ , although the metric and tangent vectors are not yet defined. However, the Benamou–Brenier formulation expresses the 2-Wasserstein distance as the kinetic energy of an optimal path, written

$$\frac{1}{2} W_2^2(\rho, \mu) = \int_0^1 \frac{1}{2} \|v_t^*\|_{L^2(\rho_t^*)}^2 dt,$$

which is essentially of the form we want, and gives us all the ingredients we need to describe a formal Riemannian structure.

The inner product that emerges from the Benamou–Brenier structure is the L^2 inner product, which writes

$$g_\rho(v, w) := \int \langle v(x), w(x) \rangle d\rho(x),$$

for $\rho \in \mathcal{P}_{2,\text{ac}}(\mathbb{R}^d)$ and two vector fields v, w . The tangent space will then consist of (the L^2 closure of the space of) gradient fields

$$T_\rho \mathcal{P}_2(\mathbb{R}^d) := \overline{\{\nabla \psi : \psi \in C_c^\infty(\mathbb{R}^d)\}}^{L^2(\rho)}.$$

This can be seen because of the continuity equation. Indeed, recall that for a curve $(\gamma_t)_{t \in [0,1]}$ and corresponding velocity $(v_t)_{t \in [0,1]}$, we have that

$$\partial_t \gamma_t = -\nabla \cdot (\gamma_t v_t).$$

Going a step further, we see that testing against a smooth function $\psi \in C_c^\infty(\mathbb{R}^d)$ and integrating by parts gives

$$\frac{d}{dt} \int \psi d\gamma_t = \int \langle \nabla \psi, v_t \rangle d\gamma_t.$$

Thus the density γ_t changes with v_t only through its inner products with gradient fields. Velocities orthogonal to all gradient fields do not affect this evolution, so we use the closure of gradient fields as our tangent space.

We now connect this to the more canonical description of a geodesic, which moves particles in a “straight line” to their destinations, which is also governed by the Monge formulation:

$$\frac{1}{2} W_2^2(\rho, \mu) = \int_0^1 \frac{1}{2} \|v_t^*\|_{L^2(\rho_t^*)}^2 dt = \frac{1}{2} \int \|T^{\rho \rightarrow \mu}(x) - x\|^2 d\rho(x).$$

We recall the *McCann interpolation* from (1.20), which suggests paths as

$$\rho_t^* := (T_t)_\# \rho := ((1-t)\text{id} + tT^{\rho \rightarrow \mu})_\# \rho.$$

Writing $T = T^{\rho \rightarrow \mu}$, the particle starting at x has constant velocity $T(x) - x$, giving

$$v_t^*(T_t(x)) = T(x) - x.$$

The path is therefore optimal by the Benamou–Brenier formula.

We now further explain the constant-speed aspect. Let $(\rho_t^*)_{t \in [0,1]}$ be the Benamou–Brenier curve from ρ to μ . We can bound the 2-Wasserstein distance between ρ_s^* and ρ_t^* as

$$\begin{aligned} W_2^2(\rho_s^*, \rho_t^*) &\leq \int \|T_t(x) - T_s(x)\|^2 d\rho(x) \\ &= (t-s)^2 \int \|T(x) - x\|^2 d\rho(x) = (t-s)^2 W_2^2(\rho, \mu), \end{aligned}$$

where in the first inequality, we used the naive coupling $(T_s, T_t)_\# \rho$, and used the fact that $(T_t - T_s)(x) = (t-s)(T(x) - x)$ by definition. Taking square roots gives $W_2(\rho_s^*, \rho_t^*) \leq |t-s|W_2(\rho, \mu)$. However, $0 \leq s \leq t \leq 1$, the triangle inequality gives

$$\begin{aligned} W_2(\rho, \mu) &\leq W_2(\rho, \rho_s^*) + W_2(\rho_s^*, \rho_t^*) + W_2(\rho_t^*, \mu) \\ &\leq sW_2(\rho, \mu) + W_2(\rho_s^*, \rho_t^*) + (1-t)W_2(\rho, \mu) \end{aligned}$$

Rearranging gives the reverse bound $W_2(\rho_s^*, \rho_t^*) \geq (t-s)W_2(\rho, \mu)$. Combining the two inequalities, we see that

$$W_2(\rho_s^*, \rho_t^*) = |t-s|W_2(\rho, \mu).$$

Then we can see that the instantaneous rate of change in the curve is constant for all $t \in [0, 1]$

$$\lim_{h \rightarrow 0} \frac{W_2(\rho_{t+h}^*, \rho_t^*)}{|h|} = W_2(\rho, \mu),$$

which implies that the McCann interpolation is indeed a constant-speed geodesic.

Putting everything together, writing $v = T - \text{id}$, the logarithmic and exponential maps along this geodesic are

$$\begin{aligned} \log_\rho(\mu) &= T^{\rho \rightarrow \mu} - \text{id}, \\ \exp_\rho(tv) &= (\text{id} + tv)_\# \rho, \quad t \in [0, 1], \\ \rho_t &= \exp_\rho(t \log_\rho(\mu)). \end{aligned}$$

4.2.3 Otto calculus

Let $\mathcal{F} : \mathcal{P}(\mathbb{R}^d) \rightarrow \mathbb{R} \cup \{+\infty\}$ be a functional of interest. The *first variation* of $\mathcal{F}$, written $\delta\mathcal{F} : \mathbb{R}^d \rightarrow \mathbb{R}$, is given by

$$\lim_{\varepsilon \rightarrow 0} \frac{\mathcal{F}(\mu + \varepsilon\eta) - \mathcal{F}(\mu)}{\varepsilon} = \int \delta\mathcal{F}(\mu) \, d\eta, \quad (4.3)$$

where η is a signed measure.

We can now define gradient flows in the Wasserstein space, in analogy with gradients in $\mathbb{R}^d$. Let $(\rho_t, w_t)_{t \in [0,1]}$ satisfy the continuity equation, i.e., ρ_t defines a curve with corresponding tangent vectors w_t . If we are interested in the rate of change of a functional $\mathcal{F}$ along this path, we arrive at

$$\begin{aligned} \partial_t \mathcal{F}(\rho_t) &= \lim_{h \searrow 0} \frac{\mathcal{F}(\rho_{t+h}) - \mathcal{F}(\rho_t)}{h} \\ &\approx \lim_{h \searrow 0} \frac{\mathcal{F}(\rho_t + h\partial_t \rho_t) - \mathcal{F}(\rho_t)}{h} \\ &= \int \delta\mathcal{F}(\rho_t) \partial_t \rho_t \\ &= - \int \delta\mathcal{F}(\rho_t) \nabla \cdot (\rho_t w_t). \end{aligned}$$

Using integration by parts, we have the relation

$$\partial_t \mathcal{F}(\rho_t) = \int \langle \nabla \delta\mathcal{F}(\rho_t), w_t \rangle \, d\rho_t.$$

Note that this relation holds for any pair $(\rho_t, w_t)_{t \in [0,1]}$. If we choose $w_t = -\nabla \delta\mathcal{F}(\rho_t)$, we arrive at the relation

$$\partial_t \mathcal{F}(\rho_t) = -\|\nabla \delta\mathcal{F}(\rho_t)\|_{L^2(\rho_t)}^2,$$

which suggests that the value of the functional is decreasing along curves ρ_t which are defined via vector fields (note, gradient fields) given by $-\nabla \delta\mathcal{F}(\rho_t)$.

5 Diffusion models

The goal of this section will be to devise a stochastic flow that starts from the standard Gaussian $\gamma = \mathcal{N}(0, I)$ and terminates at the complex distribution μ . This is most commonly known as a *diffusion generative model*, due to the connections with diffusion processes and the like.

5.1 Derivation via forward and reverse OU

Diffusion models are based on the forward-reverse process ideas from Section 3.4.2. Our forward process will be given by the Ornstein–Uhlenbeck (OU) process:

$$dX_t = -X_t dt + \sqrt{2} dB_t, \quad X_0 \sim \mu, \quad (5.1)$$

where we write $\mu_t = \text{Law}(X_t)$. Multiplying (5.1) by the integrating factor e^t and applying Itô's formula (recall Theorem 3.4) gives

$$X_t = e^{-t} X_0 + \sqrt{2} \int_0^t e^{-(t-s)} dB_s.$$

By the Itô isometry (recall Theorem 3.3), the covariance of the stochastic integral above is

$$2 \int_0^t e^{-2(t-s)} ds I = (1 - e^{-2t}) I,$$

and thus the law of X_t given X_0 is

$$\text{Law}(X_t | X_0) = \mathcal{N}(e^{-t} X_0, (1 - e^{-2t}) I),$$

for any $t \geq 0$. The t -marginal μ_t is obtained by averaging the conditional Gaussian laws over $X_0 \sim \mu$:

$$\mu_t = \int \mathcal{N}(e^{-t} x, (1 - e^{-2t}) I) d\mu(x).$$

As t increases, these means contract toward zero and the conditional covariance approaches I .

Now, as our goal is to transport from $\mathcal{N}(0, I)$ to μ , we can meaningfully define the reverse process following the manipulations in Section 3.4.2. Letting $T \in (0, \infty)$, for $t \in [0, T]$, define $X_t^{\leftarrow} := X_{T-t}$; the reverse SDE is then given by

$$dX_t^{\leftarrow} = (X_t^{\leftarrow} + 2\nabla \ln \mu_{T-t}(X_t^{\leftarrow})) dt + \sqrt{2} dB_t, \quad X_0^{\leftarrow} \sim \mu_T. \quad (5.2)$$

The continuous-time diffusion sampler underlying DDPM replaces the exact initial law by the stationary distribution (i.e., the standard Gaussian) which induces some negligible initialization error:

$$dY_t^{\leftarrow} = (Y_t^{\leftarrow} + 2\nabla \ln \mu_{T-t}(Y_t^{\leftarrow})) dt + \sqrt{2} dB_t, \quad Y_0^{\leftarrow} \sim \mathcal{N}(0, I). \quad (5.3)$$

Exponential convergence to the Gaussian. Here we briefly describe how to derive the exponentially fast convergence of the OU process to its stationary distribution using a technique called *synchronous couplings*.

We begin by choosing an optimal coupling of μ and γ , resulting in $\mathbb{E}\|X_0 - G_0\|^2 = W_2^2(\mu, \gamma)$. We then draw an independent Brownian motion and drive the OU process with this same realization, resulting in the two processes

$$dX_t = -X_t dt + \sqrt{2} dB_t, \quad dG_t = -G_t dt + \sqrt{2} dB_t.$$

Note that $G_t \sim \gamma$ for every $t > 0$. Subtracting the two equations cancels the common noise and solving the resulting ODE yields

$$X_t - G_t = e^{-t}(X_0 - G_0).$$

Consequently, (X_t, G_t) is a coupling of μ_t and γ and the definition of the 2-Wasserstein distance yields

$$W_2^2(\mu_t, \gamma) \leq \mathbb{E}\|X_t - G_t\|^2 = e^{-2t}\mathbb{E}\|X_0 - G_0\|^2 = e^{-2t}W_2^2(\mu, \gamma).$$

5.2 Denoising score matching

Note that the score function $(\nabla \ln \mu_t)_{t>0}$ is not available a priori. Our first goal is to define a tractable objective function that lets us learn the score function from finite data.

Our ideal objective is the following: For any $t > 0$, optimize the neural network s_θ to the score function $\nabla \ln \mu_t$ i.e.,

$$\mathcal{L}_{\text{DDPM}}(\theta, t) := \|s_\theta(\cdot, t) - \nabla \ln \mu_t\|_{L^2(\mu_t)}^2. \quad (5.4)$$

However, μ_t is unknown to us. To obtain a tractable objective, we expand (5.4) and invoke integration-by-parts. Writing $C_t = \|\nabla \log \mu_t\|_{L^2(\mu_t)}^2$, we obtain

$$\begin{aligned} \|s_\theta(\cdot, t) - \nabla \ln \mu_t\|_{L^2(\mu_t)}^2 - C_t &= \|s_\theta(\cdot, t)\|_{L^2(\mu_t)}^2 - 2\langle s_\theta(\cdot, t), \nabla \ln \mu_t \rangle_{L^2(\mu_t)} \\ &= \|s_\theta(\cdot, t)\|_{L^2(\mu_t)}^2 - 2\langle s_\theta(\cdot, t), \nabla \mu_t \rangle_{L^2} \\ &= \|s_\theta(\cdot, t)\|_{L^2(\mu_t)}^2 + 2 \int \nabla \cdot s_\theta(\cdot, t) d\mu_t. \end{aligned}$$

Now, we make use of the formula for μ_t , namely, we write $\mu_t = \text{Law}(y_t)$ where $y_0 \sim \mu$, and $y_t = m_t y_0 + \sigma_t z$, where $z \sim \mathcal{N}(0, I)$ and $m_t = e^{-t}$ and $\sigma_t = (1 - e^{-2t})^{1/2}$. This perspective yields

$$\int \nabla \cdot s_\theta(y_t, t) d\mu_t(y_t) = \int \nabla \cdot s_\theta(m_t y_0 + \sigma_t z, t) d\gamma(z) d\mu(y_0).$$

Again by integration by parts, we can pass the divergence term to the density of the standard normal distribution, using the identity

$$\nabla \gamma(z) = -z \gamma(z).$$

Together, we obtain

$$-2\langle s_\theta(\cdot, t), \nabla \ln \mu_t \rangle_{L^2(\mu_t)} = \frac{2}{\sigma_t} \int \langle s_\theta(y_t, t), z \rangle d\gamma_{0,I}(z) d\mu(y_0).$$

As this term is an inner product, we can rewrite our original objective as a different quadratic loss (up to irrelevant additive constants):

$$\mathcal{L}_{\text{DDPM}}(\theta, t) = \|s_\theta(\cdot, t) - \nabla \ln \mu_t\|_{L^2(\mu_t)}^2 = \mathbb{E} \|s_\theta(Y_t, t) + (\sigma_t)^{-1} Z\|^2 + C, \quad (5.5)$$

where the expectation is such that $Y_0 \sim \mu$ and $Z \sim \mathcal{N}(0, I)$ are independent.

Now we are in a position to recast our ideal objective through an empirical-risk minimization, as we can approximate, for any $t > 0$, $\mathcal{L}_{\text{DDPM}}(\theta, t)$ by

$$\hat{\mathcal{L}}_{\text{DDPM}}(\theta, t) = \frac{1}{n} \sum_{i=1}^n \|s_\theta(Y_t^{(i)}, t) + (\sigma_t)^{-1} Z^{(i)}\|^2, \quad (5.6)$$

where we draw n i.i.d. samples from the target and the standard normal distribution. This procedure is succinctly summarized in Algorithm 5.

Algorithm 5 Training score-based diffusion models

Inputs: Sampler for target distribution μ , noise distribution sampler $\vartheta \in \mathcal{P}([t_{\min}, T])$

Initialize: Neural network $s_\theta : \mathbb{R}^d \times \mathbb{R}_+ \rightarrow \mathbb{R}^d$

while not converged **do**

 Sample $t \sim \vartheta$

 Sample $Z \sim \mathcal{N}(0, I)$ and $Y \sim \mu$

 Define $Y_t = m_t Y + \sigma_t Z$

 Compute $\hat{\mathcal{L}}_{\text{DDPM}}(\theta, t) = \|s_\theta(Y_t, t) + Z/\sigma_t\|^2$

 Update $\theta \leftarrow \text{optim}(\theta; \nabla_\theta \hat{\mathcal{L}}_{\text{DDPM}}(\theta))$

end while

5.3 Sampling with DDPM

Given a trained score approximation $s_\theta(x, t) \approx \nabla \log \mu_t(x)$, generating new samples amounts to numerically integrating the reverse SDE. Initializing with $Y_0^\leftarrow \sim \mathcal{N}(0, I)$ and choosing a schedule $0 = t_0 < t_1 < \dots < t_N = T$ and $h_k := t_{k+1} - t_k$, the resulting Euler–Maruyama discretization of (5.3) becomes

$$Y_{k+1}^\leftarrow = Y_k^\leftarrow + h_k \left(Y_k^\leftarrow + 2s_\theta(Y_k^\leftarrow, T - t_k) \right) + \sqrt{2h_k} \xi_k, \quad \xi_k \stackrel{\text{iid}}{\sim} \mathcal{N}(0, I).$$

It is possible to instead integrate the linear term exactly while freezing the score over the step; this results in an exponential-Euler update

$$Y_{k+1}^\leftarrow = e^{h_k} Y_k^\leftarrow + 2(e^{h_k} - 1)s_\theta(Y_k^\leftarrow, T - t_k) + \sqrt{e^{2h_k} - 1} \xi_k.$$

5.4 The probability flow ODE and its properties

Following Section 3.4.1, notice that we can write the Fokker–Planck equation (5.2) in the form of a continuity equation with velocity given by $x + \nabla \log \mu_{T-t}(x)$.

$$\dot{X}_t^{\leftarrow} = X_t^{\leftarrow} + \nabla \log \mu_{T-t}(X_t^{\leftarrow}). \quad (5.7)$$

As before, the ODE has the same t -marginals as the reverse SDE when initialized from μ_T , just different paths. The ODE described in (5.7) is known in the machine learning community as the “probability flow ODE”. As before, if we apply a standard forward Euler discretization, we obtain

$$Z_{k+1} = Z_k + h_k \left(Z_k + s_\theta(Z_k, T - t_k) \right), \quad Z_0 \sim \mathcal{N}(0, I).$$

Since the ODE has no Brownian motion term, it can also be integrated with standard higher-order ODE solvers (e.g., RK4).

Perhaps surprisingly, the probability flow ODE was first studied by [KM12] in a completely different context. Their goal was to generalize Caffarelli’s contraction theorem [Caf00] to more general distributions (recall Section 2.6.1). Kim and Milman [KM12] proposed to instead analyze the *flow map* of the probability flow ODE: let $T_{\text{KM}}^{T, \mu}$ denote the flow map generated by (5.7), which transports μ_T to μ ; then formally define $T_{\text{KM}}^\mu := \lim_{T \rightarrow \infty} T_{\text{KM}}^{T, \mu}$, which then transports $\mu_\infty = \gamma = \mathcal{N}(0, I)$ to μ .

The next two theorems describe intriguing mathematical properties of this map which stand similarly to those of optimal transport maps. Before stating the results we require the following assumption:

$$\nabla^2 \log Q_s \left[\frac{d\mu}{d\gamma} \right] (\cdot) = I + \nabla^2 \log q_s^\mu(\cdot) \preceq \theta_s I, \quad \text{and} \quad \sup_{t \geq 0} \int_t^\infty \theta_u du < \infty, \quad (\Theta)$$

where q_s^μ is the density, and Q_s is the Ornstein–Uhlenbeck semigroup.

A first result gives a bound on the Lipschitz constant of the Kim–Milman heat flow map under suitable conditions on the measure.

Theorem 5.1. [KM12; MS23] Suppose that μ satisfies (Θ) . Then T_{KM}^μ is Lipschitz with constant

$$\text{Lip}(T_{\text{KM}}^\mu) \leq \exp\left(\int_0^\infty \theta_s ds\right).$$

The following corollary provides a concrete example where we have a bound on the Lipschitz constant.

Corollary 5.2. [BP25] Suppose $\mu \propto \exp(-V)$, where $V = W + H$ with W α -strongly convex

and H is a smooth L -Lipschitz perturbation, with $\alpha > 0$. For $s > 0$, a valid bound in (Θ) is

$$\theta_s = \frac{1 - \alpha}{\alpha(e^{2s} - 1) + 1} + \frac{e^{2s}L^2}{(\alpha(e^{2s} - 1) + 1)^2} + \frac{2Le^{2s}}{(\alpha(e^{2s} - 1) + 1)^{3/2}\sqrt{e^{2s} - 1}},$$

and therefore

$$\text{Lip}(\mathsf{T}_{\text{KM}}^\mu) \leq \alpha^{-1/2} \exp\left(\frac{L^2}{2\alpha} + \frac{2L}{\sqrt{\alpha}}\right).$$

Notice that if $L = 0$ (i.e., μ is strongly log-concave), we recover the same Lipschitz constant as for optimal transport maps.

Our next result concerns the stability of these maps as a function of the target measure. Unlike with optimal transport maps, merely satisfying the assumptions of Theorem 5.1 is sufficient for a near-Lipschitz stability bound in Wasserstein.⁴

Theorem 5.3 (Near-Lipschitz stability of the Kim–Milman map [CEP26]). *Suppose μ satisfies (Θ) , and let $\nu \in \mathcal{P}_2(\mathbb{R}^d)$. There are finite constants $c_\mu, c'_\mu > 0$, depending only on μ , such that*

$$\|\mathsf{T}_{\text{KM}}^\mu - \mathsf{T}_{\text{KM}}^\nu\|_{L^2(\gamma)} \leq c_\mu W_2(\mu, \nu) \log\left(2 + \frac{c'_\mu}{W_2(\mu, \nu)}\right).$$

⁴Importantly, the assumptions of Theorem 5.1 do not imply that the map is bi-Lipschitz.

6 Dynamic transport via neural ODEs

We now turn our attention to a generative modeling family based on (neural) ordinary differential equations known commonly as flow matching [LCBHN+22] (also known as *stochastic interpolants* [AVE22] or *rectified flows* [LGL22]).

6.1 Flow matching

Recall that $(\rho_t, w_t)_{t \in [0,1]} \in \mathfrak{C}(\rho, \mu)$ satisfy the continuity equation if

$$\partial_t \rho_t = -\nabla \cdot (w_t \rho_t), \quad \rho_0 = \rho, \rho_1 = \mu. \quad (6.1)$$

At the level of the individual particles, this means that $X_t \sim \rho_t$ is such that

$$\dot{X}_t = w_t(X_t), \quad X_0 \sim \rho, \quad (6.2)$$

and $X_1 \sim \mu$.

We want to develop a procedure that allows us to generate new samples from μ , or transport samples from ρ to μ . Based on the above ODE, we will want to simulate

$$\dot{x}_t = w_\theta(x_t, t) \quad x_0 \sim \rho,$$

where $w_\theta : \mathbb{R}^d \times [0,1] \rightarrow \mathbb{R}^d$ is a neural network that is to be trained using finite data. With enough samples and training, etc., we suspect that simulating the above ODE will result in valid samples.

Perhaps the most ideal algorithm is to fit our vector field of interest w_t with the neural network w_θ by minimizing the following objective given ρ_t :

$$\min_{\theta} \int_0^1 \int \|w_\theta(z, t) - w_t(z)\|^2 d\rho_t(z) dt. \quad (6.3)$$

Unfortunately, this approach is intractable as the pair $(\rho_t, w_t)_{t \in [0,1]}$ is unknown a priori.

6.1.1 Algorithm derivation and pseudocode

To continue, we will first instantiate a particular path.

Fix $y \sim \mu$, and let $\rho_t(\cdot|y)$ be a path such that

$$\rho_0(\cdot|y) = \rho(\cdot), \quad \rho_1(\cdot|y) = \delta_y.$$

In words, the path completely degenerates to a point mass at time $t = 1$, completely forgetting anything about the source measure ρ .

Note that the original path can itself be factored such that this same conditional path appears:

$$\rho_t(z) = \int \rho_t(z|y) d\mu(y).$$

Now, suppose for each $y \in \text{supp}(\mu)$, there exists a pair $(\rho_t(\cdot|y), w_t(\cdot|y))_{t \in [0,1]} \in \mathfrak{C}$, i.e., there exist conditional path–vector-field pairs that each satisfy the continuity equation. What is the natural way to aggregate the vector fields? Returning to the continuity equation, we have

$$\begin{aligned}\partial_t \rho_t(z) &= \partial_t \int \rho_t(z|y) \, d\mu(y) \\ &= \int \partial_t \rho_t(z|y) \, d\mu(y) \\ &= \int -\nabla \cdot (\rho_t(z|y) w_t(z|y)) \, d\mu(y) \\ &= -\nabla \cdot \left(\int w_t(z|y) \rho_t(z|y) \, d\mu(y) \right) \\ &= -\nabla \cdot \left(\rho_t(z) \int w_t(z|y) \frac{\rho_t(z|y)}{\rho_t(z)} \, d\mu(y) \right).\end{aligned}$$

The calculation above shows that one valid choice of velocity field is

$$w_t(z) = \int w_t(z|y) \frac{\rho_t(z|y)}{\rho_t(z)} \, d\mu(y). \quad (6.4)$$

This is to say: given a conditional vector field family, the aggregate vector field for the full path can be naturally seen as an appropriate averaging of many conditional vector fields.

To obtain a tractable algorithm, we substitute (6.4) into (6.3), resulting in

$$\min_{\theta} \int_0^1 \int \left\| w_{\theta}(z, t) - \int w_t(z|y) \frac{\rho_t(z|y)}{\rho_t(z)} \, d\mu(y) \right\|^2 \, d\rho_t(z) \, dt.$$

By the law of iterated expectations, up to a constant, we can write the above loss as

$$\min_{\theta} \int_0^1 \iint \left\| w_{\theta}(x_t, t) - w_t(x_t|y) \right\|^2 \, d\rho_t(x_t|y) \, d\mu(y) \, dt. \quad (6.5)$$

These brief calculations show that if $\rho_t(\cdot|y)$ and $w_t(\cdot|y)$ are known to the practitioner by design, then we can efficiently minimize an upper bound on the flow-matching loss and still fit w_{θ} .

Recall that the basis of this whole construction is a path $\rho_t(\cdot|y)$ that equals the source distribution at time $t = 0$ (that is, $\rho_0(\cdot|y) = \rho(\cdot)$) and degenerates to $\rho_1(\cdot|y) = \delta_y$ at time $t = 1$. The most natural path is the following *straight-line path*:

$$X_t := (1 - t)X + ty, \quad (6.6)$$

where $X \sim \rho$ is a random variable but $y \in \mathbb{R}^d$ is fixed, and $t \in [0, 1]$. For example, if $\rho = \mathcal{N}(0, I)$, then $X_t \sim \rho_t(\cdot|y) = \mathcal{N}(ty, (1 - t)^2 I)$. In this case, different paths at the level of the particles give rise to different expressions for the resulting Gaussian density.

For other source measures ρ , a closed-form expression for $\rho_t(\cdot|y)$ might not be possible to obtain. Thankfully, this is not an issue as we only require that the companion velocity

field be explicitly defined. Indeed, for the straight-line path, we obtain $\dot{X}_t = y - X$. This results in the objective

$$\mathcal{L}_{\text{FM}}(\theta) := \int_0^1 \iint \|w_\theta(t, x_t) - (y - x)\|_2^2 d\rho(x) d\mu(y) dt. \quad (6.7)$$

This relation from (6.5) holds since, at the level of the particles, $\rho_t(\cdot|y) = \text{Law}(X_t)$ which is obtained by merely sampling from ρ , given $y \sim \mu$. In practice, this objective is minimized by sampling minibatches of data from the source and target distributions, and randomly drawing $t \sim \text{Unif}([0, 1])$.

6.2 Alternative derivation via stochastic interpolants

Letting ρ, μ denote two arbitrary probability distributions, write $(X_0, X_1) \sim \rho \otimes \mu$ (though in principle this can be any coupling), and draw $Z \sim \mathcal{N}(0, I)$ independently of ρ and μ . Consider a process $(X_t)_{t \in [0, 1]}$ defined as

$$X_t = I_t(X_0, X_1) + \sigma_t Z,$$

where $I_t : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}^d$ and we impose the conditions

$$I_0(x_0, x_1) = x_0, \quad I_1(x_0, x_1) = x_1, \quad \sigma_0 = \sigma_1 = 0.$$

For instance, a suitable parametrization would be

$$I_t(x_0, x_1) = \alpha_t x_0 + \beta_t x_1, \quad \alpha_0 = \beta_1 = 1, \alpha_1 = \beta_0 = 0,$$

for coefficients $(\alpha_t, \beta_t)_{t \in [0, 1]}$. Let $\gamma_t := \text{Law}(X_t)$, we want to write down a corresponding velocity field w_t such that the pair $(\gamma_t, w_t)_{t \in [0, 1]} \in \mathfrak{C}^{\rho\mu}$. To this end, letting h be a smooth test function, we have by chain rule

$$\begin{aligned} \int h \partial_t \gamma_t dx &= \partial_t \mathbb{E} h(X_t) \\ &= \partial_t \mathbb{E} h(I_t(X_0, X_1) + \sigma_t Z) \\ &= \mathbb{E} \langle \nabla h(X_t), \dot{I}_t(X_0, X_1) + \dot{\sigma}_t Z \rangle \\ &= \mathbb{E} \langle \nabla h(X_t), \mathbb{E}_{\rho \otimes \mu \otimes \gamma} [\dot{I}_t(X_0, X_1) + \dot{\sigma}_t Z | X_t] \rangle, \end{aligned}$$

where the last line follows by the law of iterated expectation. Writing

$$w_t(x_t) = \mathbb{E}_{\rho \otimes \mu \otimes \gamma} [\dot{I}_t(X_0, X_1) + \dot{\sigma}_t Z | X_t = x_t], \quad (6.8)$$

we conclude via integration by parts that $(\gamma_t, w_t)_{t \in [0, 1]}$ satisfy the continuity equation. We then regress a neural network onto this velocity field using the loss

$$\mathcal{L}_{\text{SI}}(\theta) = \int_0^1 \mathbb{E} \|w_\theta(t, X_t) - (\dot{\alpha}_t X_0 + \dot{\beta}_t X_1 + \dot{\sigma}_t Z)\|^2 dt,$$

which again leverages the law of iterated expectations. See Algorithm 6.

Algorithm 6 Stochastic interpolants (aka flow matching)

Inputs: Sampler for source distribution ρ , sampler for target distribution μ

Initialize: Neural network $w_\theta : \mathbb{R}^d \times [0, 1] \rightarrow \mathbb{R}^d$

while not converged **do**

 Sample $(X, Y, Z) \sim \rho \otimes \mu \otimes \gamma$

 Sample $t \sim \text{Unif}((0, 1))$

 Define $X_t = \alpha_t X + \beta_t Y + \sigma_t Z$

 Compute $\hat{\mathcal{L}}_{\text{SI}}(\theta) = \|w_\theta(t, X_t) - (\dot{\alpha}_t X + \dot{\beta}_t Y + \dot{\sigma}_t Z)\|^2$

 Update $\theta \leftarrow \text{optim}(\theta; \nabla_\theta \hat{\mathcal{L}}_{\text{SI}}(\theta))$

end while

6.2.1 Mini-batch sampling

Notice how in the derivation of stochastic interpolants, we did not use the fact that $(X_0, X_1) \sim \rho \otimes \mu$. In fact, one can choose any coupling $\tilde{\pi} \in \Gamma(\rho, \mu)$ and the derivation still goes through. However, $\tilde{\pi}$ is often not known to the practitioner (otherwise there are other methods to use). Also, if the dataset size is massive, even obtaining the (discrete) entropic coupling can be potentially problematic.

Still, [PBHDEA+23; TFMH+23] leveraged this idea to improve the training dynamics of flow matching at the level of *minibatches* used during training. The modified pseudocode is given in Algorithm 7.

Algorithm 7 Minibatch-OT sampling for flow matching

Inputs: Sampler for source distribution ρ , sampler for target distribution μ

Initialize: Neural network $w_\theta : \mathbb{R}^d \times [0, 1] \rightarrow \mathbb{R}^d$

while not converged **do**

 Sample $X^{(1)}, \dots, X^{(b)} \sim \rho$

 Sample $Y^{(1)}, \dots, Y^{(b)} \sim \mu$

 Compute minibatch OT coupling $\hat{\pi}$ between $(X^{(1)}, \dots, X^{(b)}; Y^{(1)}, \dots, Y^{(b)})$

 Sample $t^{(1)}, \dots, t^{(b)} \sim \text{Unif}((0, 1))$

 Define $X_t^{(i)} = (1 - t^{(i)})X^{(i)} + t^{(i)}Y^{\hat{\pi}(i)}$

 Compute $\hat{\mathcal{L}}_{\text{MB}}(\theta) = b^{-1} \sum_{i=1}^b \|w_\theta(t^{(i)}, X_t^{(i)}) - (Y^{\hat{\pi}(i)} - X^{(i)})\|^2$

 Update $\theta \leftarrow \text{optim}(\theta; \nabla_\theta \hat{\mathcal{L}}_{\text{MB}}(\theta))$

end while

Despite the persistence of the curse of dimensionality, the use of minibatch couplings has been shown to substantially *decrease* the number of function evaluations (i.e., the number of ODE steps when integrating) before obtaining a reasonable sample.

6.3 Flow matching beyond Euclidean spaces

6.3.1 Riemannian flow matching

We now describe how to modify the flow matching algorithm to take place over sufficiently nice Riemannian manifolds.

Interpolations on manifolds. Recall the exponential and logarithmic maps from our preliminary discussion on geodesics and manifolds from Section 4.2.1. To extend these ideas for the purposes of flow matching, we draw independent endpoints $x_0 \sim \rho$ and $x_1 \sim \mu$, choose the constant-speed geodesic between them as our path of interest. Concretely, for $t \in [0, 1]$, we choose

$$x_t = \exp_{x_0}(t \log_{x_0}(x_1)) = \exp_{x_1}((1-t) \log_{x_1}(x_0)),$$

$$\dot{x}_t = \frac{\log_{x_t}(x_1)}{1-t} := w_t(x_t \mid x_1).$$

Deriving the objective function. For $(X_0, X_1) \sim \rho \otimes \mu$, define X_t to be the random geodesic interpolation above and write $\rho_t = \text{Law}(X_t)$. Its marginal velocity is given by

$$w_t(x) = \mathbb{E}[\dot{X}_t \mid X_t = x] \in T_x \mathcal{M}.$$

To understand the path ρ_t , we differentiate against a smooth test function and arrive at

$$\begin{aligned} \frac{d}{dt} \mathbb{E}[h(X_t)] &= \mathbb{E}[\langle \nabla_g h(X_t), \dot{X}_t \rangle_{g(X_t)}] \\ &= \int \langle \nabla_g h(x), w_t(x) \rangle_{g(x)} d\rho_t(x), \end{aligned}$$

where recall $\langle \cdot, \cdot \rangle_{g(\cdot)}$ is the Riemannian metric inner product, and ∇_g denotes the Riemannian gradient. This implies a *Riemannian* continuity equation

$$\partial_t \rho_t + \text{div}_g(\rho_t w_t) = 0, \quad \rho_0 = \rho, \quad \rho_1 = \mu.$$

From here it is easy to convince oneself that the ideal objective is

$$\int_0^1 \int \|w_\theta(x, t) - w_t(x)\|_{g(x)}^2 d\rho_t(x) dt,$$

though, as in flow matching, this is not tractable a priori from data. Pattern-matching, one can verify that instead the right loss to use is

$$\mathcal{L}_{\text{RFM}}(\theta) = \int_0^1 \int \|w_\theta(x_t, t) - \dot{x}_t\|_{g(x_t)}^2 d\rho(x_0) d\mu(x_1) dt,$$

where $g(x_t)$ is the metric of the tangent space at the intermediate point x_t .

Example: Euclidean space. For $\mathcal{M} = \mathbb{R}^d$, $\exp_x(v) = x + v$ and $\log_x(y) = y - x$. Hence

$$x_t = (1-t)x_0 + tx_1, \quad w_t(x \mid x_1) = \frac{x_1 - x}{1-t}, \quad \dot{x}_t = x_1 - x_0.$$

Example: The sphere. Suppose $\mathcal{M} = \mathbb{S}^d \subset \mathbb{R}^{d+1}$ with the induced metric, so $T_x \mathbb{S}^d = \{v : x^\top v = 0\}$ and $\|v\|_{g(x)} = \|v\|_2$. If we define $\alpha = \arccos(x_0^\top x_1) \in (0, \pi)$, we compute

$$\begin{aligned} x_t &= \frac{\sin((1-t)\alpha)}{\sin \alpha} x_0 + \frac{\sin(t\alpha)}{\sin \alpha} x_1, \\ \dot{x}_t &= \frac{\alpha}{\sin \alpha} (-\cos((1-t)\alpha) x_0 + \cos(t\alpha) x_1). \end{aligned}$$

Sampling. At generation time, start from $z_0 \sim \rho$ and integrate $\dot{z}_t = w_\theta(z_t, t)$. As an example, a Riemannian-Euler step would be given by

$$z_{t+h} = \exp_{z_t}(hw_\theta(z_t, t)),$$

provided the exponential map is easy to compute.

6.3.2 Wasserstein flow matching

The RFM framework applies to any Riemannian manifold $\mathcal{M}$. An important instance, introduced in [HPPA24], sets $\mathcal{M} = \mathcal{P}_2(\mathbb{R}^d)$: the Wasserstein space.⁵ This is natural when each data point is itself a probability distribution—for example, a 3D shape observed as a point cloud, or a cell’s gene-expression microenvironment.

Let $\mathbf{p}_0, \mathbf{p}_1 \in \mathcal{P}(\mathcal{P}_2(\mathbb{R}^d))$ be a source and target *distribution over distributions*. For a pair $\mu \sim \mathbf{p}_0$ and $\nu \sim \mathbf{p}_1$, the conditional interpolating path is the McCann interpolation

$$\mu_t := (T_t^{\mu \rightarrow \nu})_\# \mu, \quad T_t^{\mu \rightarrow \nu} := (1-t) \text{id} + t T_\star^{\mu \rightarrow \nu},$$

which is the constant-speed geodesic in Wasserstein space. The corresponding velocity field is

$$v_t = (T_\star^{\mu \rightarrow \nu} - \text{id}) \circ (T_t^{\mu \rightarrow \nu})^{-1},$$

so that $\dot{X}_t = v_t(X_t)$ for $X_0 \sim \mu$ traces $X_t \sim \mu_t$. Substituting into the RFM objective yields the *Wasserstein flow matching* (WFM) loss

$$\mathcal{L}_{\text{WFM}}(\theta) = \int_0^1 \iint \|f_\theta(\mu_t, t) - v_t\|_{L^2(\mu_t)}^2 \, \mathbf{d}\mathbf{p}_0(\mu) \, \mathbf{d}\mathbf{p}_1(\nu) \, dt. \quad (6.9)$$

⁵Of course, the Wasserstein space over suitably regular measures is only formally the same as a Riemannian manifold, though technically it is not.

Bures–Wasserstein setting. When $\mathbf{p}_0$ and $\mathbf{p}_1$ are distributions over Gaussians, the McCann interpolation remains Gaussian, $\mu_t = \mathcal{N}(m_t, \Sigma_t)$, with

$$m_t = (1-t)m_\mu + tm_\nu, \quad \Sigma_t = T_t \Sigma_\mu T_t, \quad T_t = (1-t)I + tC^{\mu \rightarrow \nu},$$

where $C^{\mu \rightarrow \nu} = \Sigma_\mu^{-1/2}(\Sigma_\mu^{1/2}\Sigma_\nu\Sigma_\mu^{1/2})^{1/2}\Sigma_\mu^{-1/2}$. The tangent space at $\mu = \mathcal{N}(m, \Sigma)$ in the BW manifold consists of pairs $(a, S) \in \mathbb{R}^d \times \mathbb{S}^d$ with BW tangent norm

$$\|(a, S)\|_{\text{BW}(\mu)}^2 = \|a\|^2 + \text{Tr}(S^2\Sigma).$$

Along the interpolation, the Riemannian velocities are

$$\dot{m}_t = m_\nu - m_\mu, \quad \dot{\Sigma}_t^{\text{BW}} = (C^{\mu \rightarrow \nu} - I)((1-t)I + tC^{\mu \rightarrow \nu})^{-1}.$$

These differ from the Euclidean time-derivative $\dot{\Sigma}_t^E$; the two are related by the Lyapunov identity $\dot{\Sigma}_t^E = \dot{\Sigma}_t^{\text{BW}}\Sigma_t + \Sigma_t\dot{\Sigma}_t^{\text{BW}}$. Matching only the Euclidean derivative (a naïve baseline) fails to respect the manifold geometry and produces degenerate covariance estimates in practice. The BW-FM loss is

$$\mathcal{L}_{\text{BW}}(\theta) = \int_0^1 \iint \|f_\theta^{\text{BW}}((m_t, \Sigma_t), t) - (\dot{m}_t, \dot{\Sigma}_t^{\text{BW}})\|_{\text{BW}(\mu_t)}^2 \mathbf{d}\mathbf{p}_0(\mu) \mathbf{d}\mathbf{p}_1(\nu) dt, \quad (6.10)$$

where $f_\theta^{\text{BW}} : \mathbb{R}^d \times \mathbb{S}_{++}^d \times [0, 1] \rightarrow \mathbb{R}^d \times \mathbb{S}^d$.

General distributions and equivariant architectures. When μ and ν are general, $T_\star^{\mu \rightarrow \nu}$ has no closed form; one instead approximates it via entropic optimal transport on point-cloud realizations using Sinkhorn’s algorithm. The architecture for f_θ is then constrained by the geometry of the problem.

A point cloud $X = \{x_1, \dots, x_n\} \subset \mathbb{R}^d$ represents an empirical distribution $\hat{\mu} = n^{-1} \sum_i \delta_{x_i}$, which is invariant to any relabelling of its particles. The predicted velocity field must therefore be *permutation equivariant*: for any $\sigma \in S_n$,

$$f_\theta^{\text{PC}}(\sigma X, t) = \sigma f_\theta^{\text{PC}}(X, t).$$

A *transformer* (self-attention blocks without positional encodings) satisfies this requirement by construction. Letting

$$\alpha_{ij}(X) = \exp(\langle x_i, x_j \rangle / \sqrt{d}) / \sum_{\ell=1}^n \exp(\langle x_i, x_\ell \rangle / \sqrt{d})$$

denote the attention weights, the output at position i is

$$\text{Attn}(X)_i = \sum_{j=1}^n \alpha_{ij}(X) x_j V,$$

which satisfies $\text{Attn}(\sigma X) = \sigma \text{Attn}(X)$ for any $\sigma \in S_n$, since both the weights α_{ij} and the values $x_j V$ permute consistently. With an entropic OT map $\hat{T}_\varepsilon^{\mu \rightarrow \nu}$ computed by Sinkhorn, the discretised WFM loss is

$$\int_0^1 \iint \|f_\theta^{\text{PC}}(\hat{X}_t, t) - (\hat{T}_\varepsilon^{\mu \rightarrow \nu}(X) - X)\|_2^2 \mathrm{d}\mathbf{p}_0(\mu) \mathrm{d}\mathbf{p}_1(\nu) \mathrm{d}t,$$

where $\hat{X}_t = (1 - t)X + t\hat{T}_\varepsilon^{\mu \rightarrow \nu}(X)$ is the discretized McCann interpolation.

7 Schrödinger bridge problem on path space

7.1 A minimization problem on path space

We first recall some preliminary notation regarding probability measures on path space.

Let $\text{path} := \mathcal{C}([0, 1], \mathbb{R}^d)$ be the space of paths with $X_t : \text{path} \rightarrow \mathbb{R}^d$ given by the canonical mapping $X_t(h) = h_t$ for any $h \in \text{path}$ and any $t \in [0, 1]$. For a path measure $P \in \mathcal{P}(\text{path})$ and any $t \in [0, 1]$, we write $P_t := (X_t)_\# P \in \mathcal{P}(\mathbb{R}^d)$ for the t^{th} marginal of P . Similarly, for $s, t \in [0, 1]$, we can define the joint probability measure $P_{st} := (X_s, X_t)_\# P$. Restrictions to $[0, t]$ are written $P_{[0, t]}$ (which restricts P to $\mathcal{C}([0, t], \mathbb{R}^d)$) and we define regular conditional probabilities for the law of a path given its value at time t , which we denote $P_{|t}$ (which is permitted since path is a Polish space).

For $\varepsilon > 0$, let R denote the law of the reversible Brownian motion on $\mathbb{R}^d$ with diffusion coefficient $\sqrt{\varepsilon}$, with the Lebesgue measure as the initial distribution.⁶ The joint distribution of the initial and final positions under R is then given by

$$dR_{01}(x, y) = (2\pi\varepsilon)^{-d/2} \exp(-\frac{1}{2}\|x - y\|^2/\varepsilon) dx dy.$$

With our reference measure in hand, we arrive at the Schrödinger bridge problem defined over path measures where the initial and final marginals are given by μ and ν , which we assume to have finite second moment with finite entropy:

$$\min_{P \in \mathcal{P}(\text{path})} \varepsilon \text{KL}(P \parallel R) \quad \text{s.t.} \quad P_0 = \mu, P_1 = \nu. \quad (\text{SB})$$

Let $P^{\mu\nu}$ be the minimizer of (SB); the goal of this section is to characterize and estimate $P^{\mu\nu}$. (We omit the ε subscript for path measures to not conflate with the time-indexing.)

7.1.1 Equivalence to entropic optimal transport

If we now apply the chain rule for the KL divergence by conditioning on times $t = 0, 1$, the objective function (SB) decomposes into

$$\varepsilon \text{KL}(P \parallel R) = \varepsilon \text{KL}(P_{01} \parallel R_{01}) + \varepsilon \mathbb{E}_{P_{01}}[\text{KL}(P_{|01} \parallel R_{|01})].$$

Under the assumption that μ and ν have finite entropy, it can be shown that the first term on the right-hand side is equivalent to the objective for the entropic optimal transport problem in (2.3). Indeed, if $\pi = P_{01} \in \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d)$, then this term can be written as

$$\varepsilon \text{KL}(P_{01} \parallel R_{01}) = \varepsilon \mathcal{H}(\pi) + \frac{\varepsilon d}{2} \log(2\pi\varepsilon) + \frac{1}{2} \iint \|x - y\|^2 d\pi(x, y).$$

⁶Note that this is not in fact a probability distribution.

If we have the finite-entropy condition, then indeed this can be recast as the EOT objective, resulting in

$$\min_{\pi \in \Gamma(\mu, \nu)} \varepsilon \text{KL}(\pi \parallel R_{01}) = \varepsilon(\mathcal{H}(\mu) + \mathcal{H}(\nu)) + \frac{\varepsilon d}{2} \log(2\pi\varepsilon) + \text{OT}_\varepsilon(\mu, \nu).$$

Minimizing, we see that $P_{01} = \pi_\varepsilon^{\mu\nu}$. If we now choose the measure P such that the conditional measure $P_{|01}$ is the same as $R_{|01}$, then the second term vanishes. This can be accomplished by connecting the source and target by a Brownian bridge.

Therefore, the objective function in (SB) is minimized when $P_{01}^{\mu\nu} = \pi_\varepsilon^{\mu\nu}$ and when P is a mixture of Brownian bridges whose initial and final points are drawn according to $\pi_\varepsilon^{\mu\nu}$:

$$P^{\mu\nu} = \iint R(\cdot | X_0 = x, X_1 = y) d\pi_\varepsilon^{\mu\nu}(x, y). \quad (7.1)$$

7.1.2 Computing entropic interpolations

We now compute the distribution $P_t^{\mu\nu}$, sometimes called the *entropic interpolation* between μ and ν . We compute directly from (7.1):

$$\begin{aligned} P_t^{\mu\nu}(z) &= \iint \mathcal{N}(z | (1-t)x + ty, \varepsilon t(1-t)I) d\pi_\varepsilon^{\mu\nu}(x, y) \\ &= (2\pi\varepsilon)^{d/2} \left(\int e^{g_\varepsilon^{\mu\nu}(y)/\varepsilon} \mathcal{N}(z | y, (1-t)\varepsilon I) d\nu(y) \right) \\ &\quad \times \left(\int e^{f_\varepsilon^{\mu\nu}(x)/\varepsilon} \mathcal{N}(z | x, t\varepsilon I) d\mu(x) \right) \\ &= (2\pi\varepsilon)^{d/2} \mathfrak{H}_{(1-t)\varepsilon}[e^{g_\varepsilon^{\mu\nu}/\varepsilon}\nu](z) \mathfrak{H}_{t\varepsilon}[e^{f_\varepsilon^{\mu\nu}/\varepsilon}\mu](z). \end{aligned} \quad (7.2)$$

where we write the *heat semigroup* acting on a measure Q as

$$Q \mapsto \mathfrak{H}_s[Q](z) := (2\pi s)^{-d/2} \int e^{-\frac{1}{2s}\|\xi - z\|^2} dQ(\xi).$$

Let's now write

$$\mathfrak{g}_t(z) := \mathfrak{H}_{(1-t)\varepsilon}[\exp(g_\varepsilon^{\mu\nu}/\varepsilon)\nu](z), \quad \mathfrak{f}_t(z) := \mathfrak{H}_{t\varepsilon}[\exp(f_\varepsilon^{\mu\nu}/\varepsilon)\mu](z).$$

7.2 Formulation as a drift minimization problem

A third formulation of the Schrödinger bridge problem is in terms of drifts. Indeed, as R is a path measure, an application of Girsanov's theorem yields the following optimization problem.

Proposition 7.1. (SB) has the same minimizers as the following drift problem:

$$\min_{\rho_t, w_t} \int_0^1 \int \frac{1}{2} \|w_t(z)\|^2 d\rho_t(z) dt \quad \text{s.t.} \quad \begin{cases} \partial_t \rho_t + \nabla \cdot (\rho_t w_t) = \frac{\varepsilon}{2} \Delta \rho_t, \\ \rho_0 = \mu, \quad \rho_1 = \nu. \end{cases} \quad (7.3)$$

Proof. Let R be the Brownian motion reference process, and let P be the process of interest (with drift w_t and marginal density ρ_t). By Girsanov's theorem [LG16], we have

$$\begin{aligned} \varepsilon \text{KL}(P \parallel R) &= \varepsilon \mathbb{E}_P \log \left(\frac{dP}{dR} \right) \\ &= \int_0^1 \int \frac{\varepsilon}{2} \|w_t(z)\|^2 d\rho_t(z) dt + \sqrt{\varepsilon} \mathbb{E}_P \left[\int_0^1 \langle w_t(X_t), dB_t^P \rangle \right] + \varepsilon \mathcal{H}(\mu) \\ &= \int_0^1 \int \frac{1}{2} \|w_t(z)\|^2 d\rho_t(z) dt + \varepsilon \mathcal{H}(\mu), \end{aligned}$$

The term in expectation vanishes, and minimizing over all quantities completes the proof. $\square$

We now derive the optimal Schrödinger bridge drift.

Theorem 7.2. The optimal solution to (7.3) is attained at $(\rho_t^\varepsilon, b_t^\varepsilon)$, where

$$b_t^\varepsilon(z) = \varepsilon \nabla \log \mathfrak{H}_{(1-t)\varepsilon}[\exp(g_\varepsilon^{\mu\nu}/\varepsilon)\nu](z) = \varepsilon \nabla \log g_t(z),$$

and ρ_t^ε is given by (7.2).

Proof. We proceed by introducing Lagrange multipliers in the form of $\lambda_t : \mathbb{R}^d \rightarrow \mathbb{R}$; these are meant to encode the Fokker–Planck equation as a constraint. Dropping the arguments (of the functions and densities) for convenience, the resulting optimization problem is

$$\int_0^1 \int \frac{1}{2} \|w_t\|^2 d\rho_t dt + \int_0^1 \int \lambda_t (\partial_t \rho_t + \nabla \cdot (\rho_t w_t) - \frac{\varepsilon}{2} \Delta \rho_t) dz dt.$$

The only difference between this case and the derivation of the Benamou–Brenier formulation is the additional Laplacian term, for which two applications of integration by parts yield

$$\int \lambda_t (\Delta \rho_t) dz = \int (\Delta \lambda_t) \rho_t dz.$$

Combining all the terms, we arrive at

$$\int_0^1 \int \left(\frac{1}{2} \|w_t\|^2 - \langle w_t, \nabla \lambda_t \rangle - \partial_t \lambda_t - \frac{\varepsilon}{2} \Delta \lambda_t \right) \rho_t dz dt + \int \lambda_1 d\nu - \int \lambda_0 d\mu.$$

Optimizing in w_t results in

$$w_t^* = b_t^\varepsilon = \nabla \lambda_t,$$

and substituting this into the objective gives

$$\int_0^1 \int \left(-\frac{1}{2} \|\nabla \lambda_t\|^2 - \partial_t \lambda_t - \frac{\varepsilon}{2} \Delta \lambda_t \right) \rho_t \, dz \, dt + \int \lambda_1 \, d\nu - \int \lambda_0 \, d\mu.$$

We now choose the family of Lagrange multipliers such that the first integrand is always zero:

$$-\frac{1}{2} \|\nabla \lambda_t\|^2 - \partial_t \lambda_t - \frac{\varepsilon}{2} \Delta \lambda_t = 0.$$

Now choose the following parametrization, called the *Hopf–Cole* transform

$$\lambda_t(z) = \varepsilon \log(\eta_t(z)),$$

for a smooth function η_t . Under this parametrization, we have that

$$\partial_t \lambda_t = \varepsilon \frac{\partial_t \eta_t}{\eta_t}, \quad \nabla \lambda_t = \varepsilon \frac{\nabla \eta_t}{\eta_t}, \quad \Delta \lambda_t = \varepsilon \left(\frac{\Delta \eta_t}{\eta_t} - \frac{\|\nabla \eta_t\|^2}{\eta_t^2} \right).$$

Performing the necessary substitutions, we arrive at the *backward* heat equation

$$\partial_t \eta_t + \frac{\varepsilon}{2} \Delta \eta_t = 0.$$

Since time is between $t = 0$ and $t = 1$, η_t is written as

$$\eta_t = \mathfrak{H}_{(1-t)\varepsilon}[\eta_1],$$

where η_1 is a terminal condition that remains to be chosen. Note that $\lambda_0 = \varepsilon \log \eta_0 = \varepsilon \log \mathfrak{H}_\varepsilon[\eta_1]$. A candidate choice is $\eta_1 = \exp(g_\varepsilon^{\mu\nu}/\varepsilon)\nu$, which yields

$$\lambda_0 = \varepsilon \log(\eta_0) = -f_\varepsilon^{\mu\nu} - \frac{\varepsilon d}{2} \log(2\pi\varepsilon).$$

Putting everything together, we see that our objective is then

$$\begin{aligned} \int \lambda_1 \, d\nu - \int \lambda_0 \, d\mu &= \int (\varepsilon \log(\exp(g_\varepsilon^{\mu\nu}/\varepsilon)\nu)) \, d\nu - \int (-f_\varepsilon^{\mu\nu} - \frac{\varepsilon d}{2} \log(2\pi\varepsilon)) \, d\mu \\ &= \varepsilon \mathcal{H}(\nu) + \int f_\varepsilon^{\mu\nu} \, d\mu + \int g_\varepsilon^{\mu\nu} \, d\nu + \frac{\varepsilon d}{2} \log(2\pi\varepsilon). \end{aligned}$$

Adding the initial entropy term $\varepsilon \mathcal{H}(\mu)$ from Girsanov's formula, we recover the path-space value in (SB):

$$\begin{aligned} \varepsilon \text{KL}(\mathbf{P}^{\mu\nu} \parallel \mathbf{R}) &= \varepsilon \mathcal{H}(\mu) + \int \lambda_1 \, d\nu - \int \lambda_0 \, d\mu \\ &= \varepsilon (\mathcal{H}(\mu) + \mathcal{H}(\nu)) + \text{OT}_\varepsilon(\mu, \nu) + \frac{\varepsilon d}{2} \log(2\pi\varepsilon). \end{aligned}$$

□

7.3 Algorithms

7.3.1 Plug-in estimation of Schrödinger bridges

Further developing the expression for b_t^ε , we obtain

$$\begin{aligned} b_t^\varepsilon(z) &= (1-t)^{-1} \left(-z + \frac{\int y e^{(g_\varepsilon^{\mu\nu}(y) - \frac{1}{2(1-t)} \|z-y\|^2)/\varepsilon} d\nu(y)}{\int e^{(g_\varepsilon^{\mu\nu}(y) - \frac{1}{2(1-t)} \|z-y\|^2)/\varepsilon} d\nu(y)} \right) \\ &=: (1-t)^{-1} (-z + \nabla \varphi_{(1-t)\varepsilon}^{\mu\nu}(z)). \end{aligned} \quad (7.4)$$

From this, we can see the drift is expressed as a function of a particular entropic Brenier map, namely, the entropic Brenier map between ρ_t^ε and ν where the regularization parameter changes with time, namely $(1-t)\varepsilon$.

Let $X^{(1)}, \dots, X^{(n)} \sim \mu$ and $Y^{(1)}, \dots, Y^{(m)} \sim \nu$, and let $\mu_n := n^{-1} \sum_{i=1}^n \delta_{X^{(i)}}$, and similarly $\nu_m := m^{-1} \sum_{j=1}^m \delta_{Y^{(j)}}$. Let $(\hat{f}, \hat{g}) \in \mathbb{R}^n \times \mathbb{R}^m$ be the optimal entropic potentials associated with $\text{OT}_\varepsilon(\mu_n, \nu_m)$, which can be computed efficiently via Sinkhorn's algorithm [Cut13; PC19]. A natural plug-in estimator is then

$$\begin{aligned} \hat{b}_t(z) &:= \varepsilon \nabla \log \mathfrak{H}_{(1-t)\varepsilon}[\exp(\hat{g}/\varepsilon) \nu_m] \\ &= (1-t)^{-1} \left(-z + \frac{\sum_{j=1}^m Y^{(j)} \exp((\hat{g}_j - \frac{1}{2(1-t)} \|z - Y^{(j)}\|^2)/\varepsilon)}{\sum_{j=1}^m \exp((\hat{g}_j - \frac{1}{2(1-t)} \|z - Y^{(j)}\|^2)/\varepsilon)} \right) \\ &=: (1-t)^{-1} (-z + \nabla \hat{\varphi}_{(1-t)\varepsilon}(z)) \end{aligned} \quad (7.5)$$

8 Few- and one-step generative algorithms

8.1 Flow map matching

Recall that for a family of vector fields $(w_t)_{t \in [0,1]}$, we write the solution map between times $s, t \in [0, 1]$ as

$$\Phi_{s,t}(x_s) = x_s + \int_s^t w_r(\Phi_{s,r}(x_s)) \, dr = x_t,$$

which solves the ODE $\dot{x}_t = w_t(x_t)$.

We collect some easy-to-derive properties below.

Proposition 8.1. *For a suitably smooth flow map, the following hold for $s, t, \tau \in [0, 1]$:*

$$\begin{aligned}\Phi_{t,\tau} \circ \Phi_{s,t} &= \Phi_{s,\tau}, \\ \partial_t \Phi_{s,t}(x) - w_t(\Phi_{s,t}(x)) &= 0, \\ \partial_s \Phi_{s,t} + \langle w_s, \nabla \Phi_{s,t} \rangle &= 0.\end{aligned}$$

Recall that for $\rho, \mu \in \mathcal{P}_2(\mathbb{R}^d)$, we use the following notation to denote stochastic interpolants:

$$x_t = I_t(x_0, x_1, Z) := \alpha_t x_0 + \beta_t x_1 + \sigma_t Z.$$

8.1.1 Distillation

Suppose we already have a pre-trained model in the form of a family of vector fields $(b_t)_{t \in [0,1]}$ and we want to learn the flow map between arbitrary times, or simply improve the discretization and require fewer function evaluations for generation. This is known as “distilling” the model.

Let $\Phi_{s,t}^\theta$ be some parametrization of the flow map, which we will always assume to be of the following tractable form

$$\Phi_{s,t}^\theta(x) = x + (t - s)v^\theta(x, s, t),$$

where $v^\theta : \mathbb{R}^d \times [0, 1]^2 \rightarrow \mathbb{R}^d$ is a parametric neural network.

We will now write down some loss functions as a result of the computations in Proposition 8.1. The first is known as *Lagrangian map distillation*.

$$\mathcal{L}_{\text{LMD}}(\theta) := \mathbb{E} \|\partial_t \Phi_{s,t}^\theta(I_s(X_0, X_1, Z)) - b_t(\Phi_{s,t}^\theta(I_s(X_0, X_1, Z)))\|^2, \quad (8.1)$$

where the expectation is over $(s, t) \sim \omega \in \mathcal{P}([0, 1]^2)$, and $(X_0, X_1, Z) \sim \rho \otimes \mu \otimes \gamma$.

The next loss uses the other optimality condition, called *Eulerian map distillation*.

$$\mathcal{L}_{\text{EMD}}(\theta) := \mathbb{E} \|\partial_s \Phi_{s,t}^\theta(I_s(X_0, X_1, Z)) + \langle b_s(I_s(X_0, X_1, Z)), \nabla \Phi_{s,t}^\theta(I_s(X_0, X_1, Z)) \rangle\|^2. \quad (8.2)$$

8.1.2 Flow map matching loss

Flow map matching was introduced by [BAVE24]. Recall the formula for w_t from (6.8):

$$w_t(\cdot) = \mathbb{E}[\dot{I}_t(X_0, X_1, Z) \mid X_t = \cdot],$$

where the expectation is over $(X_0, X_1, Z) \sim \rho \otimes \mu \otimes \gamma$. If we make use of the law of conditional expectations, we arrive at the following natural loss function that *directly* learns a flow map $\Phi_{s,t}$.

$$\mathcal{L}_{\text{FMM}}(\theta) := \mathbb{E} \left\{ \|\partial_t \Phi_{s,t}^\theta(\Phi_{t,s}^\theta(I_t)) - \dot{I}_t\|^2 + \|\Phi_{s,t}^\theta(\Phi_{t,s}^\theta(I_t)) - I_t\|^2 \right\}. \quad (8.3)$$

The second term encourages the flow map to be invertible, and the first asks the instantaneous velocities to match the usual loss in (6.5). Another aspect to keep in mind is that $\omega \in \mathcal{P}([0, 1]^2)$ ought to be symmetric, as we want to enforce invertibility.

8.1.3 Eulerian estimation (aka consistency models)

[AAP: TODO]

8.2 An optimal transport perspective on generative drifting models

Drifting models

We adopt the standard notation from drifting models. Let $\epsilon \sim \gamma = \mathcal{N}(0, I)$ and let f_θ be the one-step mapping we hope to optimize. Let $\mathcal{F} : \mathcal{P}(\mathbb{R}^d) \rightarrow \mathbb{R}$ be a functional of interest, and write $q \mapsto \mathbb{W}\mathcal{F}[q](\cdot) : \mathbb{R}^d \rightarrow \mathbb{R}^d$ to be the Wasserstein gradient of $\mathcal{F}$ at q . Finally, let $\text{sg}(\cdot)$ denote the `stopgrad`($\cdot$) operation, which removes the argument from, e.g., PyTorch’s computational graph.

Let $\theta^{(k)}$ be the parameters of the generative model for iterations $k = 0, 1, \dots, K_{\max}$, writing $q_k := q_{\theta^{(k)}}$. Drifting models iterate between the following steps:

1. Compute $\mathbb{W}\mathcal{F}[q_k]$ and, for a parameter $\eta > 0$, define $x \mapsto g^{(k)}(x) := x - \eta \mathbb{W}\mathcal{F}[q_k](x)$
2. Compute the loss $\mathcal{L}(\theta) = \mathbb{E}_{\epsilon \sim \gamma} \|f_\theta(\epsilon) - \text{sg}(g^{(k)}(f_{\theta^{(k)}}(\epsilon)))\|^2$
3. Update the parameters $\theta^{(k+1)} \leftarrow \text{optim}(\theta^{(k)}; \nabla_\theta \mathcal{L}(\theta)|_{\theta^{(k)}})$.

Prior works have motivated this objective through the lens of Wasserstein gradient flows, where η is supposed to play the role of a “time discretization parameter.” However, to our knowledge, all works use the parameter choice $\eta = 1$. The use of the `sg` operator is also curious.

Choices of $\mathcal{F}(q)$ include $\text{KL}(q \parallel \mu)$, $\text{MMD}^2(q, \mu)$, and the Sinkhorn divergence $S_\epsilon(q, \mu)$.

Minimum Wasserstein estimators (MWEs)

For a parametric model q_θ and a target distribution μ , minimum Wasserstein estimators (MWEs) are solutions to the following objective

$$\min_{\theta} \frac{1}{2} W_2^2(q_\theta, \mu) \quad q_\theta = (f_\theta)_\# \gamma. \quad (8.4)$$

For many years, such a methodology (as well as its approximation) was deemed unfavorable in the machine learning community as they require backpropagating through exact or approximate solvers.

MWEs as idealized drifting models

Let $\mathcal{W}(q) := \frac{1}{2} W_2^2(q, \mu)$ for μ fixed. It is straightforward to derive the following relation

$$\nabla \mathcal{W}(q)(x) = x - T_0^{q \rightarrow \mu}(x),$$

where $T_0^{q \rightarrow \mu}$ is the optimal transport map from q to μ . If we choose $\eta = 1$ as suggested from prior work, then the drifting model loss becomes

$$\begin{aligned} \mathcal{L}(\theta) &= \mathbb{E}_{\epsilon \sim \gamma} \|f_\theta(\epsilon) - \text{sg}(f_{\theta^{(k)}}(\epsilon) - (f_{\theta^{(k)}}(\epsilon) - T_0^{q_k \rightarrow \mu}(f_{\theta^{(k)}}(\epsilon)))\|^2 \\ &= \mathbb{E}_{\epsilon \sim \gamma} \|f_\theta(\epsilon) - \text{sg}(T_0^{q_k \rightarrow \mu}(f_{\theta^{(k)}}(\epsilon)))\|^2, \end{aligned}$$

where we recall $q_k = q_{\theta^{(k)}}$. The update proceeds as normal.

To relate this to MWEs, we rewrite the objective using the Monge formulation of optimal transport

$$\min_{\theta} \frac{1}{2} W_2^2(q_\theta, \mu) = \min_{\theta} \min_T \int \frac{1}{2} \|x - T(x)\|^2 dq_\theta(x) \quad \text{s.t.} \quad T_\# q_\theta = \mu.$$

From here, we define the objective

$$\mathcal{A}(\theta, h) := \mathbb{E}_{\epsilon \sim \gamma} \|f_\theta(\epsilon) - h(\epsilon)\|^2 \quad \text{s.t.} \quad h_\# \gamma = \mu.$$

Fix $\theta^{(k)}$. Optimizing in h , we want to solve

$$\min_h \mathcal{A}(\theta^{(k)}, h) = \min_h \mathbb{E}_{\epsilon \sim \gamma} \|f_{\theta^{(k)}}^{(k)}(\epsilon) - h(\epsilon)\|^2 \quad \text{s.t.} \quad h_\# \gamma = \mu.$$

Notice that for any $h : \mathbb{R}^d \rightarrow \mathbb{R}^d$, it holds that

$$\frac{1}{2} W_2^2(q_{\theta^{(k)}}, \mu) \leq \frac{1}{2} \mathbb{E}_{\epsilon \sim \gamma} \|f_{\theta^{(k)}}^{(k)}(\epsilon) - h(\epsilon)\|^2,$$

by choosing the suboptimal coupling $(f_{\theta^{(k)}}^{(k)}, h)_\# \gamma$. If $q_{\theta^{(k)}}$ has a density, then minimizing over all functions h implies that the minimization is attained at

$$h_k^* = T_0^{q_k \rightarrow \mu} \circ f_{\theta^{(k)}}.$$

This follows from the definition of $\frac{1}{2}W_2^2(q_{\theta^{(k)}}, \mu)$ and the pushforward identity.

Now if h_k^* is fixed, and we optimize in θ , then

$$\min_{\theta} \mathcal{A}(\theta, h_k^*) = \min_{\theta} \mathbb{E}_{\epsilon \sim \gamma} \|f_{\theta}(\epsilon) - h_k^*(\epsilon)\|^2 = \min_{\theta} \mathbb{E}_{\epsilon \sim \gamma} \|f_{\theta}(\epsilon) - T_0^{q_k \rightarrow \mu}(f_{\theta^{(k)}}(\epsilon))\|^2,$$

which is precisely the update obtained from the drifting model formulation. Here, the `stopgrad` aspect naturally emerges from viewing minimum Wasserstein estimation as a two-stage optimization problem.

Moreover, one can also view the `stopgrad` mechanism from the envelope theorem. Indeed, we can write

$$\nabla_{\theta} \frac{1}{2}W_2^2(q_{\theta}, \mu) = \frac{1}{2} \nabla_{\theta} \mathcal{A}(\theta, h_{\theta}^*),$$

for h_{θ}^* which solves the nested optimization problem, where the derivative on the right holds h_{θ}^* fixed. The factor $\frac{1}{2}$ accounts for the unscaled squared loss in $\mathcal{A}$.

References

- [Amo22] Brandon Amos. “On amortizing convex conjugates for optimal transport”. In: *arXiv preprint arXiv:2210.12153* (2022) (cit. on pp. 24, 25).
- [AVE22] Michael S Albergo and Eric Vanden-Eijnden. “Building normalizing flows with stochastic interpolants”. In: *arXiv preprint arXiv:2209.15571* (2022) (cit. on p. 76).
- [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. 38).
- [AXK17] Brandon Amos, Lei Xu, and J Zico Kolter. “Input convex neural networks”. In: *International conference on machine learning*. PMLR. 2017, pp. 146–155 (cit. on p. 25).
- [BAVE24] Nicholas M. Boffi, Michael S. Albergo, and Eric Vanden-Eijnden. “Flow map matching”. In: *arXiv preprint arXiv:2406.07507* (2024) (cit. on p. 90).
- [BP25] Giovanni Brigati and Francesco Pedrotti. “Heat flow, log-concavity, and Lipschitz transport maps”. In: *Electronic Communications in Probability* 30 (2025), pp. 1–12 (cit. on p. 74).
- [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).
- [Caf00] Luis A Caffarelli. “Monotonicity properties of optimal transportation and the FKG and related inequalities”. In: *Communications in Mathematical Physics* 214.3 (2000), pp. 547–563 (cit. on p. 74).
- [CEP26] Sinho Chewi, Katharina Eichinger, and Aram-Alexandre Pooladian. “Near-Lipschitz stability of the Kim–Milman flow map”. In: *arXiv preprint arXiv:2606.23383* (2026) (cit. on p. 75).
- [CGP21] Yongxin Chen, Tryphon T Georgiou, and Michele Pavon. “Stochastic control liaisons: Richard Sinkhorn meets Gaspard Monge on a Schrödinger bridge”. In: *Siam Review* 63.2 (2021), pp. 249–313 (cit. on p. 65).
- [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. 48).

- [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. 39).
- [Csi75] I. Csiszár. “ I -divergence geometry of probability distributions and minimization problems”. In: *Ann. Probability* 3 (1975), pp. 146–158 (cit. on p. 30).
- [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. 39).
- [Cut13] Marco Cuturi. “Sinkhorn distances: Lightspeed computation of optimal transport”. In: *Advances in Neural Information Processing Systems* 26 (2013) (cit. on pp. 31, 88).
- [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. 21).
- [FSVA+19] Jean Feydy, Thibault Séjourné, François-Xavier Vialard, Shun-ichi Amari, Alain Trounev, 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. 44).
- [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. 16).
- [GM96] Wilfrid Gangbo and Robert J McCann. “The geometry of optimal transportation”. In: (1996) (cit. on p. 12).
- [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).
- [HPPA24] Doron Haviv, Aram-Alexandre Pooladian, Dana Pe’er, and Brandon Amos. “Wasserstein Flow Matching: Generative modeling over families of distributions”. In: *arXiv preprint arXiv:2411.00698* (2024) (cit. on p. 81).

- [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. 20).
- [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. 39).
- [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, 21).
- [KM12] Young-Heon Kim and Emanuel Milman. “A generalization of Caffarelli’s contraction theorem via (reverse) heat flow”. In: *Mathematische Annalen* 354.3 (2012), pp. 827–862 (cit. on p. 74).
- [LCBHN+22] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. “Flow matching for generative modeling”. In: *arXiv preprint arXiv:2210.02747* (2022) (cit. on p. 76).
- [LG16] Jean-François Le Gall. *Brownian motion, martingales, and stochastic calculus*. Springer, 2016 (cit. on p. 86).
- [LGL22] Xingchao Liu, Chengyue Gong, and Qiang Liu. “Flow straight and fast: Learning to generate and transfer data with rectified flow”. In: *arXiv preprint arXiv:2209.03003* (2022) (cit. on p. 76).
- [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. 16).
- [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. 16).
- [Mér26] Quentin Mérigot. “Sharp stability of Brenier maps via quantitative regularity of potentials”. In: (2026) (cit. on p. 17).
- [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. 39).
- [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. 30).

- [MS23] Dan Mikulincer and Yair Shenfeld. “On the Lipschitz properties of transportation along heat flows”. In: *Geometric Aspects of Functional Analysis: Israel Seminar (GAFA) 2020-2022*. Springer. 2023, pp. 269–290 (cit. on p. 74).
- [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. 30).
- [PBHDEA+23] Aram-Alexandre Pooladian, Heli Ben-Hamu, Carles Domingo-Enrich, Brandon Amos, Yaron Lipman, and Ricky TQ Chen. “Multisample flow matching: Straightening flows with minibatch couplings”. In: *arXiv preprint arXiv:2304.14772* (2023) (cit. on p. 79).
- [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 pp. 21, 88).
- [PNW21] Aram-Alexandre Pooladian and Jonathan Niles-Weed. “Entropic estimation of optimal transport maps”. In: *arXiv preprint arXiv:2109.12004* (2021) (cit. on pp. 37, 41).
- [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. 37).
- [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. 31).
- [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. 22).
- [TFMH+23] Alexander Tong, Kilian Fatras, Nikolay Malkin, Guillaume Huguet, Yanlei Zhang, Jarrod Rector-Brooks, Guy Wolf, and Yoshua Bengio. “Improving and generalizing flow-based generative models with minibatch optimal transport”. In: *arXiv preprint arXiv:2302.00482* (2023) (cit. on p. 79).