AS.110.503 : Undergraduate Research in Mathematics

Wasserstein Gradient Flow

James Guo

December 16, 2025

Contents

I Wasserstein Space **1**
I.1 Probability Space 1
I.2 Wasserstein Distance 2
I.3 Wasserstein Space 4

II Optimal Transport **7**
II.1 Geodesics in Wasserstein Space 7
II.2 From Gradient Descent to Gradient Flow 10
II.3 First Variation and JKO Scheme 13

III Interactive Particle System **15**
III.1 Towards IPS Simulations 15
III.2 Applications of Interactive Particle System 17
III.3 Attention and Self-Attention 18

References **20**

Acknowledgements:

This collection of notes serves as a summary of the materials related to topics of gradient flow of measure and optimal transport. We will assume adequate backgrounds in measure theory and probability theory in going through the materials.
This also serves as the report for the undergraduate research on such topic guided by Dr. Lu, Fei at *Johns Hopkins University* for Fall 2025 semester.

I Wasserstein Space

In this chapter, we want to extend the idea of metric space to a metric representation over "measures," which can be achieved with a similar method.

I.1 Probability Space

Definition I.1.1. Measure Space.
We define a **measure space** $(\Omega, \mathscr{F}, \mathbb{P})$ composed of:

- Ω as a set,
- $\mathscr{F}$ as a σ-algebra over Ω, and
- $\mathbb{P} : \mathscr{F} \to \mathbb{R}_{\geq 0}$ as the measure such that $\mu(\varnothing) = 0$ and being σ-additive.

The measure space is called a **probability space** if we additionally assume $\mu(\Omega) = 1$. ⌟

Unless otherwise stated, we will assume that $(\Omega, \mathscr{F}, \mathbb{P})$ will be a probability space for the notation of this text.

In many probability problems, a crucial problem is to consider the problem with multiple probability spaces and attempting to consider them at the same times. Thereby, a handy tool could be the product space.

Definition I.1.2. Product Space.
Suppose we have two probability space $(\Omega_1, \mathscr{F}_1, \mathbb{P}_1)$ and $(\Omega_2, \mathscr{F}_2, \mathbb{P}_2)$, the product probability space is defined as $(\Omega_1 \times \Omega_2, \mathscr{F}_1 \otimes \mathscr{F}_2, \mathbb{P})$, where:

- $\Omega_1 \times \Omega_2$ is the natural set product,
- endowed with the product σ-algebra that:

$$\mathscr{F}_1 \otimes \mathscr{F}_2 := \sigma \left\{ A \times B : A \in \mathscr{F}_1, B \in \mathscr{F}_2 \right\},$$

- and with the product measure $\mu \times \nu$ satisfying that:

$$\mathbb{P}(A \times B) = \mathbb{P}_1(A) \times \mathbb{P}_2(B) \qquad \text{for all } A \in \mathscr{F}_1 \text{ and } B \in \mathscr{F}_2.$$

⌟

Remark I.1.3. Although the definition of the product measure is only defined on *finite* unions of rectangles, it is well defined over $\Sigma := \{A \times B : A \in \mathscr{F}_1, B \in \mathscr{F}_2\}$ as a algebra, hence by the **Carathéodory theorem**, there exists a unique extension to the σ-algebra generated by Σ, namely $\sigma(\Sigma)$. ⌟

The rigorous definition of a probability space allows us to define more structures over the probability space.

Definition I.1.4. Random Variables.
Let $(\Omega, \mathscr{F}, \mathbb{P})$ be a probability space and let $(S, \mathscr{S})$ be a measurable space, a random variable is a **measurable function** $X : \Omega \to S$ such that for all $B \in \mathscr{S}$, we have $X^{-1}(B) \in \mathscr{F}$.
Moreover, the distribution of X induces a probability measure μ over $B \in (S, \mathscr{S})$, namely by:

$$\mu(B) = \mathbb{P}\big(X^{-1}(B)\big) = \mathbb{P}(X \in B).$$

⌟

Imagine constructing different random variables could result in completely different probability measure. Given a measurable space, there could be different ways to cast a probability measure on it, and we will then attempt to find ways measuring the differences between these probability measures.

I.2 Wasserstein Distance

Before getting to the definition of *distance* over a measurable space $(\Omega, \mathscr{F})$, we would first be thinking about a similar idea like the product measure in this case, *i.e.*, the coupling of different measures.

Definition I.2.1. Coupling.
Consider the probability spaces $(X, \mathscr{X}, \mu)$ and $(Y, \mathscr{Y}, \nu)$, the coupling of the measures (denoted $\Gamma(\mu, \nu)$) is a probability measure γ defined over the product measurable space $(X \times Y, \mathscr{X} \otimes \mathscr{Y})$ such that for any $A \in \mathscr{X}$ and $B \in \mathscr{Y}$, we have:

$$\gamma(A \times Y) = \mu(A) \qquad \text{and} \qquad \gamma(X \times B) = \nu(B).$$

⌟

Remark I.2.2. The product measure (defined in Definition I.1.2) is a coupling. ⌟

Now, we consider a simple example of some possible coupling on the Lebesgue measure over $[0, 1]$.

Example I.2.3. Coupling of $[0, 1]$.
Let $\big([0, 1], \mathcal{B}([0, 1])\big)$ be the interval $[0, 1]$ with Borel measurable sets, consider the Lebesgue measure m, we can define some of the coupling of m with itself as follows:

- The Lebesgue measure over $\mathbb{R}^2$ is a coupling.
- We can also consider the point mass $\gamma(x, y) = \delta_{x,y}$, where the mass is uniformly distributed over the set $\{(x, x) : x \in [0, 1]\}$. ◊

Now, given two probability measures, we therefore can define the **Wasserstein p-distance** so we know how much the measures differ.

To make the argument rigid, we consider the general probability measures.

Remark I.2.4. Probability Measures.
Given a measurable space $(\Omega, \mathscr{F})$, we denote the set of all probability measures (satisfying Definition I.1.1) as $\mathcal{P}(\Omega)$. ⌟

Now, let's consider this distance in this space.

Definition I.2.5. Wasserstein *p*-distance.
Given $\mu, \nu \in \mathcal{P}(\Omega)$ and $p \in [1, \infty]$, the Wasserstein p-distance between μ and ν is defined as:

$$W_p(\mu, \nu) = \inf_{\gamma \in \Gamma(\mu,\nu)} \left(\int d(x,y)^p d\gamma(x,y) \right)^{\frac{1}{p}},$$

where we format $\Gamma(\mu, \nu)$ as the coupling of the two measures. ⌟

A very simple example could be dealing with atomic measures, *i.e.*, little point masses.

Example I.2.6. Consider that we have two data points $x_1, x_2 \in \mathbb{R}^d$ and $\mu = \delta_{x_1}$ and $\nu = \delta_{x_2}$. There will be one single coupling $\delta_{(x_1,x_2)}$ as all the masses are at x_1 and x_2, respectively, for μ and ν. Therefore, what we can think of is that:

$$W_p(\mu, \nu) = \left(d(x_1, x_2)^p\right)^{\frac{1}{p}} = d(x_1, x_2),$$

which is exactly the distance between the two data points. ♢

To consider some setup that is mildly harder, we will be considering the Wasserstein 2 distances between normal distributions.

Example I.2.7. Wasserstein Distances between Normal Distributions.
Let $\mu_0 = \mathcal{N}(m_0, \sigma_0^2)$ and $\mu_1 = \mathcal{N}(m_1, \sigma_1^2)$ be two probability measures on $\mathbb{R}$.
Recall the definition for the W_2 distance:

$$W_2^2(\mu_0, \mu_1) = \inf_{\gamma \in \Gamma(\mu_0,\mu_1)} \int_{\mathbb{R}^2} |x - y|^2 d\gamma(x,y),$$

where $\Gamma(\mu_0, \mu_1)$ is the set of couplings with marginals μ_0 and μ_1.
Any joint Gaussian distribution with marginals μ_0 and μ_1 has the form:

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} m_0 \\ m_1 \end{pmatrix}, \begin{pmatrix} \sigma_0^2 & c\,\sigma_0\sigma_1 \\ c\,\sigma_0\sigma_1 & \sigma_1^2 \end{pmatrix} \right).$$

for some correlation $c \in [-1, 1]$.

Compute the expectation of squared difference:

$$\begin{aligned}\mathbb{E}[|X-Y|^2] &= \mathbb{E}[(X-Y)^2] = \mathbb{E}[X^2] + \mathbb{E}[Y^2] - 2\mathbb{E}[XY] \\ &= \sigma_0^2 + m_0^2 + \sigma_1^2 + m_1^2 - 2(c\,\sigma_0\sigma_1 + m_0 m_1) = (m_0 - m_1)^2 + \sigma_0^2 + \sigma_1^2 - 2c\,\sigma_0\sigma_1.\end{aligned}$$

To find the minimal possible value, minimize over c:

$$W_2^2(\mu_0, \mu_1) = \min_{c \in [-1,1]} \left[(m_0 - m_1)^2 + \sigma_0^2 + \sigma_1^2 - 2c\,\sigma_0\sigma_1 \right]$$

The minimum is achieved when $c = 1$, so:

$$W_2^2(\mu_0, \mu_1) = (m_0 - m_1)^2 + (\sigma_0 - \sigma_1)^2,$$

and thus:

$$W_2(\mu_0, \mu_1) = \sqrt{(m_0 - m_1)^2 + (\sigma_0 - \sigma_1)^2}.$$

◊

Then, we will show that this is a metric, which allows us with the further constructions here. However, we would need to postpone a little bit by defining the space for the metric acting on.

I.3 Wasserstein Space

Definition I.3.1. Wasserstein Space.
The Wasserstein space of order p is defined as:

$$\mathcal{P}_p(\Omega) = \left\{ \mu \in P(\Omega) : W_p(\mu, \nu) < +\infty \right\},$$

where ν is a measure that has all mass on some $x_0 \in \mathbb{R}^n$, and the choice of x_0 is arbitrary. ⌟

In particular, we are making this finite p-th moment to prevent the transport cost being $+\infty$, which degenerates the distance in the space. Then, we can consider this modified space in which the Wasserstein distance is a metric space.

Proposition I.3.2. $\left(\mathcal{P}_p(\Omega), W_p\right)$ is a metric space.

To show the triangular inequality, we would need an additional lemma.

Lemma I.3.3. Gluing Lemma.
Suppose we are given probability spaces $(X, \mathscr{X}, \mu)$, $(Y, \mathscr{Y}, \lambda)$, and $(Z, \mathscr{Z}, \nu)$, and coupling γ_1 of μ and λ, γ_2 of λ and ν. Then, there exists a probability measure π on $X \times Y \times Z$ such that:

(i) For any $A \in \mathscr{X}$ and $B \in \mathscr{Y}$, $\pi(A \times B \times Z) = \gamma_1(A \times B)$.

(ii) For any $B \in \mathscr{Y}$ and $C \in \mathscr{Z}$, $\pi(X \times B \times C) = \gamma_2(B \times C)$.

Proof. Let $f : X \times Y \to [0,1]$ and $g : Y \times Z \to [0,1]$ be measurable functions (densities) for γ_1 and γ_2 with respect to reference measures, or consider the general case where these couplings exist as probability measures.

Disintegration theory guarantees that we can write $\gamma_1(dx,dy) = \mu(dx)\gamma_1^x(dy)$, where γ_1^x is a regular conditional probability (the conditional law of y given x). Likewise, $\gamma_2(dy,dz) = \lambda(dy)\gamma_2^y(dz)$.

Define π on $X \times Y \times Z$ by:

$$\pi(dx,dy,dz) = \mu(dx)\gamma_1^x(dy)\gamma_2^y(dz)$$

That is: sample x from μ, y from γ_1^x, z from γ_2^y. Then, we would wish to discuss each marginal:

- Marginal on $X \times Y$:

$$\begin{aligned}\pi(A \times B \times Z) &= \int_A \mu(dx) \int_B \gamma_1^x(dy) \int_Z \gamma_2^y(dz) = \int_A \mu(dx) \int_B \gamma_1^x(dy) \\ &= \gamma_1(A \times B)\end{aligned}$$

- Marginal on $Y \times Z$:

$$\begin{aligned}\pi(X \times B \times C) &= \int_X \mu(dx) \int_B \gamma_1^x(dy) \int_C \gamma_2^y(dz) = \int_B \left(\int_X \mu(dx)\gamma_1^x(dy) \right) \int_C \gamma_2^y(dz) \\ &= \lambda(B) \int_C \gamma_2^y(dz) = \gamma_2(B \times C)\end{aligned}$$

Hence, π has the required marginals. Then, consider the marginal on $X \times Z$:

$$\pi(A \times Y \times C) = \int_A \mu(dx) \left[\int_Y \gamma_1^x(dy)\gamma_2^y(C) \right]$$

Hence, the construction is a coupling between μ and ν. □

After this result, we would be returning to the proof that $(\mathcal{P}_p(\Omega), W_p)$ is a metric space.

Proof. We would verify each conditions for a metric.

(i) Positivity: For any coupling $\gamma \in \Gamma(\mu,\nu)$, $d(x,y)^p \geq 0$, therefore the integral is non-negative. Taking the infimum does not change this:

$$W_p(\mu,\nu) \geq 0$$

(ii) Definiteness: Suppose $\mu = \nu$. Consider the coupling γ_0 where all mass is put on the diagonal, i.e., $\gamma_0(A \times B) = \mu(A \cap B)$. Then,

$$\int_{M \times M} d(x,y)^p \, d\gamma_0(x,y) = \int_M d(x,x)^p \, d\mu(x) = 0$$

so $W_p(\mu,\mu) = 0$.

Conversely, suppose $W_p(\mu,\nu) = 0$. Then there exists couplings γ_n such that

$$\int d(x,y)^p \, d\gamma_n(x,y) \to 0$$

Since $d(x,y)^p = 0$ if and only if $x = y$, it follows that for any continuous bounded function f,

$$\int f(x)\,d\mu(x) = \int f(x)\,d\gamma(x,y) = \int f(y)\,d\gamma(x,y) = \int f(y)\,d\nu(y)$$

so $\mu = \nu$.

(iii) Symmetry: By definition,

$$W_p(\mu,\nu) = \left(\inf_{\gamma \in \Gamma(\mu,\nu)} \int d(x,y)^p\,d\gamma(x,y) \right)^{1/p}$$

Note that if $\gamma \in \Gamma(\mu,\nu)$, then its "transpose" $\gamma'(A \times B) := \gamma(B \times A)$ is in $\Gamma(\nu,\mu)$, and $d(x,y) = d(y,x)$. Thus,

$$W_p(\mu,\nu) = W_p(\nu,\mu)$$

(iv) Triangle inequality: Let $\mu,\nu,\lambda \in \mathcal{P}_p(M)$. Let $\gamma_1 \in \Gamma(\mu,\lambda)$ and $\gamma_2 \in \Gamma(\lambda,\nu)$. By the Gluing Lemma (see e.g. Villani, *Optimal Transport*), there exists a probability measure π on $M \times M \times M$ with marginals γ_1 on (x,y) and γ_2 on (y,z).

Define a plan $\gamma \in \Gamma(\mu,\nu)$ by its marginal on (x,z):

$$\gamma(A \times C) = \int_M \mathbf{1}_A(x)\mathbf{1}_C(z)\,d\pi(x,y,z)$$

Now, by the metric property of d and Minkowski's inequality:

$$d(x,z) \leq d(x,y) + d(y,z)$$

Therefore,

$$d(x,z)^p \leq [d(x,y) + d(y,z)]^p$$

Take expectations:

$$\int d(x,z)^p\,d\gamma(x,z) = \int d(x,z)^p\,d\pi(x,y,z) \leq \int [d(x,y) + d(y,z)]^p\,d\pi(x,y,z)$$

By Minkowski's inequality,

$$\left(\int [d(x,y) + d(y,z)]^p\,d\pi \right)^{1/p} \leq \left(\int d(x,y)^p\,d\gamma_1(x,y) \right)^{1/p} + \left(\int d(y,z)^p\,d\gamma_2(y,z) \right)^{1/p}$$

Therefore,

$$W_p(\mu,\nu) \leq W_p(\mu,\lambda) + W_p(\lambda,\nu)$$

as desired.

□

II Optimal Transport

As we have developed a “metric space” structure, we have the distances and a space well-defined. Now, we can push forward on the structures, *i.e.*, we consider the movements of the measures inside the space, and how we can evaluate such movements.

From now on, unless otherwise specified, we consider the Wasserstein 2 distance, so we have $\mathcal{P} := (\mathcal{P}_2(M), W_2)$ denoting the probability space, that is the metric space of functions with finite second moment, equipped with the Wasserstein 2 distance.

II.1 Geodesics in Wasserstein Space

In the study of Riemmanian geometry and manifolds, when we have a manifold structures, we would be concerning the shortest path connecting between two points (known as **geodesics**), and we also want a parallel definition of geodesics in terms of the Wasserstein space.

Definition II.1.1. Geodesics.
A **geodesic** in $(\mathcal{P}_2(M), W_2)$ is a continuous curve $\{\mu_t\}_{t\in[0,1]}$ such that, for any $0 \leq s < t \leq 1$,

$$W_2(\mu_s, \mu_t) = |t-s| W_2(\mu_0, \mu_1).$$

┘

Now, let’s consider an example for geodesics.

Example II.1.2. Geodesics between Normal Distributions.
Consider the measures:

$$\mu_0 = \mathcal{N}(m_0, \sigma_0^2) \qquad \text{and} \qquad \mu_1 = \mathcal{N}(m_1, \sigma_1^2).$$

Recall from Example I.2.7, we know that:

$$W_2(\mu_0, \mu_1) = \sqrt{(m_1 - m_0)^2 + (\sigma_1 - \sigma_0)^2}.$$

For $t \in [0, 1]$, define:

$$\mu_t = \mathcal{N}(m_t, \sigma_t^2),$$

where:

$$m_t = (1-t)m_0 + tm_1, \quad \sigma_t = (1-t)\sigma_0 + t\sigma_1.$$

Now, for any $0 \leq s < t \leq 1$, the Wasserstein distance between μ_s and μ_t is:

$$\begin{aligned} W_2(\mu_s, \mu_t) &= \sqrt{(m_t - m_s)^2 + (\sigma_t - \sigma_s)^2} \\ &= \sqrt{\left(\left[(1-t)m_0 + tm_1\right] - \left[(1-s)m_0 + sm_1\right]\right)^2 + \left((1-t)\sigma_0 + t\sigma_1 - (1-s)\sigma_0 - s\sigma_1\right)^2} \\ &= \sqrt{\left((t-s)(m_1 - m_0)\right)^2 + \left((t-s)(\sigma_1 - \sigma_0)\right)^2} \\ &= |t-s|\sqrt{(m_1 - m_0)^2 + (\sigma_1 - \sigma_0)^2} \\ &= |t-s|\, W_2(\mu_0, \mu_1) \end{aligned}$$

which is exactly the definition of a constant-speed geodesic in Wasserstein space.
Hence, the path $t \mapsto \mu_t$ is a constant-speed geodesic in $(\mathcal{P}_2(\mathbb{R}), W_2)$ between μ_0 and μ_1. ◊

Theorem II.1.3. Brenier's Theorem.
Let μ and ν be probability measures on $\mathbb{R}^d$ with finite second moments, and suppose that μ is absolutely continuous with respect to Lebesgue measure.
Then there exists a unique (almost everywhere) measurable map $T : \mathbb{R}^d \to \mathbb{R}^d$ such that:

(i) $T_\# \mu = \nu$, *i.e.*, T pushes forward μ to ν, so that for every measurable $B \subseteq \mathbb{R}^d$, $\nu(B) = \mu(T^{-1}(B))$,

(ii) T minimizes the total transport cost:

$$\int_{\mathbb{R}^d} |x - T(x)|^2 \, d\mu(x)$$

among all measurable maps pushing μ to ν, and

(iii) T is the gradient of a convex function, *i.e.*,

$$T(x) = \nabla\phi(x)$$

for some convex function $\phi : \mathbb{R}^d \to \mathbb{R}$.

In particular, the push forward measure can be thought of as a optimization problem, called the *Optimal transport problem.*

Definition II.1.4. Optimal Transport Problem.
Given probability measures μ and ν on $\mathbb{R}^d$, and cost function $c(x, y)$, the **optimal transport problem** is to find the optimal solution to:

$$\min_{\gamma \in \Gamma(\mu,\nu)} \int c(x, y) \, d\gamma(x, y),$$

where $\Gamma(\mu, \nu)$ is the set of couplings with marginals μ, ν. ⌟

Proof of Theorem II.1.3 in sketch. Here, we sketch the steps to the proof the problem.

(i) **Kantorovichs Problem and Duality.**

- The Kantorovich optimal transport problem for quadratic cost admits a dual formulation:

$$\sup_{\varphi,\psi} \left\{ \int \varphi(x) d\mu(x) + \int \psi(y) d\nu(y) \ : \ \forall x, y,\ \varphi(x) + \psi(y) \leq |x-y|^2 \right\}$$

- For quadratic cost, one can write $\psi(y) = \inf_x \left\{ |x-y|^2 - \varphi(x) \right\}$, so optimal ψ is a c-transform of φ.

(ii) **Existence of Optimal Plan and Concentration on a Map.**

- By standard theory, there exists an optimal plan γ^*.
- For quadratic cost and absolutely continuous μ, the optimal plan is unique and is induced by a measurable map T; i.e., $\gamma^*(dx, dy) = \mu(dx)\delta_{T(x)}(dy)$.

(iii) **T is the Gradient of a Convex Function.**

- Show that optimal T must be monotone (cyclically monotone for quadratic cost).
- By a classical theorem (Rockafellar), any cyclically monotone map is the gradient of a convex function.
- Thus, $T = \nabla\phi$ for some convex ϕ.

(iv) **Uniqueness.**

- The convex function ϕ is unique up to an additive constant.
- The transport T is unique μ-almost-everywhere.

□

Then, with this theorem, we can think of the geodesics of measures.

Corollary II.1.5. Geodesics of Measures.
By Brenier's theorem, the optimal transport map $T : M \to M$ pushing μ_0 to μ_1 (i.e., $T_\# \mu_0 = \mu_1$) exists and is unique (almost everywhere).
Then, the geodesic is given by

$$\mu_t = ((1-t)\mathrm{Id} + tT)_\# \, \mu_0,$$

where Id is the identity map and $(\cdot)_\# \mu_0$ denotes the push-forward of μ_0.
That is, at time t, the mass originally at x under μ_0 will be moved to:

$$x_t = (1-t)x + tT(x).$$

Remark II.1.6.

- Geodesics are **constant speed shortest paths** in Wasserstein space.
- Any absolutely continuous curve $\{\mu_t\}$ in $\mathcal{P}_2(M)$ can be represented by a velocity field v_t solving the continuity equation:

$$\frac{\partial \mu_t}{\partial t} + \nabla \cdot (v_t \mu_t) = 0,$$

and along a geodesic, $v_t(x) = T(x) - x$ (constant in t).

⌟

The geodesics in Wasserstein space will be helpful for the theory of gradient flows for probability measures, *e.g.*, in diffusions and PDEs.

II.2 From Gradient Descent to Gradient Flow

In classical gradient descent, a point x_t in Euclidean space evolves according to the steepest descent of a function $f(x)$:

$$\frac{dx_t}{dt} = -\nabla f(x_t).$$

The gradient descent method helps us to find local minimum by getting towards the direction of the fastest descent direction.

Similarly, we want to apply such ideas of how to decrease the value of a linear functional $\mathcal{F} : \mathcal{P}_2(M) \to \mathbb{R}$.

In the Wasserstein space $(\mathcal{P}_2(M), W_2)$, we generalize this notion to flows of probability measures.

Definition II.2.1. Gradient Flow from Linear Functional.
Given a functional $\mathcal{F}\colon \mathcal{P}_2(M) \to \mathbb{R}$, the gradient flow seeks a curve $\{\mu_t\}_{t\geq 0}$ such that:

$$\frac{d\mu_t}{dt} = -\nabla_{W_2}\mathcal{F}(\mu_t)$$

where ∇_{W_2} denotes the Wasserstein-space gradient. ⌟

A linear functional might seems unrelated at the first glance, but this will be very useful to portrait in a statistical learning problem.

Example II.2.2. Expected Error in Statistical Learning.
A core concept in statistical learning is on probability measures. Given a model distribution μ over a data space S, and a measurable function $f : M \to \mathbb{R}$ (such as a loss or feature-extraction function). We can consider the linear functional defined over a measure as:

$$\mathcal{F}(\mu) = \int_S f(x)\, d\mu(x).$$

It is a *linear functional* of μ, since for any probability measures μ, ν and $\alpha \in [0, 1]$, we can show that:

$$\mathcal{F}(\alpha\mu + (1-\alpha)\nu) = \alpha\mathcal{F}(\mu) + (1-\alpha)\mathcal{F}(\nu).$$

In supervised learning, the goal is to minimize the expected loss:

$$\mathcal{R}(\mu) = \mathbb{E}_{x\sim\mu}[\ell(x,y)]$$

where $\ell(x,y)$ is the loss function. Here, $\mathcal{R}$ is a linear functional on the joint distribution of (x,y).
In the language of statistical learning, the ℓ can be interpreted as how much our predicted measure μ

differs from the actual measure. ◊

Proposition II.2.3. Continuity Equation.
In terms of probability densities $\rho_t = \frac{d\mu_t}{dx}$, the flow can be written as the continuity (transport) equation:

$$\frac{\partial \rho_t}{\partial t} + \nabla \cdot (\rho_t v_t) = 0,$$

where the velocity field v_t is determined by the variation of the functional $\mathcal{F}$, denoted $v_t = -\nabla \frac{\delta \mathcal{F}}{\delta \rho_t}$.

Proof. Consider a curve of probability measures $(\mu_t)_{t\geq 0}$ on $\mathbb{R}^d$, absolutely continuous with respect to Lebesgue measure, so that $\mu_t(dx) = \rho_t(x)\,dx$.
Suppose the measure evolves as a gradient flow of a functional $\mathcal{F}$ on the Wasserstein space, i.e.,

$$\frac{d\mu_t}{dt} = -\nabla_{W_2}\mathcal{F}(\mu_t)$$

In the formalism developed by Otto (see Villani, Ch. 15.1), the tangent space at μ_t can be identified with velocity fields v_t satisfying

$$\frac{d\mu_t}{dt} + \nabla \cdot (v_t \mu_t) = 0$$

which, in terms of densities, becomes the continuity (transport) equation:

$$\frac{\partial \rho_t}{\partial t} + \nabla \cdot (\rho_t v_t) = 0.$$

Now, the velocity field v_t represents the direction in which the measure flows to steepest descent of $\mathcal{F}$, and is given by

$$v_t = -\nabla \frac{\delta \mathcal{F}}{\delta \rho_t}$$

where $\frac{\delta \mathcal{F}}{\delta \rho_t}$ denotes the first variation (functional derivative) of $\mathcal{F}$ with respect to the density ρ_t.
This gradient flow structure in Wasserstein space mirrors the classical gradient descent in Euclidean spaces, but for the evolution of measures. The negative sign ensures movement in the direction of decreasing $\mathcal{F}$.
In sum, the gradient flow in the space of probability densities is governed by

$$\frac{\partial \rho_t}{\partial t} + \nabla \cdot \left[\rho_t \left(-\nabla \frac{\delta \mathcal{F}}{\delta \rho_t}\right)\right] = 0,$$

which is equivalent to the statement in the proposition. □

Remark II.2.4. Connections to PDEs.
Many partial differential equations describing the evolution of probability densities, which can be interpreted as gradient flows in Wasserstein space. ⌟

Proposition II.2.5. The gradient flow of the entropy functional gives the heat equation.

Proof. Let $\rho_t(x)$ be a family of probability densities on $\mathbb{R}^d$. Consider the entropy functional:

$$\mathcal{F}(\rho) = \int_{\mathbb{R}^d} \rho(x) \log \rho(x)\, dx.$$

(i) **Compute First Variation**: The first variation (functional derivative) of $\mathcal{F}$ with respect to ρ is:

$$\frac{\delta \mathcal{F}}{\delta \rho}(x) = \log \rho(x) + 1.$$

(ii) **Write Gradient Flow in Wasserstein Space**: In Wasserstein 2 space, the gradient flow of $\mathcal{F}$ gives the evolution equation:

$$\frac{\partial \rho_t}{\partial t} = \nabla \cdot \left(\rho_t \nabla \frac{\delta \mathcal{F}}{\delta \rho_t} \right).$$

(iii) **Substitutions**: Substitute the computed first variation:

$$\frac{\partial \rho_t}{\partial t} = \nabla \cdot (\rho_t \nabla (\log \rho_t + 1))$$

But, $\nabla(\log \rho_t + 1) = \frac{\nabla \rho_t}{\rho_t}$, so:

$$\rho_t \nabla(\log \rho_t + 1) = \rho_t \frac{\nabla \rho_t}{\rho_t} = \nabla \rho_t$$

Therefore, the evolution equation becomes:

$$\frac{\partial \rho_t}{\partial t} = \nabla \cdot (\nabla \rho_t) = \Delta \rho_t,$$

where Δ is the Laplacian.

Hence, the gradient flow of the entropy functional in Wasserstein 2 space is the heat equation:

$$\frac{\partial \rho_t}{\partial t} = \Delta \rho_t.$$

□

Proposition II.2.6. The gradient flow of the KL divergence gives the Fokker-Planck equation.

Proof. Let $\nu(dx) = e^{-V(x)} dx$ be a reference probability measure on $\mathbb{R}^d$ with potential $V(x)$, and let $\rho_t(x)$ be a probability density evolving over time.

KL Divergence Functional:

$$\mathcal{F}(\rho) = \mathrm{KL}(\rho \| \nu) = \int_{\mathbb{R}^d} \rho(x) \log \left(\frac{\rho(x)}{e^{-V(x)}} \right) dx = \int_{\mathbb{R}^d} \rho(x) \log \rho(x)\, dx + \int_{\mathbb{R}^d} V(x) \rho(x)\, dx$$

(i) **Compute First Variation**: The first variation with respect to ρ is:

$$\frac{\delta \mathcal{F}}{\delta \rho}(x) = \log \rho(x) + 1 + V(x).$$

(ii) **Write the Gradient Flow Equation**: The gradient flow in Wasserstein-2 space is:

$$\frac{\partial \rho_t}{\partial t} = \nabla \cdot \left(\rho_t \nabla \frac{\delta \mathcal{F}}{\delta \rho_t} \right).$$

(iii) **Substitutions**: Substitute the first variation:

$$\frac{\partial \rho_t}{\partial t} = \nabla \cdot \left(\rho_t \nabla (\log \rho_t + 1 + V(x))\right).$$

But $\nabla(\log \rho_t + 1 + V(x)) = \frac{\nabla \rho_t}{\rho_t} + \nabla V(x)$, so

$$\rho_t \nabla(\log \rho_t + 1 + V(x)) = \nabla \rho_t + \rho_t \nabla V(x).$$

(iv) **Format as Fokker-Planck Equation**: Therefore, the gradient flow equation becomes:

$$\frac{\partial \rho_t}{\partial t} = \nabla \cdot (\nabla \rho_t + \rho_t \nabla V(x)) = \Delta \rho_t + \nabla \cdot (\rho_t \nabla V(x)).$$

This is the **Fokker-Planck equation** (also called the forward Kolmogorov equation):

$$\frac{\partial \rho_t}{\partial t} = \Delta \rho_t + \nabla \cdot (\rho_t \nabla V(x)).$$

Thus, the gradient flow of the KL divergence functional in Wasserstein 2 space yields the Fokker-Planck equation. □

II.3 First Variation and JKO Scheme

Definition II.3.1. First Variation.
The "direction" in which the measure should flow to decrease $\mathcal{F}$ most rapidly is given by the **first variation** of the functional with respect to the measure:

$$\frac{d}{dt}\Big|_{t=0} \mathcal{F}(\mu_t),$$

where $(\mu_t)_t$ is a curve in $\mathcal{P}_2(M)$, starting at $\mu_0 = \mu$ and moving infinitesimally towards another. ⌟

The first variation identifies a potential function governing the local cost of redistributing mass, and, in physical terms, corresponds to a gradient in the space of measures.

Remark II.3.2. This formalism endows the space of probability measures with a geometric structure, allowing us to interpret the evolution of distributions as moving "downhill" (along geodesics or gradient flows) in a landscape defined by functionals such as entropy, likelihood, or KL divergence.
This perspective is essential in modern generative modeling, variational inference, and understanding learning dynamics at the level of distributions. ⌟

In the previous examples, we have seen the gradient flow for some specific functions related to PDEs. However, we will then consider the JKO Scheme for a non-closed form approximation of the gradient flow.

Definition II.3.3. Jordan-Kinderlehrer-Otto Scheme.
The Jordan-Kinderlehrer-Otto (JKO) scheme provides a time-discretized way to compute gradient flows

in Wasserstein space:

$$\mu_{k+1} \in \underset{\mu \in \mathcal{P}_2(M)}{\arg\min} \left\{ \mathcal{F}(\mu) + \frac{1}{2\tau} W_2^2(\mu, \mu_k), \right\}$$

where this sequence converges to the true gradient flow as $\tau \to 0$. ⌟

Proof of sketch. We provide the main idea for the sketches leading to the scheme:

(i) **Existence of Minimizers**: By lower semicontinuity and geodesic convexity of $\mathcal{F}$, for each k, the minimization problem has a (possibly non-unique) solution.

(ii) **A Priori Estimates and Compactness**: By construction and its minimizing property, the sequence μ_k^τ stays in a set with uniform bound on the second moment, and $\mathcal{F}(\mu_k^\tau)$ is nonincreasing. This provides compactness with respect to the narrow topology and W_2.

(iii) **Discrete Energy Inequality**: Summing up the optimality conditions yields energy dissipation estimates, which allow control over the "speed" of the path.

Using these estimates and compactness, as $\tau \to 0$, the interpolated paths $\mu^\tau(t)$ converge to a limit curve μ_t, which can be shown (via variational inequalities in the sense of De Giorgi) to be an absolutely continuous curve that satisfies

$$\frac{d\mu_t}{dt} = -\nabla_{W_2} \mathcal{F}(\mu_t)$$

in the distributional sense of the continuity equation in Wasserstein space. That is, μ_t is a gradient flow of $\mathcal{F}$. □

III Interactive Particle System

The theory of Wasserstein gradient flows and the JKO scheme provides a powerful framework to describe the evolution of probability measures driven by the steepest descent of a functional in the space $(\mathcal{P}_2(M), W_2)$. In practice, these measure-valued flows can be represented and simulated using **Interactive Particle Systems (IPS)**.

III.1 Towards IPS Simulations

The IPS Simulation will be a discrete approximation of the representation of the real measure.

Definition III.1.1. IPS represented Measure.
Given a probability measure μ_t on M, we can approximate it by an **empirical measure** consisting of N particles:

$$\mu_t \approx \hat{\mu}_t^N := \frac{1}{N}\sum_{i=1}^{N} \delta_{X_i(t)}$$

where $X_i(t)$ is the position of the i-th particle at time t. ⌟

The Wasserstein gradient flow of a functional $\mathcal{F}(\mu)$ leads to a density evolution given by the continuity equation:

$$\frac{\partial \rho_t}{\partial t} + \nabla \cdot (\rho_t v_t) = 0,$$

where $v_t = -\nabla \frac{\delta \mathcal{F}}{\delta \rho_t}$ is the velocity field derived from the first variation of $\mathcal{F}$.

In the particle viewpoint, the dynamics are given by:

$$\frac{dX_i}{dt} = -\nabla \frac{\delta \mathcal{F}}{\delta \mu_t}(X_i(t))$$

The JKO scheme computes the time-discretized gradient flow by solving

$$\mu_{k+1} \in \arg\min_{\mu \in \mathcal{P}_2(M)} \left\{ \mathcal{F}(\mu) + \frac{1}{2\tau} W_2^2(\mu, \mu_k) \right\}.$$

Remark III.1.2. In practice, this is implemented by updating the particles so that their empirical measure minimizes the above quantity:

$$\hat{\mu}_{k+1}^N \in \arg\min_{\hat{\mu}^N} \left\{ \mathcal{F}(\hat{\mu}^N) + \frac{1}{2\tau} W_2^2(\hat{\mu}^N, \hat{\mu}_k^N) \right\}$$

This approach provides a discrete approximation to the true gradient flow as $\tau \to 0$ and $N \to \infty$. ⌟

In generative modeling, IPS provide a flexible method to learn complex data distributions by simulating flows in measure space, *e.g.*, via score-based diffusion models, Fokker-Planck particle systems, or optimal transport maps. These dynamics are governed by the same principles as Wasserstein gradient flows and

JKO minimization.

Remark III.1.3. Noises Interference.
For stochastic systems, a noise term may also be present, yielding interacting SDEs. ⌟

Here, we consider the Itô SDE formatted as:

$$dX_t = -\nabla\Phi(X_t)\,dt + \sqrt{2\beta^{-1}}\,dW_t,$$

where $\Phi : \mathbb{R}^D \to \mathbb{R}$ is a potential, W_t is standard Brownian motion, and $\beta > 0$. The initial condition $X_0 \sim \rho^0$ describes the law of the initial particle.

Consider the Fokker-Planck Equation as the measure revolution, we have thee law (probability density) ρ_t of X_t satisfies the Fokker-Planck equation:

$$\frac{\partial \rho_t}{\partial t} = \nabla \cdot \big(\nabla\Phi(x)\rho_t\big) + \beta^{-1}\Delta\rho_t,$$

with initial condition $\rho_0 = \rho^0$. This equation describes the evolution of the probability distribution under drift ($-\nabla\Phi$) and diffusion (β^{-1}).
The Fokker-Planck equation is in fact a Wasserstein-2 gradient flow for the **free energy functional**:

$$\mathcal{F}_{\mathrm{FP}}(\rho) = \int \Phi(x)d\rho(x) - \beta^{-1}\int \log\frac{d\rho}{dx}(x)d\rho(x),$$

where the first term is the potential energy, and the second (up to sign) is the entropy functional.

Then, we can think about how to utilize the functional and the IPS to apprxoimate the measure.

Example III.1.4. Relation to IPS.
An **interactive particle system** (IPS) can be constructed by simulating N particles, each evolving as:

$$dX_t^i = -\nabla\Phi(X_t^i)dt + \sqrt{2\beta^{-1}}dW_t^i,$$

so that the empirical measure

$$\hat{\rho}_t^N = \frac{1}{N}\sum_{i=1}^N \delta_{X_t^i}$$

approximates the solution of the Fokker-Planck equation as $N \to \infty$. ◊

From here, this demonstrates that:

- The evolution of the law (distribution) under the SDE is given by a PDE that is itself a Wasserstein gradient flow.
- The free energy $\mathcal{F}_{\mathrm{FP}}$ encodes both the energy landscape and the effect of entropy/diffusion.
- Interactive particle systems provide a tractable way to simulate the Wasserstein gradient flow: particle positions evolve stochastically and interact through the empirical measure.

III.2 Applications of Interactive Particle System

Then, we will consider an interactive particle system controlled by a specific kernel in this scenario.

Example III.2.1. Consider the interactive particle system governed as follows:

$$dX_i(t) = -\nabla V(X_i(t))\,dt - \frac{1}{N}\sum_{j=1}^{N} \nabla_x K(X_i(t), X_j(t))\,dt + \sqrt{2\beta^{-1}}dW_i(t),$$

where the system is constructed with:

- $V : \mathbb{R}^d \to \mathbb{R}^d$ is the *drift*, or the external potential.
- $K : \mathbb{R}^d \to \mathbb{R}^d$ is the *interaction kernel*, modeling the pairwise potential energy between particles, and the component:

  $$-\frac{1}{N}\sum_{j=1}^{N} \nabla_x K(X_i(t), X_j(t))\,dt$$

  is the total interaction force felt by i from all other particles. $\Diamond$

Here, we consider the empirical measure as:

$$\mu_t^N = \frac{1}{N}\sum_{i=1}^{N} \delta_{X_i(t)}$$

as $N \to \infty$, converges (in law) to a deterministic measure $\rho_t(x)\,dx$ evolving according to the nonlinear Fokker-Planck equation:

$$\frac{\partial \rho_t}{\partial t} = \beta^{-1}\Delta\rho_t + \nabla \cdot \big(\rho_t \nabla V(x)\big) + \nabla \cdot \left(\rho_t \int \nabla_x K(x,y)\,\rho_t(y)\,dy\right)$$

This PDE is precisely the **Wasserstein gradient flow** of the free energy functional:

$$\mathcal{F}(\rho) = \int V(x)\,d\rho(x) + \frac{1}{2}\int K(x,y)\,d\rho(x,y) + \beta^{-1}\int \rho(x)\log\rho(x)\,dx$$

Thus:

- The paths of particles in the IPS correspond to a "microscopic" realization of the gradient flow of $\mathcal{F}$ in the space of probability measures.
- In the large-N limit, the collective movement of particles *transports* the empirical measure optimally (in the sense of steepest-descent for $\mathcal{F}$ under the W_2 metric).
- The "noise" strength $2\beta^{-1}$ appears in the entropy contribution to the free energy and determines the strength of diffusion.

Remark III.2.2.

- The IPS approximates evolution of the measure by the action of drift, interaction, and diffusion forces at the particle level.

- The limiting measure evolution, given by the PDE above, describes the optimal (steepest-descent under Wasserstein metric) flow that decreases the free energy $\mathcal{F}$ most efficiently.
- Thus, the IPS is a particle-based scheme for simulating the optimal transport (in gradient flow sense) of the probability measure under the combined effects of V, K, and entropy (via temperature β^{-1}). ⌟

Example III.2.3. Measure Transport via Quadratic Mean-Field Kernel.
Let $K(x,y) = \frac{1}{2}|x-y|^2$, $V(x) = 0$, and $\beta = 1$. Then the interactive particle system can be represented via:

$$dX_i = -\frac{1}{N}\sum_{j=1}^{N}(X_i - X_j)\,dt + \sqrt{2}dW_i = -\left(X_i - \frac{1}{N}\sum_j X_j\right)dt + \sqrt{2}dW_i.$$

Physically interpreting, each particle is pulled toward their mean.
The associated PDE for the density ρ_t is

$$\frac{\partial\rho_t}{\partial t} = \Delta\rho_t + \nabla\cdot\left(\rho_t(x - m(t))\right), \quad m(t) = \int x\,\rho_t(x)\,dx.$$

This models measure transport where every "grain" of mass moves toward the mean of the current density. Furthermore, if we assume $\beta = \infty$, we effectively also neglect the random term, so we will effectively have any measure μ transported to $\delta_{\int d\mu}$. ♢

Remark III.2.4. An easier example could be when $K \equiv 0$ and V represents a pure translation, then it is the same as translating the measure with that direction. ⌟

Then, we will investigate an example of such model applied to machine learning problem.

Example III.2.5. Consensus-Based Optimization for Machine Learning.
Given a loss function $F(x)$ (e.g., neural network training loss), consider the IPS

$$dX_i(t) = -\lambda\left(X_i(t) - m_F(t)\right)\,dt + \sigma\,dW_i(t)$$

where

$$m_F(t) = \frac{\sum_{j=1}^{N} X_j(t)\,e^{-\alpha F(X_j(t))}}{\sum_{j=1}^{N} e^{-\alpha F(X_j(t))}}$$

Each model X_i is attracted toward the consensus, emphasizing better performers. The empirical measure of $\{X_i(t)\}$ thus concentrates around minimizers of $F(x)$, modeling optimal parameter transport in a machine learning ensemble. ♢

III.3 Attention and Self-Attention

Now, we take out attention to the Transformer model, which is a more complicated model in machine learning, consider that we use the flow map on $(\mathbb{R}^d)^n$, we have the input sequence $x_i(0)$, we have the dynamic that:

$$\dot{x}_i(t) = P^{\perp}_{x_i(t)} \left(\frac{1}{Z_{\beta,i}(t)} \sum_{j=1}^{n} e^{\beta \langle Q(t)x_i(t), K(t)x_j(t) \rangle} V(t)x_j(t) \right),$$

in which we have:

$$P^{\perp}_x y = y - \langle x, y \rangle x$$

is the projection of $y \in \mathbb{R}^d$ onto the tangent space $T_x\mathbb{R}^d$ at x. Notice the $Z_{\beta,i}(t)$ function is the part where we do the `softmax` function:

$$Z_{\beta,i}(t) = \sum_{k=1}^{n} e^{\beta \langle Q(t)x_i(t), K(t)x_k(t) \rangle}.$$

Definition III.3.1. Self-Attention.
A common feature used is the self attention mechanism, where we have:

$$A_{i,j}(t) := \frac{e^{\beta \langle Q(t)x_i(t), K(t)x_j(t) \rangle}}{Z_{\beta,i}(t)}, \qquad \text{for } i, j = 1, \cdots, n.$$

⌟

The dynamics defined above can be interpreted as an **interacting particle system** in high dimensions, where each $x_i(t)$ is both influenced by its own position and by *all other particles* in the system, through the weighted sum encoded by the self-attention matrix $A_{i,j}(t)$. In particular, we consider the rule:

$$\dot{x}_i(t) = P^{\perp}_{x_i(t)} \left(\sum_{j=1}^{n} A_{i,j}(t)\, V(t)x_j(t) \right)$$

being the mean-field interaction depending on learned similarity and value features.

Remark III.3.2. Consider the empirical measure at time t,

$$\mu_t^n := \frac{1}{n} \sum_{i=1}^{n} \delta_{x_i(t)}.$$

As $n \to \infty$ and under suitable assumptions, the behavior of the system can be described by the evolution of μ_t. The attention-weighted interactions (summing j's) for a *transport of measure* in $(\mathbb{R}^d)^n$. This is a flow of the measure. ⌟

Hence, the self-attention mechanism defines dynamic flows for the positions $x_i(t)$, which in the limit correspond to a transport map on the measure μ_t. The weights $A_{i,j}(t)$ can be seen as a discrete transport plan, and the evolution resembles (entropic-regularized) optimal transport between empirical measures at different times. The projection $P^{\perp}_{x_i(t)}$ enforces the flow to happen on the tangent space of the sphere, reflecting constraints reminiscent of those in manifold optimal transport problems.

Here, the self-attention mechanism in Transformers may be interpreted dynamically as an interacting particle system, whose evolution induces a transport of the empirical measure, and whose interaction kernel is adaptively learned via attention inputs. This connection opens the way for the analysis of Transformer dynamicsfeature transport, clustering, and information mixingusing the mathematical tools of optimal transport and collective dynamics.

References

[AGS08] Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient Flows: In Metric Spaces and in the Space of Probability Measures*. 2nd. Basel: Birkhäuser, 2008. ISBN: 978-3-7643-8722-4.

[Bar+23] Omer Bar-Tal et al. *MultiDiffusion: Fusing Diffusion Paths for Controlled Image Generation*. 2023. arXiv: 2302.08113 [cs.CV]. URL: https://arxiv.org/abs/2302.08113.

[CTV21] J. A. Carrillo, C. Totzeck, and U. Vaes. *Consensus-based Optimization and Ensemble Kalman Inversion for Global Optimization Problems with Constraints*. 2021. arXiv: 2111.02970 [math.OC]. URL: https://arxiv.org/abs/2111.02970.

[GBC16] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. http://www.deeplearningbook.org. MIT Press, 2016.

[Ges+25] Borjan Geshkovski et al. *A mathematical perspective on Transformers*. 2025. arXiv: 2312.10794 [cs.LG]. URL: https://arxiv.org/abs/2312.10794.

[JKO98] R. Jordan, D. Kinderlehrer, and F. Otto. “The Variational Formulation of the Fokker-Planck Equation”. In: *SIAM Journal on Mathematical Analysis* 29.1 (1998), pp. 1–17. DOI: 10.1137/S0036141096303359.

[Lei+17] Na Lei et al. *A Geometric View of Optimal Transportation and Generative Model*. 2017. arXiv: 1710.05488 [cs.LG]. URL: https://arxiv.org/abs/1710.05488.

[Mok+21] Petr Mokrov et al. *Large-Scale Wasserstein Gradient Flows*. 2021. arXiv: 2106.00736 [cs.LG]. URL: https://arxiv.org/abs/2106.00736.

[Vil08] Cédric Villani. *Optimal transport, old and new*. 2008. URL: http://elenaher.dinauz.org/B07D.StFlour.pdf.

[Wik] Wiki. *Wasserstein metric*. URL: https://en.wikipedia.org/wiki/Wasserstein_metric.