Wasserstein Gradient Flows and Statistical Applications

Mohamad Altrabulsi

August 2025

These notes are written as an introduction to the wonderful and increasingly useful world of optimal transport, with a particular emphasis on ideas and techniques that are relevant to computer science. The goal is to present the subject in a way that is both mathematically principled and accessible to advanced undergraduate students, without assuming extensive background in measure theory or functional analysis.

Optimal transport offers a powerful framework for comparing probability distributions and understanding the geometry of data. Originally developed in the context of analysis and economics, it now plays a central role in modern machine learning, statistics, optimization, and generative modeling. Throughout these notes, abstract concepts are motivated through concrete examples, algorithmic interpretations, and connections to familiar computational problems.

The presentation is largely self-contained and assumes only basic knowledge of linear algebra, calculus, probability, and algorithms. Technical details are introduced gradually and only when needed, with the emphasis placed on intuition, structure, and practical relevance. Pointers to more advanced treatments are provided for readers who wish to explore the theory in greater depth.

These notes were written by an undergraduate computer science student, primarily for other undergraduate students in computer science who are interested in gaining an intuitive and mathematically grounded introduction to optimal transport.

I would like to sincerely thank Professor Michel Pavon and Raghavendra Tripathi for their guidance and insights throughout the development of these notes. I am also grateful to Rashed Al Neyadi and Bikash Shah for their engagement and for attending some of my lectures.

Chapter 1

Monge Problem

1.0.1 Background Knowledge

Definition **1.0.1**

A **Sigma Algebra** $\sigma\text{-}algebra \subseteq \mathcal{P}(X)$ on a set X is a non-empty collection of subsets of X such that

1. $X \in \sigma\text{-}algebra$ (equivalently $\emptyset \in \sigma\text{-}algebra$),
2. if $A \in \sigma\text{-}algebra$ then $A^c := X \setminus A \in \sigma\text{-}algebra$ (closed under complements),
3. if $(A_i)_{i\in\mathbb{N}} \subseteq \sigma\text{-}algebra$, then $\bigcup_{i=1}^{\infty} A_i \in \sigma\text{-}algebra$ (closed under countable unions).

Definition **1.0.2**

A **Measure Space** (X, A, μ) is a tuple such that

1. $X \neq \emptyset$,
2. A is a $\sigma\text{-}algebra$ of X,
3. $\mu : A \to [0, +\infty) \cup \{+\infty\}$ with the following properties:
   (a) $\mu(\emptyset) = 0$,
   (b) $\mu(\cup_i^{\infty} A_i) = \sum_i^{\infty} \mu(A_i)$ whenever $A_i \in A$ and $i \neq j \Rightarrow A_i \cap A_j = \emptyset$.

μ is called a **measure** on X and (X, A) is called a **measurable space**.

***Example* 1.0.1**

Measures can be very different. To illustrate this, compare

1. **The Lebesgue Measure on** $\mathbb{R}$. Let $X = \mathbb{R}$ and $\sigma\text{-}algebra = \mathcal{B}(\mathbb{R})$ (the Borel σ-algebra). The Lebesgue measure λ is defined to generalize the concept of length. For any interval (a, b), its measure is $\lambda((a, b)) = b - a$. Note that a single point has measure zero, $\lambda(\{x\}) = 0$, while the whole line has infinite measure, $\lambda(\mathbb{R}) = +\infty$.
2. **The Counting Measure on** $\mathbb{N}$. Let $X = \mathbb{N}$ and $\sigma\text{-}algebra = \mathcal{P}(\mathbb{N})$ (the power set of $\mathbb{Z}$). The counting measure c is defined for any set $A \subseteq$

To see how different these measures are, consider the singleton set $\{0\}$. Lebesgue measure assigns it zero length:

$$\lambda(\{0\}) = 0,$$

while the counting measure assigns it one measure:

$$c(\{0\}) = 1.$$

***Definition* 1.0.3**

The **Dirac Measure** at $x \in X$ is defined by

$$\delta_x(A) = \begin{cases} 1, & \text{if } x \in A, \\ 0, & \text{if } x \notin A, \end{cases}$$

1.1 Theory

Let α, β be discrete measures on the measurable spaces (X, A) and (Y, B)

Discrete measures can be written as sums of point masses, i.e.

$$\alpha = \sum_{i}^{n} a_i \delta_{x_i}, \qquad \beta = \sum_{j}^{m} b_j \delta_{y_j} \tag{1.1}$$

Let $T : X \to Y$ be a function with constraint

$$b_j = \sum_{x_i \in T^{-1}(y_j)} a_i \quad \forall j \in \{1, \ldots, m\} \tag{1.2}$$

Equation (1.2) ensures that the total mass assigned to each y_j by the transport map T matches the measure β at that point. In the classical mines-and-factories example, this condition guarantees that the supply transported to each factory exactly meets its demand (Example 1.1).

Note 1.1.1

We shorten the writing of Equation (1.2) with the following notation:

$$T_{\#\alpha} = \beta.$$

Imagine the **pushforward measure** $T_{\#\alpha}$ as an extension of T, acting on measures rather than individual densities.
In the continuous case, the constraint in Equation (1.2) becomes

$$\beta(b) = \alpha(T^{-1}(b)) \quad \forall b \in B \tag{1.3}$$

Example 1.1.1

Let

$$\alpha = 0.2\,\delta_{x_1} + 0.3\,\delta_{x_2} + 0.5\,\delta_{x_3}, \qquad \beta = 0.5\,\delta_{y_1} + 0.5\,\delta_{y_2},$$

(see Figures (1.1) and (1.2) for a visualization).
Observe that there are multiple transport maps

$$T : X \to Y \quad \text{such that} \quad T_{\#}\alpha = \beta.$$

On the other hand, the inverse problem of finding a measurable map

$$Q : Y \to X \quad \text{with} \quad Q_{\#}\beta = \alpha$$

has no solution because with a deterministic T, each point mass x stays a single point mass after transformation $T(x)$. So, $Q_{\#}\beta$ can only have 2 non-zero point masses as opposed to α which has 3.

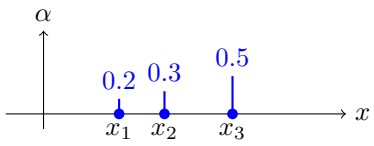

Figure 1.1: Discrete measure α

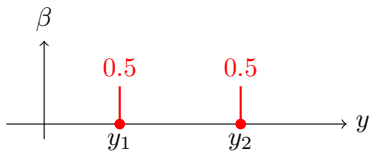

Figure 1.2: Discrete measure β

Let C denote a cost function such that $C : X \times Y \to [0, \infty)$. Finally, now we can formulate the Monge problem as:

$$\min_T \{E(c) : T_{\#\alpha} = \beta\} \tag{1.4}$$

where E(c) is the expectation of c. In the discrete case, it becomes

$$\min_T \left\{ \sum_{i=1}^{n} a_i C(x_i, T(x_i)) : T_{\#\alpha} = \beta \right\} \tag{1.5}$$

In the continuous case, it becomes

$$\min_T \left\{ \int_X C(x, T(x))\, d\alpha(x) : T_{\#\alpha} = \beta \right\} \tag{1.6}$$

Generally, you want to find the mapping T that minimizes the cost function while still pushing α to β.

***Example* 1.1.2 Mines and Factories I**

Consider n mines and m factories. Mine $i \in \{1, \ldots, n\}$ supplies $a_i \geq 0$ units of a resource, and factory $j \in \{1, \ldots, m\}$ requires $b_j \geq 0$, with $\sum_{i=1}^{n} a_i = \sum_{j=1}^{m} b_j$. Transporting one unit from mine i to factory j costs $C(i, j) \geq 0$.
In Monge's formulation, a transport map is a plan $T : \{1, \ldots, n\} \to \{1, \ldots, m\}$ that assigns each mine to a single factory (no splitting of a mine's supply) – with the condition that $T_{\#\alpha} = \beta$ (as in Equation (1.2)).

For example, assume that $n = 1$, that there is only one mine A, so if $\mu = \delta_A$ then the set of possible transport maps is those of the form $T_{\#\delta_A} = \delta_{T(A)}$.
The total cost of a feasible plan T is

$$\text{Cost}(T) \;=\; \sum_{i=1}^{n} a_i\, C\big(i,\, T(i)\big),$$

and the Monge problem seeks T that minimizes Cost(T) (as in Equation (1.1)).

***Example* 1.1.3 Book Shifting**

Consider a bookshelf filled with identical books, each book being given a position index $i \in \mathbb{Z}_+$.
At the start, the books occupy positions indexed by $\{0, n\}$, which corresponds to the uniform measure

$$\mu = \frac{1}{n}\chi_{\{0,n\}}\,\mathcal{L}^1,$$

while the target configuration is obtained by shifting the block of books one unit to the right, i.e.,

$$\nu = \frac{1}{n}\chi_{\{1,n+1\}}\,\mathcal{L}^1.$$

μ and ν can be thought of as uniform Lebesgue measures on a restricted interval.

A natural transport map is simply

$$T(t) = t + 1, \qquad t \in [0, n],$$

which pushes μ onto ν. Its cost is

$$\frac{1}{n}\int_0^n |t - (t+1)|\,dt = 1.$$

With the cost function $C(x, y) = |y - x|$
T can be shown to be optimal. However, optimality does not imply uniqueness. Indeed, the map

$$T_2(t) = \begin{cases} t + n, & t = 0, \\ t, & t \in \{1, n\}, \end{cases}$$

also transports μ to ν with the same total cost 1

Examples 1.1 and 1.1 show some downsides of the Monge formulation, such as the inability to split masses – which is unrealistic in the case of the mines-and-factories problem – and which leads to the inverse problem mentioned in 1.1 having no solution. In addition, 1.1 shows that Monge problem, when feasible, doesn't necessarily have a unique optimal solution.

Chapter 2

Kantorovich Relaxation

2.1 Background Knowledge

Definition **2.1.1**

If (X, A, μ) is a measure space with the conditions,

$$\mu(X) = 1.$$

Then (X, A, μ) is called a **probability space** and μ is called a **probability measure**.
In the case when μ is referred to as a probability measure on $\mathbb{R}^d$, the measure space is implicitly $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$, where $\mathcal{B}(\mathbb{R}^d)$ is the Borel σ-algebra: the smallest σ-algebra containing all open sets of $\mathbb{R}^d$.

2.2 Theory

Let μ and ν be probability measures[1] on $\mathbb{R}^d$.

Let γ denote the *coupling* of μ and ν s.t $\forall$ Borel sets B $\in \mathcal{B}(\mathbb{R}^d)$

$$\gamma(B \times R^d) = \mu(N) \quad \text{and} \quad \gamma(R^d \times B) = \nu(M) \tag{2.1}$$

Coupling γ can be understood as a joint distribution of random variables $X \sim \mu$ and $Y \sim \nu$, where

$$P(a \leq X \leq b) = \mu([a, b]) = \int_a^b d\mu, \tag{2.2}$$

$$P(a \leq Y \leq b) = \nu([a, b]) = \int_a^b d\nu. \tag{2.3}$$

[1]The Kantorovich relaxation can be understood generally for any measures μ, ν - with the same total mass - but for the sake of brevity only probability measures are considered.

And so $\gamma(A \times B) = J(A, B)\ \forall A, B \in \mathcal{B}(\mathbb{R}^d)$

Let $\Gamma_{\mu,\nu}$ represent the set of all possible valid couplings (satisfying Equations (2.3) and (2.2)) between μ and ν. Note that the set $\Gamma_{\mu,\nu}$ is *non-empty* since the coupling $\gamma_1 = \mu \otimes \nu$ *(or in joint distribution terms* $J_1(A, B) = X(A) \cdot Y(B)$*)* always exists. This represents one advantage of the Kantorovich formulation with respect to the Monge formulation.

Let $C : \mathbb{R}^d \times \mathbb{R}^d \to [0, \infty)$ be a cost function. Then, the Kantorovich formulation is:

$$\min_{\gamma \in \Gamma_{\mu,\nu}} \left\{ \int_{\mathbb{R}^d \times \mathbb{R}^d} C(x, y)\, d\gamma(x, y) \right\} \tag{2.4}$$

A solution to (2.4) is called an *optimal transport map* or an *optimal coupling.*

***Note* 2.2.1**

For the Monge problem, the constraint that α is pushed to β was represented by Equation (1.3). In the Kantorovich formulation, this is represented by Equations (2.2) and (2.3).

***Example* 2.2.1 Mines and Factories II**

Same setup as Example 1.1. Now, in addition, define cost matrix C $\in \mathbb{R}_+^{n \times m}$ *s.t* C_{ij} is the cost of moving 1 resource from mine i to factory j and a plan matrix $J \in \mathbb{R}_+^{n \times m}$ *s.t* J_{ij} resources are moved from mine i to factory j.

Finally, define U as the set of plans (couplings) that meet the marginal constraints

$$U = \left\{ J \in \mathbb{R}_+^{n \times m} : \sum_j (J_{ij}) = a_i\ \forall i \in [1, n] \text{ and } \sum_i (J_{ij}) = b_j\ \forall j \in [1, m] \right\} \tag{2.5}$$

Then, the most cost-effective plan can be found by solving

$$\min_{J \in U} \left\{ \sum_{i,j} c_{ij} J_{ij} \right\} \tag{2.6}$$

Chapter 3

Wasserstein distance

3.1 Background Knowledge

Definition **3.1.1**

A pair (X,d) is called a **Metric space** if X is a set and the **Metric** $d : X \times X \to \mathbb{R}$ is a function that satisfies $\forall x, y, z \in X$ the following:

1. $d(x, y) \geq 0$,
2. $d(x, y) = d(y, x)$,
3. $d(x, y) = 0 \Leftrightarrow x = y$,
4. $d(x, y) \leq d(x, z) + d(z, y)$ (triangle equality)

***Example* 3.1.1**

The geometry of metric spaces may be quite difficult and possibly non-intuitive. One of the simplest examples is the *discrete metric*, defined on any set by

$$d(x,y) = \begin{cases} 0, & \text{if } x = y, \\ 1, & \text{if } x \neq y. \end{cases}$$

This metric assigns distance 1 between any two distinct points.

More familiar examples include metrics on $\mathbb{R}^n$:

- the Euclidean distance (the ℓ^2 metric),

$$d(x,y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2},$$

- the ℓ^1 (taxicab) distance,

$$d(x,y) = \sum_{i=1}^{n} |x_i - y_i|,$$

- more generally, any norm $\|\cdot\|$ induces the metric $d(x,y) = \|x-y\|$.

Below are several further examples illustrating the diversity of metric spaces:

- Graphs with shortest-path distance. Consider a connected undirected graph $G = (V, E)$. The distance between vertices $u, v \in V$ is defined as the length of the shortest path joining them. With this distance, the vertex set V becomes a metric space.

- Matrices with the Frobenius metric. On $M = \mathbb{C}^{n\times n}$, define

$$d(A,B) = \|A - B\|_F = \sqrt{\operatorname{trace}((A-B)^*(A-B))}.$$

  This makes the space of matrices a metric space.

- The unit circle with arc-length distance. If A and B lie on a circle of radius r, and θ is the (shortest) angle between them, then

$$d(A,B) = r\theta.$$

  This metric measures distance along the circle, not through its interior.

For more details and examples of metric spaces, see the lecture notes by Andersson, Björn, and Wiman [1].

In machine learning, the goal is often to learn a function from a finite set of points, called training data, with the hope that this learned function will generalize to unseen testing data. A fundamental difficulty is overfitting, where the learned function matches the training data too closely and therefore performs poorly on new samples. One classical fix is regularization, which penalizes the magnitude of the model parameters. The idea behind this is that large model weights make the model too sensitive to small changes in the input data, and thus forcing the weights to stay small can prevent such sensitivity.

The two most common choices are the ℓ^2 and ℓ^1 norms:

- ℓ^2 **regularization (Ridge).**

$$\|w\|_2^2 = \sum_{i=1}^{d} w_i^2,$$

  which corresponds to the geometry of the Euclidean metric.

- ℓ^1 **regularization (Lasso).** This penalizes the taxicab norm,

$$\|w\|_1 = \sum_{i=1}^{d} |w_i|,$$

  whose geometry corresponds to the ℓ^1 metric.

For more details, see Russell and Norvig [8, Chapter 19.6].

***Definition* 3.1.2**

A **Polish space** is a metric space that is:

1. *Complete*: every Cauchy sequence converges in the space.
2. *Separable*: it contains a countable dense subset.

3.2 Theory

Let $\mathcal{P}_p(\mathbb{R}^d)$ be the set of probability measures over $\mathbb{R}^d$ with finite p-th moment

$$\mu \in \mathcal{P}_p(\mathbb{R}^d) \leftrightarrow \int ||x||^p \mu(dx) < \infty \tag{3.1}$$

The p-Wasserstein distance between two probability measures $\mu, \nu \in \mathcal{P}_p(\mathbb{R}^{d)}$ is defined to be

$$W_p(\mu, \nu) = \inf_{\gamma \in \Gamma_{\mu,\nu}} (\int ||x - y||^p \gamma(dx, dy))^{\frac{1}{p}} \tag{3.2}$$

Then, the pair $(\mathcal{P}_p(\mathbb{R}^d), W_p)$ forms a metric space, and W_p is a mathematical distance or a metric.

In addition, it turns out that $(\mathcal{P}_p(\mathbb{R}^d), W_p)$ is a particularly good metric space, a Polish space. For more details, see Villani [9, Chapter 1.1] in Topics in Optimal Transportation (2003).

Chapter 4

The One-dimensional case

4.1 Background Knowledge

Definition **4.1.1**

A **Cumulative Distribution Function (CDF)** associated with a real-valued random variable X is a function

$$F(x) = \mathbb{P}(X \leq x), \quad x \in \mathbb{R}.$$

The CDF $F : \mathbb{R} \to [0, 1]$ satisfies:

- F is non-decreasing.
- $\lim_{x \to -\infty} F(x) = 0$ and $\lim_{x \to +\infty} F(x) = 1$.
- F is right-continuous.

$F(x)$ then gives the probability that the random variable takes a value less than or equal to x.

Definition **4.1.2**

A **Pseudo-inverse** F^{-1} of a cumulative distribution function (CDF) $F : \mathbb{R} \to [0, 1]$ is defined as

$$F^{-1}(t) = \inf\{x \in \mathbb{R} \mid F(x) \geq t\}.$$

In words, $F^{-1}(t)$ is the smallest value x such that the distribution assigns probability at least t to $(-\infty, x]$.

A special property of pseudo-inverses of CDFs can be used to simulate any random variable X with pseudo-inverse F_X^{-1}.

Let $U \sim Unif([0,1])$ denote a uniform random variable, then one can construct a random variable $Z \sim F_X^{-1}(U)$ with CDF F. This can be shown by using the following:

$$P(Z \leq t) = P(F_X^{-1}(U) \leq t) = P(U \leq F(t)) = F(t)$$

Where the second equality hold because of the non-decreasing nature of CDFs, and the third equality holds because U has the CDF of a uniform distribution.

4.2 Theory

Let $\mu, \nu \in \mathcal{P}_1(\mathbb{R})$ be probability measures with cumulative distribution functions (CDFs) F_μ and F_ν. Consider a random variable $U \sim \text{Unif}([0,1])$. Then, the optimal coupling γ_1 is given by the joint distribution of

$$\big(X,\, F_\nu^{-1}(F_\mu(X))\big), \quad X \sim \mu.$$

This induces the transport map

$$T_1 = F_\nu^{-1} \circ F_\mu.$$

This shows that, in the one-dimensional case, the Monge and Kantorovich problems converge to the same solutions—a result which, surprisingly, holds in any dimension.

For more details, see Chewi, Niles-Weed, and Rigollet [3, Chapter 1.3] in Statistical Optimal Transport (2019).

Chapter 5

Kantorovich duality

5.1 Background Knowledge

Definition **5.1.1**

To find the extreme points of a function $f(x, y)$ under the constraint $g(x, y) = k$, one solves the system

$$\nabla f(x, y) = \lambda \nabla g(x, y), \quad g(x, y) = k,$$

where ∇f and ∇g denote the gradient vectors of f and g, and $\lambda \in \mathbb{R}$ is a scalar known as the **Lagrange multiplier**.
This information can be expressed through the following function, known as the **Lagrangian**

$$L(x, y, \lambda) = f(x, y) + \lambda \, (g(x, y) - k)$$

One can then solve the following system of equations

$$\frac{\partial L}{\partial x} = 0, \ \frac{\partial L}{\partial y} = 0, \ \frac{\partial L}{\partial \lambda} = 0$$

***Definition* 5.1.2**

Lagrangian duality provides a framework for solving constrained optimization problems. Consider the **primal problem**

$$\inf_{x \in X} f(x) \quad \text{subject to } g_i(x) \leq 0,\ i = 1, \ldots, m.$$

The associated *Lagrangian* is

$$L(x, \lambda) = f(x) + \sum_{i=1}^{m} \lambda_i g_i(x), \quad \lambda \in \mathbb{R}_+^m$$

The **dual function** is defined as

$$\phi(\lambda) = \inf_{x \in X} L(x, \lambda).$$

Note that $\phi(\lambda)$ is always concave, even without assuming convexity of f or g_i.

If p^* denotes the optimal value of the primal problem, then

$$\phi(\lambda) \leq p^* \quad \text{for all } \lambda \in \mathbb{R}_+^m,$$

because for any feasible $x_p \in X$,

$$\phi(\lambda) = \inf_{x \in X} L(x, \lambda) \ \leq \ L(x_p, \lambda) \ \leq \ f(x_p).$$

Thus, $\phi(\lambda)$ provides a lower bound on the primal optimum. A natural continuation is to maximize this lower bound, leading to the **dual problem**:

$$\sup_{\lambda \in \mathbb{R}_+^m} \phi(\lambda) \ = \ \sup_{\lambda \in \mathbb{R}_+^m} \inf_{x \in X} L(x, \lambda).$$

By construction,

$$\sup_{\lambda \in \mathbb{R}_+^m} \phi(\lambda) \ \leq \ \inf_{x \in X} f(x),$$

which is the principle of **weak duality**.
If f and each g_i are convex and constraint qualifications hold, then **strong duality** applies, i.e., the two values coincide.

The dual Kantorovich problem can be derived as a direct application of Lagrangian duality to the primal Kantorovich formulation. For simplicity, we focus on the 2-Wasserstein distance, although Kantorovich duality extends to more general cost functions.

An important step is to consider the squared 2-Wasserstein distance, $W_2^2(\mu, \nu)$. This is because minimizing W_2 is equivalent to minimizing W_2^2, but the latter

corresponds to an optimization problem with an objective function that is linear in the optimization variable γ. The primal problem is:

$$W_2^2(\mu,\nu) = \inf_{\gamma\in\Gamma_{\mu,\nu}} \int_{\mathbb{R}^d\times\mathbb{R}^d} \|x-y\|^2 \,\gamma(dx,dy), \tag{5.1}$$

where the feasible set $\Gamma_{\mu,\nu}$ is the set of all couplings with marginals μ and ν. This constraint is defined as:

$$\gamma\in\Gamma_{\mu,\nu} \iff \begin{cases}\gamma(A\times\mathbb{R}^d)=\mu(A),\\ \gamma(\mathbb{R}^d\times B)=\nu(B),\end{cases} \qquad \forall A,B\in\mathbb{B}(\mathbb{R}^d). \tag{5.2}$$

An equivalent formulation of these constraints, which is more convenient for duality, is

$$\int_{\mathbb{R}^d\times\mathbb{R}^d} (f(x)+g(y))\,\gamma(dx,dy) = \int_{\mathbb{R}^d} f(x)\,\mu(dx) + \int_{\mathbb{R}^d} g(y)\,\nu(dy), \qquad \forall f,g\in C_b, \tag{5.3}$$

where C_b denotes the space of bounded continuous functions on $\mathbb{R}^d$.

Let M_+ denote the set of all positive measures on $\mathbb{R}^d\times\mathbb{R}^d$. We can now form the *Lagrangian* by introducing the functions f and g as Lagrange multipliers for the marginal constraints. The Lagrangian $L(\gamma,f,g)$ is a function of the primal variable $\gamma\in M_+$ and the dual variables $f,g\in C_b$.

$$L(\gamma,f,g) = \int \|x-y\|^2\,d\gamma - \left(\int (f(x)+g(y))\,d\gamma - \int f(x)\,d\mu - \int g(y)\,d\nu\right). \tag{5.4}$$

The dual problem is found by maximizing the dual function $D(f,g)=\inf_{\gamma\in M_+} L(\gamma,f,g)$. We can rearrange the Lagrangian to isolate terms involving γ:

$$L(\gamma,f,g) = \int f(x)\,d\mu + \int g(y)\,d\nu + \int \left(\|x-y\|^2 - f(x) - g(y)\right)\,d\gamma. \tag{5.5}$$

The dual function is therefore:

$$\begin{aligned} D(f,g) = \inf_{\gamma\in M_+} L(\gamma,f,g) = &\int f(x)\,d\mu + \int g(y)\,d\nu \\ &+ \inf_{\gamma\in M_+}\left\{\int \left(\|x-y\|^2 - f(x) - g(y)\right)\gamma(dx,dy)\right\}. \end{aligned} \tag{5.6}$$

Now, we analyze the above infimum over all positive measures γ.

- If there exists any pair (x,y) such that $\|x-y\|^2 - f(x) - g(y) < 0$, one could choose γ to be a Dirac mass concentrated at (x,y) and scale it to make the integral arbitrarily negative. Thus, the infimum would be $-\infty$.
- To obtain a non-trivial dual problem, we must therefore constrain the dual variables (f,g) to prevent this. This requires that the integrand be non-negative everywhere:

$$\|x-y\|^2 - f(x) - g(y) \geq 0 \quad \implies \quad f(x)+g(y) \leq \|x-y\|^2 \quad \forall x,y. \tag{5.7}$$

The dual problem is to maximize the dual function, $\sup_{f,g \in C_b} D(f,g)$. Incorporating the derived constraint, this becomes:

$$\sup_{\substack{f,g \in C_b \\ f(x)+g(y) \leq \|x-y\|^2}} \left\{ \int f(x)\, \mu(dx) + \int g(y)\, \nu(dy) \right\}. \tag{5.8}$$

This guarantees weak duality. For optimal transport, strong duality also holds, meaning the value of this dual problem is equal to the value of the primal problem, $W_2^2(\mu, \nu)$. For a proof of this, as well as additional details omitted here, see Chewi, Niles-Weed, and Rigollet [3, Chapters 1.5.2–1.5.3] in Statistical Optimal Transport (2019).

Chapter 6

Brenier Theorem

6.1 Background Knowledge

Definition **6.1.1**

A function $\varphi : \mathbb{R}^d \to \mathbb{R}$ is **convex** if for all $x, y \in \mathbb{R}^d$ and $t \in [0,1]$,

$$\varphi\big(tx + (1-t)y\big) \;\leq\; t\,\varphi(x) + (1-t)\,\varphi(y).$$

An important quality of convex functions is that, when φ is differentiable, its gradient is **monotone**:

$$\langle \nabla\varphi(x) - \nabla\varphi(y),\, x - y\rangle \;\geq\; 0 \quad \forall\, x, y \in \mathbb{R}^d.$$

***Definition* 6.1.2**

A measure $\mu \in \mathcal{P}(\mathbb{R}^d)$ is said to be **absolutely continuous** with respect to the Lebesgue measure $\mathcal{L}^d$, written

$$\mu \ll \mathcal{L}^d,$$

if for every measurable set $A \subset \mathbb{R}^d$,

$$\mathcal{L}^d(A) = 0 \quad \Rightarrow \quad \mu(A) = 0.$$

In this case, there exists a density $\rho \in L^1(\mathbb{R}^d)$ such that

$$d\mu(x) = \rho(x)\, dx,$$

which follows from the Radon–Nikodym theorem.

For a detailed discussion of absolute continuity of measures, see Royden and Fitzpatrick [7, Chapter 6].

6.2 Theory

Brenier's theorem extends the one-dimensional result of Section 4 to any dimension d.

***Theorem* 6.2.1**

Let $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$ be two probability measures, with μ absolutely continuous and $X \sim \mu$. Then there exists a convex function $\varphi : \mathbb{R}^d \to \mathbb{R}$ such that

$$(X, \nabla\varphi(X)) \sim \gamma_1 \in \Gamma_{\mu,\nu},$$

where γ_1 is the optimal coupling for $W_2^2(\mu, \nu)$.

The mapping $T(X) = \nabla\varphi(X)$ may seem arbitrary at first, but any optimal transport map must be monotone, and gradients of convex functions are monotone by construction. To see why monotonicity is necessary, consider points $x_1 < x_2$ and $y_1 < y_2$. The squared W_2 cost of assigning them via a monotone map, C_{mono}, versus a cross map, C_{cross}, can then be compared

$$C_{mono} = (x_1 - y_1)^2 + (x_2 - y_2)^2 = x_1^2 - 2x_1y_1 - y_1^2 + x_2^2 - 2x_2y_2 - y_2^2$$
$$C_{cross} = (x_1 - y_2)^2 + (x_2 - y_1)^2 = x_1^2 - 2x_1y_2 - y_2^2 + x_2^2 - 2x_2y_1 - y_1^2$$

Then the difference could be examined

$$C_{cross} - C_{mono} = -2x_1y_2 - 2x_2y_1 + 2x_1y_1 + 2x_2y_2 =$$
$$2(x_1y_1 + x_2y_2 - x_1y_2 - x_2y_1) = 2(x_1(y_1 - y_2) + x_2(y_2 - y_1)) =$$
$$2(x_1(y_1 - y_2) - x_2(y_1 - y_2)) = 2(y_1 - y_2)(x_1 - x_2) > 0$$

So, a monotone map will always produce a smaller cost than one that is not.

Chapter 7

Entropic Regularization

7.1 Background Knowledge

Definition **7.1.1**

The **Kullback-Leibler (KL) divergence** is a way to measure how different two probability measures are, originating from statistical mechanics, where it is known as *relative entropy*.
Let α and β be probability measures on a measurable space such that α is absolutely continuous with respect to β (denoted $\alpha \ll \beta$). Then the KL divergence is defined as

$$KL(\alpha \,\|\, \beta) \;=\; \begin{cases} \displaystyle\int \frac{d\alpha}{d\beta}(x) \log\!\left(\frac{d\alpha}{d\beta}(x)\right) d\beta(x), & \text{if } \alpha \ll \beta, \\ +\infty, & \text{otherwise.} \end{cases}$$

If α and β are absolutely continuous with with respect to lebegsue measure, then α, β have densities p, q and

$$KL(\alpha \,\|\, \beta) \;=\; \int p(x) \log\!\left(\frac{p(x)}{q(x)}\, dx\right).$$

In discrete form, if $\alpha = (a_i)$ and $\beta = (b_i)$ are two probability vectors with $a_i > 0 \implies b_i > 0$, then

$$KL(\alpha \,\|\, \beta) \;=\; \sum_i a_i \log\!\left(\frac{a_i}{b_i}\right).$$

The KL divergence satisfies several fundamental properties:

- **Non-negativity and identity of indiscernibles.**

$$KL(\alpha \,\|\, \beta) \;\geq\; 0,$$

with

$$KL(\alpha \,\|\, \beta) = 0 \quad \text{if and only if} \quad \alpha = \beta \text{ almost everywhere.}$$

This follows from Jensen's inequality.

- **Asymmetry.** In general,

$$KL(\alpha \,\|\, \beta) \;\neq\; KL(\beta \,\|\, \alpha),$$

so the KL divergence is not symmetric. While symmetry can be enforced by symmetrization (e.g. by considering $KL(\alpha\|\beta) + KL(\beta\|\alpha)$), this does not resolve the lack of a triangle inequality.

- **Not a metric.** Even after symmetrization, the KL divergence fails to satisfy the triangle inequality, and therefore cannot be turned into a true distance.

- **Convention.** In the discrete setting, the definition uses the convention

$$0 \log 0 := 0,$$

which is consistent with the limit $\lim_{x \to 0^+} x \log x = 0$.

KL divergence has a lot of uses in Machine Learning. One of the most important, however, is it's presence in the cross-entropy loss function. The cross-entropy $H(p, q)$ between two probability distributions p and q is defined by

$$H(p, q) = H(p) + KL(p \,||\, q) \tag{7.1}$$

Where H(p) is the Shannon-entropy p defined by

$$H(p) = -\int p(x) \log p(x)\, dx \tag{7.2}$$

Since $H(p)$ is a constant, minimizing the cross-entropy with respect to q becomes equivalent to simply minimizing the KL divergence between p and q. For more details on cross-entropy and its minimization, see Russell and Norvig [8, Chapter 19.6].

7.2 Theory

The main idea of entropic optimal transport is to add a penalty term to the original problem to encourage the coupling to be more spread out than the solution to the original problem. For the primal problem, this could be formulated

as the following

$$We_2^2(\mu,\nu) = \inf_{\gamma\in\Gamma_{\mu,\nu}} \int_{\mathbb{R}^d\times\mathbb{R}^d} \|x-y\|^2\,\gamma(dx,dy) + \epsilon KL(\gamma\,||\,\mu\otimes\nu), \tag{7.3}$$

Where ϵ controls how spread out the best coupling is. There exists a dual formulation for equations (7.3) as well

$$De(\mu,\nu) = \sup_{f,g\in C_b} \left\{ \int f(x)\,\mu(dx) + \int g(y)\,\nu(dy) - \epsilon \iint e^{\frac{(f(x))+g(y)-||x-y||^2)}{\epsilon}} - 1\,\mu(dx)\nu(dy) \right\}. \tag{7.4}$$

In addition, the optimal optimal solution (f_1, g_1) to (7.4) can be used to get an optimal solution γ_1 to (7.3)

$$\gamma_1(dx,dy) = e^{\frac{f_1(x)+g_1(y)-||x-y||^2}{\epsilon}}\,\mu(dx)\nu(dy) \tag{7.5}$$

(7.3) is strictly convex - a solution always exists and has only one local/global minimum - a solution can be found always using the Sinkhorn algorithm

1. initialize $f^1 = 0$
2. perform update $g^t(y) = -\epsilon\log\int(\frac{f^{t-1}-||x-y||^2}{\epsilon})\,\mu(dx)$
3. perform update $f^t(y) = -\epsilon\log\int(\frac{g^{t}-||x-y||^2}{\epsilon})\,\nu(dy)$

In the discrete setting, let

$$\alpha = \sum_{i=1}^n a_i\,\delta_{x_i}, \qquad \beta = \sum_{j=1}^m b_j\,\delta_{y_j}, \qquad \sum_i a_i = \sum_j b_j = 1.$$

The entropically regularized Kantorovich problem is

$$\inf_{\gamma\in\Gamma_{\alpha,\beta}} \sum_{i=1}^n\sum_{j=1}^m c_{ij}\,\gamma_{ij} \;+\; \varepsilon\,\mathrm{KL}(\gamma\,\|\,\alpha\otimes\beta).$$

This problem is strictly convex and admits a unique optimizer.

The optimal coupling has the so-called Gibbs form (see Peyré–Cuturi, Chapter 4):

$$\gamma_{ij} = u_i\,K_{ij}\,v_j, \qquad K_{ij} := e^{-c_{ij}/\varepsilon},$$

where $(u,v)\in\mathbb{R}_+^n\times\mathbb{R}_+^m$ are scaling vectors chosen so that γ satisfies the marginal constraints. They can be computed using the **Sinkhorn algorithm**:

1. Initialize $v^{(0)}\in\mathbb{R}_+^m$ as the vector of ones.
2. Update

$$u_i^{(t+1)} \;=\; \frac{\alpha_i}{(Kv^{(t)})_i}.$$

3. Update

$$v_j^{(t+1)} = \frac{\beta_j}{(K^\top u^{(t+1)})_j}.$$

***Example* 7.2.1 Mines and Factories III — Entropic Regularization**

Suppose there are two mines and two factories with equal supply and demand:

$$a_1 = a_2 = b_1 = b_2 = \tfrac{1}{2},$$

and cost matrix

$$C = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

Then

$$K = \begin{pmatrix} 1 & q \\ q & 1 \end{pmatrix}, \qquad q = e^{-1/\varepsilon}.$$

By symmetry $u_1 = u_2$ and $v_1 = v_2$, giving

$$\gamma_\varepsilon = \frac{1}{2(1+q)} \begin{pmatrix} 1 & q \\ q & 1 \end{pmatrix}.$$

$$\varepsilon \to 0 \quad \Rightarrow \quad \gamma_\varepsilon \to \begin{pmatrix} \frac{1}{2} & 0 \\ 0 & \frac{1}{2} \end{pmatrix} \quad \text{(deterministic matching)},$$

$$\varepsilon \to \infty \quad \Rightarrow \quad \gamma_\varepsilon \to \begin{pmatrix} \frac{1}{4} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{4} \end{pmatrix} = \alpha \otimes \beta \quad \text{(fully diffused coupling)}.$$

For small ε, each mine sends all of its supply to the cheapest factory: mine 1 $\to$ factory 1 and mine 2 $\to$ factory 2. As ε grows, each mine starts splitting its resources between the two factories, creating a "soft assignment." This reflects the stabilizing effect of entropic regularization: the plan becomes smoother, more robust, and easier to compute, at the cost of deviating from the exact cheapest assignment.

Chapter 8

Dynamic Formulation

8.1 Background Knowledge

Definition **8.1.1**

A **vector field** on a domain $D \subseteq \mathbb{R}^d$ is a function

$$v : D \ \rightarrow \ \mathbb{R}^d, \tag{8.1}$$

that assigns to each point $x \in D$ a vector $v(x) \in \mathbb{R}^d$. One can think of a vector field as describing a flow or velocity attached to each point in space. For example, in fluid dynamics, $v(x)$ may represent the velocity of the fluid particle located at position x.

Definition **8.1.2**

The **divergence** of a vector field $v = (v_1, \ldots, v_d) : \mathbb{R}^d \rightarrow \mathbb{R}^d$ is a scalar field defined by

$$div(v(x)) = \sum_{i=1}^{d} \frac{\partial v_i}{\partial x_i}(x) = (\nabla \cdot v)(x).$$

Divergence measures the net rate of *outflow* of the vector field at a point.

- If $div(v(x)) > 0$, the point x is a *source* (mass is spreading out).
- If $div(v(x)) < 0$, the point x is a *sink* (mass is concentrating).
- If $div(v(x)) = 0$, the field is locally incompressible at x (mass is preserved).

8.2 Theory

Consider a particle X evolving under the influence of a time-dependent vector field $(v_t)_{t\geq 0}$. Its motion is described by the ODE

$$\frac{d}{dt}X_t = v_t(X_t), \tag{8.2}$$

where X_t is the position of the particle at time t and $\frac{d}{dt}X_t$ its velocity.

If v_t is sufficiently regular, then for any initial condition X_0, this ODE has a unique solution on sime time interval $[0, t_{max}]$. If X_0 is drawn from a distribution $\mu_0 \in \mathcal{P}_2(\mathbb{R}^d)$, then the random trajectory $(X_t)_{t\geq 0}$ induces a family of probability measures $(\mu_t)_{t\geq 0}$, where each μ_t is the distribution of X_t. In other words, μ_t is not chosen independently, but is determined by the evolution of the particle system. The evolution of $(\mu_t)_{t\geq 0}$ is governed by the continuity equation

$$\partial_t \mu_t + div(\mu_t v_t) = 0, \tag{8.3}$$

which enforces conservation of mass.

The dynamic formulation of optimal transport (Benamou–Brenier, 2000) expresses the squared 2-Wasserstein distance as the minimal kinetic energy needed to move μ_0 to μ_1:

$$W_2^2(\mu_0, \mu_1) = \inf_{(\mu_t, v_t)} \left\{ \int_0^1 \int_{\mathbb{R}^d} \|v_t(x)\|^2 \, \mu_t(dx)\, dt \;\middle|\; \begin{array}{ll} \partial_t \mu_t + \operatorname{div}(\mu_t v_t) = 0, & \\ \mu_{t=0} = \mu_0, & \mu_{t=1} = \mu_1 \end{array} \right\}. \tag{8.4}$$

Here v_t describes the velocity field transporting the mass distribution, while the objective measures the total expected kinetic energy of the transport. In this way, $W_2^2(\mu_0, \mu_1)$ quantifies the most efficient way to continuously deform one probability distribution into another.

Moreover, the minimizing curve $(\mu_t)_{t\in[0,1]}$ is uniquely realized as follows: if (X_0, X_1) is sampled from an optimal coupling $\bar{\gamma} \in \Gamma(\mu_0, \mu_1)$, then setting

$$X_t = (1-t)X_0 + tX_1$$

defines a random trajectory whose distribution is μ_t. In other words, μ_t is obtained as the distribution of particles moving in straight lines at constant speed from X_0 to X_1. This interpolation (μ_t) is called the *displacement interpolation* or *Wasserstein geodesic*; for more details see Villani [9, Chapter 5.2] in Topics in Optimal Transportation (2003).

Chapter 9

Gradient Flows

9.1 Background Knowledge

Definition **9.1.1**

A **Geodesic** in a metric space (X, d) is the locally shortest path between two points $x, y \in X$. Formally, a curve $(\gamma_t)_{t\in[0,1]}$ is a geodesic if

$$d(\gamma_s, \gamma_t) = |s - t|\, d(x, y), \quad \gamma_0 = x,\, \gamma_1 = y.$$

For example:

- In a standard Euclidean plane ($\mathbb{R}^n$), the geodesic between two points is simply the straight line segment connecting them.
- On the surface of a sphere, the geodesic between two points is an arc of a great circle - a circle on the sphere whose center is also the center of the sphere. Taking the earth as a sphere, this is why long-distance flight paths appear curved on a flat map.

A related notion that will be important for gradient flows on Wasserstein space is *geodesic convexity*.

***Definition* 9.1.2**

A set $C \subset X$ in a metric space (X, d) is called **geodesically convex** if for any $x_0, x_1 \in C$ and any geodesic $(\gamma_t)_{t\in[0,1]}$ joining them,

$$\gamma_t \in C, \qquad \forall\, t \in [0,1].$$

A functional $\mathcal{F} : X \to \mathbb{R}$ is called **α-geodesically convex**, $\alpha \in \mathbb{R}$, if for any $x_0, x_1 \in X$ and any geodesic $(\gamma_t)_{t\in[0,1]}$ joining them,

$$\mathcal{F}(\gamma_t) \;\le\; (1-t)\,\mathcal{F}(x_0) + t\,\mathcal{F}(x_1) - \frac{\alpha}{2}\,t(1-t)\,d(x_0,x_1)^2, \qquad \forall\, t \in [0,1].$$

When $\alpha = 0$, this reduces to ordinary geodesic convexity. When $\alpha > 0$, the functional has a curvature-like property analogous to strong convexity in Euclidean space.

For more details on strong-convexity, see Boyd and Vandenberghe [2, Chapter 9.1]

***Definition* 9.1.3**

Let $V : \mathbb{R}^d \to \mathbb{R}$ be differentiable. The **gradient** of V at x is the unique vector $g_x \in \mathbb{R}^d$ such that for all directions $h \in \mathbb{R}^d$, a first-order Taylor expansion yields

$$V(x + \varepsilon h) = V(x) + \varepsilon\langle g_x, h\rangle + o(\varepsilon), \quad \varepsilon \to 0,$$

where $\langle\cdot,\cdot\rangle$ is the chosen inner product.

For example, if $\langle\cdot,\cdot\rangle$ is the standard Euclidean inner product, then $g_x = \nabla V(x)$, which is the vector of all partial derivates at x.

***Definition* 9.1.4**

Consider the optimization problem

$$\inf_{x\in\mathbb{R}^d} V(x),$$

where $V : \mathbb{R}^d \to \mathbb{R}$ is sufficiently smooth.
The associated **gradient flow** is the trajectory x_t solving the ODE

$$\dot{x}_t = -\nabla V(x_t), \quad t \ge 0,$$

with x_0 as the initial condition. This continuous-time flow is the steepest descent path of V.

To see why the final statement is true, one can inspect the directional derivative

$$D_v V(x_t) = \langle \nabla V(x_t), v \rangle,$$

which measures the instantaneous change of V when moving in direction v. Among all unit directions $\|v\| = 1$, the Cauchy–Schwarz inequality gives

$$-\|\nabla V(x_t)\| \leq \langle \nabla V(x_t), v \rangle \leq \|\nabla V(x_t)\|.$$

Using the definition of the inner product, with θ the angle between $\nabla V(x_t)$ and v,

$$\langle \nabla V(x_t), v \rangle = \|\nabla V(x_t)\| \|v\| \cos\theta = \|\nabla V(x_t)\| \cos\theta \geq -\|\nabla V(x_t)\|.$$

Equality holds only when v is antiparallel to the gradient (i.e. $\cos\theta = -1$), that is,

$$v = -\frac{\nabla V(x_t)}{\|\nabla V(x_t)\|}.$$

Thus, moving in direction $-\nabla V(x_t)$ yields the maximal instantaneous decrease of V, and the ODE $\dot{x}_t = -\nabla V(x_t)$ describes the steepest descent flow.

A natural quantity to examine is the rate of change of the function value along the flow:

$$\frac{d}{dt} V(x_t) = \sum_{i=1}^{d} \frac{\partial V}{dt}(x_t)\, \dot{x}_i = \langle \nabla V(x_t), \dot{x}_t \rangle = \langle \nabla V(x_t), -\nabla V(x_t) \rangle = -\|\nabla V(x_t)\|^2 \leq 0,$$

so $V(x_t)$ decreases monotonically with time.

***Example* 9.1.1**

Assume $x : \mathbb{R} \to \mathbb{R}^2$ and $f : \mathbb{R}^2 \to \mathbb{R}$. Then take $f(x_t) = [x_t]_2$ - f takes the second element of the vector x_t, its height.
The gradient flow ODE is

$$\dot{x}_t = -\begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

And the goal would be to find a trajectory x_t that minimizes f and solves the ODE.
Integrating the gradient of f, one finds taht

$$x_t = \begin{bmatrix} x_0^1 \\ x_0^2 - t \end{bmatrix}$$

where x_0^1, x_0^2 represent the initial points of the flow.

The discrete case of gradient flows on $\mathbb{R}^d$ is of particular importance to machine learning. It forms the basis of the gradient descent algorithm, which is

used to optimize a weight matrix $w \in \mathbb{R}^{d\times d}$ (or any parameter vector) in order to minimize a loss function

$$L(y, \hat{y}, w),$$

where y denotes the ground-truth label and $\hat{y}$ the model prediction.

To relate gradient descent to gradient flows, recall the continuous-time equation

$$\dot{x}_t = -\nabla V(x_t).$$

When time is discretized with step size $\eta > 0$, the time derivative is replaced by a finite-difference approximation. Using the forward (explicit) Euler scheme,

$$\dot{x}_t \approx \frac{x_{k+1} - x_k}{\eta},$$

and evaluating the gradient at x_k, we obtain

$$\frac{x_{k+1} - x_k}{\eta} = -\nabla V(x_k), \qquad \Longrightarrow \qquad x_{k+1} = x_k - \eta \nabla V(x_k).$$

From an optimization perspective, much more details on forward Euler can be found in Boyd and Vandenberghe [2, Chapter 9.3].

In machine learning, this is exactly the gradient descent update

$$w_{k+1} = w_k - \eta \, \nabla_w L(y, \hat{y}, w_k).$$

The step size, or *learning rate*, η_k may depend on the iteration index k, as in adaptive optimization methods such as the Adam optimizer [5].

For more details on gradient descent in ML, see Russell and Norvig [8, Chapter 19.6].

A different discretization uses the backward (implicit) Euler method, where the gradient is evaluated at the future point x_{k+1}:

$$\frac{x_{k+1} - x_k}{\eta} = -\nabla V(x_{k+1}), \qquad \Longleftrightarrow \qquad x_{k+1} = x_k - \eta \nabla V(x_{k+1}).$$

This update is implicit, since x_{k+1} appears on both sides, and typically requires solving a nonlinear equation at each iteration. Its main appeal is the associated variational characterization:

$$x_{k+1} = \arg \inf_{x\in\mathbb{R}^d} \left\{ V(x) + \frac{1}{2\eta}\|x - x_k\|^2 \right\},$$

in which the right-hand side is known as the proximal operator of V. To learn more, see Section 4.1 of Parikh and Boyd [6].

Forward Euler leads directly to the standard gradient descent algorithm widely used in practice, while backward Euler serves as a more stable but implicit proximal update.

9.2 Theory

9.2.1 The General Case

We now turn to the notion of gradient flows on the Wasserstein space $\mathcal{P}_2(\mathbb{R}^d)$ of probability measures with finite second moment endowed with the Wasserstein 2-distance. The goal is to extend the Euclidean definition of gradient flows to a space of probability measures.

Given a functional $\mathcal{F} : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$, the first variation of $\mathcal{F}$ at μ, denoted $\delta\mathcal{F}(\mu) : \mathbb{R}^d \to \mathbb{R}$, is defined through

$$\lim_{\varepsilon\to 0} \frac{\mathcal{F}(\mu + \varepsilon\chi) - \mathcal{F}(\mu)}{\varepsilon} = \int_{\mathbb{R}^d} \delta\mathcal{F}(\mu)(x)\, d\chi(x), \tag{9.1}$$

for all signed measures χ satisfying $\int d\chi = 0$, which ensures the perturbation stays within the space of probability measures. This plays the same role as the directional derivative in $\mathbb{R}^d$:

$$\lim_{\varepsilon\to 0} \frac{F(x + \varepsilon h) - F(x)}{\varepsilon} = \nabla F(x) \cdot h, \tag{9.2}$$

for smooth $F : \mathbb{R}^d \to \mathbb{R}$ and vectors $h \in \mathbb{R}^d$.

Now let $(\mu_t)_{t\geq 0}$ be a curve in $\mathcal{P}_2(\mathbb{R}^d)$. The time derivative of $\mathcal{F}(\mu_t)$ is defined by

$$\partial_t \mathcal{F}(\mu_t) = \lim_{\varepsilon\to 0} \frac{\mathcal{F}(\mu_{t+\varepsilon}) - \mathcal{F}(\mu_t)}{\varepsilon}. \tag{9.3}$$

For small $\varepsilon > 0$, we write the first-order expansion

$$\mu_{t+\varepsilon} \approx \mu_t + \varepsilon\, \partial_t \mu_t, \tag{9.4}$$

where $\partial_t \mu_t$ is determined by the continuity equation

$$\partial_t \mu_t + \nabla\cdot (\mu_t v_t) = 0, \tag{9.5}$$

for some velocity field v_t.

Using this approximation gives

$$\partial_t \mathcal{F}(\mu_t) = \lim_{\varepsilon\to 0} \frac{\mathcal{F}(\mu_t + \varepsilon\, \partial_t \mu_t) - \mathcal{F}(\mu_t)}{\varepsilon}. \tag{9.6}$$

By the definition of the first variation (with $\chi = \partial_t \mu_t$),

$$\partial_t \mathcal{F}(\mu_t) = \int_{\mathbb{R}^d} \delta\mathcal{F}(\mu_t)(x)\, \partial_t \mu_t(x)\, dx. \tag{9.7}$$

Since $\partial_t \mu_t = -\nabla\cdot (\mu_t v_t)$,

$$\partial_t \mathcal{F}(\mu_t) = \int \delta\mathcal{F}(\mu_t)\, \partial_t \mu_t\, dx \tag{9.8}$$

$$= -\int \delta\mathcal{F}(\mu_t)(x)\, \nabla\cdot (\mu_t v_t)(x)\, dx. \tag{9.9}$$

Applying integration by parts,

$$\partial_t \mathcal{F}(\mu_t) = \int \nabla \delta \mathcal{F}(\mu_t)(x) \cdot v_t(x)\, \mu_t(dx) \tag{9.10}$$

$$= \left\langle \nabla \delta \mathcal{F}(\mu_t),\, v_t \right\rangle_{\mu_t}. \tag{9.11}$$

Thus the rate of change of $\mathcal{F}$ along the curve μ_t satisfies

$$\partial_t \mathcal{F}(\mu_t) = \int_{\mathbb{R}^d} \langle \nabla \delta \mathcal{F}(\mu_t)(x),\, v_t(x) \rangle\, \mu_t(dx) \;=\; \left\langle \nabla \delta \mathcal{F}(\mu_t),\, v_t \right\rangle_{\mu_t}. \tag{9.12}$$

The expression above is the exact analogue of the Euclidean identity

$$\frac{d}{dt} V(x_t) = \langle \nabla V(x_t), \dot{x}_t \rangle,$$

where the rate of change of a function along a curve is given by the inner product between its gradient and the curve's velocity. In Wasserstein space, the role of $\dot{x}_t$ is played by the vector field v_t, and the analogue of the Euclidean gradient is the vector field $\nabla \delta \mathcal{F}(\mu_t)$, which we call the Wasserstein gradient ∇_{W2}

***Example* 9.2.1**

1. Potential Energy. Let

$$\mathcal{F}(\mu) := \int_{\mathbb{R}^d} V(x)\, d\mu(x),$$

for a smooth potential $V : \mathbb{R}^d \to \mathbb{R}$. Then

$$\delta\mathcal{F}(\mu) = V, \qquad \nabla_{W_2}\mathcal{F}(\mu) = \nabla V.$$

2. Internal Energy. Let

$$\mathcal{F}(\mu) := \int_{\mathbb{R}^d} U(\mu(x))\, dx,$$

for a smooth function $U : \mathbb{R}_+ \to \mathbb{R}$. Then

$$\delta\mathcal{F}(\mu) = U'(\mu), \qquad \nabla_{W_2}\mathcal{F}(\mu) = \nabla\big(U'(\mu)\big).$$

3. Entropy as a Special Case of Internal Energy. The Boltzmann–Shannon entropy is

$$\mathrm{Ent}(\mu) := -\int_{\mathbb{R}^d} \mu(x) \log \mu(x)\, dx.$$

This corresponds to the internal energy functional with

$$U(s) = -\, s \log s.$$

Then

$$U'(s) = -(\log s + 1),$$

so that

$$\delta\,\mathrm{Ent}(\mu) = U'(\mu) = -(\log \mu + 1), \qquad \nabla_{W_2}\,\mathrm{Ent}(\mu) = \nabla\big(U'(\mu)\big) = -\nabla \log \mu.$$

As in Euclidean space, this gradient gives the direction of steepest increase of $\mathcal{F}$.

Consequently, the direction of steepest descent for $\mathcal{F}$ is the negative Wasserstein gradient:

$$v_t^* = -\,\nabla_{W_2}\,\mathcal{F}(\mu_t) = -\,\nabla\delta\mathcal{F}(\mu_t).$$

Substituting this optimal velocity into the continuity equation

$$\partial_t \mu_t + \nabla\cdot\,(\mu_t v_t) = 0,$$

we obtain the Wasserstein gradient flow associated with $\mathcal{F}$:

$$\begin{aligned}\partial_t \mu_t &= -\nabla\cdot(\mu_t v_t^*)\\ &= -\nabla\cdot(\mu_t\cdot(-\nabla\delta\mathcal{F}(\mu_t)))\\ &= \nabla\cdot(\mu_t\nabla\delta\mathcal{F}(\mu_t))\,.\end{aligned}$$

If $\mathcal{F}$ is α-convex on $\mathcal{P}_2(\mathbb{R}^d)$, $\alpha>0$, then its Wasserstein gradient flow $(\mu_t)_{t\geq 0}$ converges exponentially fast to a minimizer of $\mathcal{F}$; if $\mu^\star$ is the unique minimizer of $\mathcal{F}$, then for all $t\geq 0$,

$$\mathcal{F}(\mu_t)-\mathcal{F}(\mu^\star)\ \leq\ e^{-2\alpha t}\left(\mathcal{F}(\mu_0)-\mathcal{F}(\mu^\star)\right).$$

9.2.2 The Gaussian Case

Take $\mu_0,\mu_1\in\mathcal{P}_2(\mathbb{R}^d)$ to be non-degenerate normal distributions with means $m_0,m_1\in\mathbb{R}^d$ and covariance matrices $\Sigma_0,\Sigma_1\in\mathbb{R}^{d\times d}$:

$$\mu_0=\mathcal{N}(m_0,\Sigma_0),\qquad \mu_1=\mathcal{N}(m_1,\Sigma_1).$$

By Brenier's theorem, the optimal transport map pushing μ_0 to μ_1 is of the form

$$T(x_0)=Ax_0+b, \tag{9.13}$$

where $A\in\mathbb{R}^{d\times d}$, $b\in\mathbb{R}^d$. A is a positive definite symmetric matrix, and x_0 is a random vector distributed according to μ_0.

To represent the constraint that $T_\#\mu_0=\mu_1$, we compute the mean and covariance of the transported random variable $T(x_0)$:

$$\begin{aligned}\mathbb{E}[T(x_0)] &= \mathbb{E}[Ax_0+b]\\ &= A\,\mathbb{E}[x_0]+b\\ &= Am_0+b.\end{aligned}$$

Matching this with the mean of μ_1 gives the condition

$$Am_0+b=m_1. \tag{9.14}$$

Similarly, the covariance of the pushforward distribution is

$$\begin{aligned}\mathrm{Var}(T(x_0)) &= \mathrm{Var}(Ax_0+b)\\ &= A\ \mathrm{Var}(x_0)\,A^\top\\ &= A\Sigma_0A^\top.\end{aligned}$$

Matching this with the covariance of μ_1 yields the second condition

$$A\Sigma_0A^\top=\Sigma_1. \tag{9.15}$$

Solving the covariance constraint yields the unique symmetric positive semidefinite solution

$$A=\Sigma_0^{-1/2}\left(\Sigma_0^{1/2}\Sigma_1\Sigma_0^{1/2}\right)^{1/2}\Sigma_0^{-1/2}. \tag{9.16}$$

Using the mean constraint $Am_0 + b = m_1$, we get

$$b = m_1 - Am_0.$$

Therefore the optimal transport map is

$$T(x) = \Sigma_0^{-1/2}\big(\Sigma_0^{1/2}\Sigma_1\Sigma_0^{1/2}\big)^{1/2}\Sigma_0^{-1/2}\,(x - m_0) + m_1. \tag{9.17}$$

This explicit formula reveals the fact that the squared Wasserstein-2 distance between Gaussian measures has a closed-form:

$$W_2^2(\mu_0, \mu_1) = \|m_0 - m_1\|^2 + \operatorname{tr}\Big[\Sigma_0 + \Sigma_1 - 2\left(\Sigma_0^{1/2}\Sigma_1\Sigma_0^{1/2}\right)^{1/2}\Big]. \tag{9.18}$$

To see the derivations and more details, see Villani [9, Example 1.19] in Topics in Optimal Transportation (2003).

To formulate gradient flows on Gaussian measures, we first introduce the structure of the space they form. The family of non-degenerate Gaussian measures on $\mathbb{R}^d$ constitutes a smooth submanifold of the Wasserstein space $\mathcal{P}_2(\mathbb{R}^d)$. Endowed with the Wasserstein-2 metric, this manifold is known as the Bures-Wasserstein space, denoted by $\mathrm{BW}(\mathbb{R}^d)$.

Since a Gaussian measure is fully characterized by its mean vector and covariance matrix, we may identify $\mathrm{BW}(\mathbb{R}^d)$ with the product space

$$\mathrm{BW}(\mathbb{R}^d) \;\cong\; \mathbb{R}^d \times \mathcal{S}_{++}^d, \tag{9.19}$$

where $\mathcal{S}_{++}^d$ denotes the set of $d \times d$ symmetric positive definite matrices. Thus every element of $\mathrm{BW}(\mathbb{R}^d)$ can be written as a pair (m, Σ) representing the Gaussian $\mathcal{N}(m, \Sigma)$.

From the explicit formula (49) for the affine optimal transport map between any two Gaussian measures μ_0, μ_1, the W_2–geodesic connecting them is given by

$$\mu_t = \big[(1-t)\,\mathrm{id} + t\,T\big]_{\#}\mu_0, \qquad t \in [0,1], \tag{9.20}$$

where T is the optimal affine transport map pushing μ_0 to μ_1.

Since the pushforward of a Gaussian measure under any affine map is again Gaussian, each μ_t is Gaussian for all $t \in [0,1]$. This shows that the Bures–Wasserstein space $\mathrm{BW}(\mathbb{R}^d)$ is geodesically convex: any two points in $\mathrm{BW}(\mathbb{R}^d)$ are joined by a geodesic that stays entirely within the space.

This geodesic convexity has an immediate consequence. Any functional $\mathcal{F}$ that is α–geodesically convex on the full Wasserstein space $\mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$ remains α–geodesically convex when restricted to the Bures–Wasserstein space. This allows one to study convexity properties of functionals directly on $\mathrm{BW}(\mathbb{R}^d)$ without having to reprove them.

Before defining gradient flows on this space, we first introduce the analogue of the Wasserstein gradient restricted to Gaussian measures. Recall that the

Wasserstein gradient of a functional $\mathcal{F}$ at a measure μ is the vector field $\nabla\delta\mathcal{F}(\mu)$. To obtain the Buresx–Wasserstein gradient, denoted $\nabla_{\mathrm{BW}}\mathcal{F}(\mu)$, we project this vector field onto the tangent space of $\mathrm{BW}(\mathbb{R}^d)$ at the Gaussian $\mu = \mathcal{N}(m, \Sigma)$.

This projection produces an affine vector field of the form

$$\nabla_{\mathrm{BW}}\mathcal{F}(\mu)(x) = \Big(\int \nabla^2\delta\mathcal{F}(\mu)\, d\mu\Big)(x - m) \; + \; \int \nabla\delta\mathcal{F}(\mu)\, d\mu.$$

In other words, the BW gradient is the closest affine vector field that approximates the full Wasserstein gradient while remaining tangent to the Gaussian manifold. This is the natural gradient direction available when the dynamics are constrained to evolve within the space of Gaussian measures.

Given this gradient, the corresponding gradient flow on $\mathrm{BW}(\mathbb{R}^d)$ is described by the evolution equation

$$\dot{X}_t = -\nabla_{\mathrm{BW}}\mathcal{F}(\mu_t)(X_t),$$

where $X_t \sim \mu_t$. Substituting the expression for the BW gradient gives

$$\dot{X}_t = -\Big(\int \nabla^2\delta\mathcal{F}(\mu_t)\, d\mu_t\Big)(X_t - m_t) \; - \; \int \nabla\delta\mathcal{F}(\mu_t)\, d\mu_t,$$

with m_t the mean of μ_t.

Since each μ_t remains Gaussian for all $t \geq 0$, it is more convenient to track the flow directly through the mean m_t and covariance Σ_t. Taking expectations yields the evolution of the mean,

$$\dot{m}_t = -\int \nabla\delta\mathcal{F}(\mu_t)\, d\mu_t,$$

and differentiating the covariance identity $\Sigma_t = \mathbb{E}[(X_t - m_t)(X_t - m_t)^\top]$ gives

$$\dot{\Sigma}_t = -\Big(\int \nabla^2\delta\mathcal{F}(\mu_t)\, d\mu_t\Big)\Sigma_t - \Sigma_t\Big(\int \nabla^2\delta\mathcal{F}(\mu_t)\, d\mu_t\Big).$$

Thus the Bures–Wasserstein gradient flow of $\mathcal{F}$ within the Gaussian manifold is characterized by the coupled system

$$\mu_t = \mathcal{N}(m_t, \Sigma_t), \qquad t \geq 0,$$

with (m_t, Σ_t) evolving according to the equations above.

For the full derivations, see Chewi, Niles-Weed, and Rigollet [3, Chapter 5.5] in Statistical Optimal Transport (2019).

Chapter 10

Statistical Applications

10.1 Background Knowledge

***Definition* 10.1.1**

A large class of probability distributions on $\mathbb{R}^d$ can be written in the form

$$\mu(dx) = \frac{1}{Z}\, e^{-V(x)}\, dx,$$

where $V : \mathbb{R}^d \to \mathbb{R}$ is a measurable function, called the **potential**, and

$$Z = \int_{\mathbb{R}^d} e^{-V(x)}\, dx$$

is the normalizing constant (or **partition function**), assumed to be finite.
Such distributions arise naturally in statistics, statistical mechanics, and Bayesian inference. When V is convex, the corresponding measure is log-concave; many commonly used distributions (including Gaussian and exponential families) belong to this class.

***Definition* 10.1.2**

A probability measure π is called a **stationary distribution** for a time-evolving probability measure $(\mu_t)_{t\geq 0}$ if

$$\mu_t = \pi \quad \text{for all } t \geq 0,$$

or equivalently, if π is an invariant solution of the associated evolution equation.
For a general introduction to stationary processes and invariant distributions, see Karlin and Taylor [4, Chapter 9].

10.2 Theory

10.2.1 Variational Inference

In Variational Inference (VI), the goal is to approximate a target distribution π by selecting q from a simpler family of probability measures $\mathcal{Q}$. This is usually formulated as the optimization problem

$$q^* = \arg\min_{q\in\mathcal{Q}} KL(q \,\|\, \pi).$$

Throughout, we assume that the target distribution π admits a density of the form $\pi(x) \propto e^{-V(x)}$, where $V : \mathbb{R}^d \to \mathbb{R}$ is called the potential function. Writing the normalized form explicitly,

$$\pi(x) = \frac{e^{-V(x)}}{Z}, \qquad Z = \int_{\mathbb{R}^d} e^{-V(y)}\, dy,$$

Define the functional

$$\mathcal{F}(\mu) = KL(\mu \,\|\, \pi) = \int_{\mathbb{R}^d} \mu(x)\, \log \frac{\mu(x)}{\pi(x)}\, dx.$$

Substituting the expression for π gives

$$\mathcal{F}(\mu) = \int \mu(x) \left[\log \mu(x) + V(x) + \log Z\right] dx.$$

Since $\log Z$ does not depend on μ, it may be discarded. Thus the KL divergence decomposes as

$$\mathcal{F}(\mu) = \underbrace{\int V(x)\, d\mu(x)}_{\mathcal{V}(\mu)} + \underbrace{\int \mu(x) \log \mu(x)\, dx}_{\mathcal{H}(\mu)} + \text{const.}$$

The KL divergence therefore splits into a potential energy term $\mathcal{V}(\mu)$ and an entropy term $\mathcal{H}(\mu)$. Both terms have simple Wasserstein gradients:

$$\nabla\mathcal{V}(\mu) = \nabla V, \qquad \nabla\mathcal{H}(\mu) = \nabla \log \mu.$$

Hence the Wasserstein gradient of the KL divergence is

$$\nabla \mathcal{F}(\mu) = \nabla V + \nabla \log \mu = \nabla \log \frac{\mu}{\pi}.$$

Consequently, the associated Wasserstein gradient flow satisfies the Fokker-Plank PDE

$$\partial_t \mu_t = \nabla \cdot \Big(\mu_t \nabla \log \frac{\mu_t}{\pi} \Big) = \nabla \cdot \big(\mu_t (\nabla V + \nabla \log \mu_t) \big),$$

If the potential V is α-convex and the set of admissible measures Q is geodesically convex, then the functional $KL(\cdot \,\|\, \pi)$ is α-geodesically convex over Q. In particular, the minimizer q^* of the VI problem is unique, and the Wasserstein gradient flow q_t converges to q^* exponentially fast:

$$KL(q_t \,\|\, \pi) - KL(q^* \,\|\, \pi) \;\le\; e^{-2\alpha t} \big(KL(q_0 \,\|\, \pi) - KL(q^* \,\|\, \pi) \big).$$

For complete set of derivations, see Chewi, Niles-Weed, and Rigollet [3, Chapter 6.1] in Statistical Optimal Transport (2019).

10.2.2 Sampling

Consider the task of sampling from a target distribution π with density proportional to e^{-V}, where $V : \mathbb{R}^d \to \mathbb{R}$ is a smooth potential function. The goal is to produce a random variable whose distribution is close to π, a problem of central importance in fields like Bayesian inference and Machine Learning.

A classical method is the Langevin diffusion, defined by the stochastic differential equation

$$dX_t = -\nabla V(X_t)\, dt + \sqrt{2}\, dB_t,$$

where $(B_t)_{t \ge 0}$ is standard Brownian motion. Under mild regularity assumptions on V, this diffusion has a unique strong solution, and its stationary distribution is exactly π.

Let μ_t denote the distribution of the random variable X_t. Although X_t evolves randomly, its distribution evolves deterministically according to the Fokker–Planck equation

$$\partial_t \mu_t = \Delta \mu_t + \nabla \cdot (\mu_t \nabla V).$$

A key observation is that the right-hand side of this equation is precisely the Wasserstein gradient of the KL divergence. From the identity

$$\nabla \mathcal{F}(\mu) = \nabla \log \frac{\mu}{\pi} = \nabla V + \nabla \log \mu,$$

the Wasserstein gradient flow of $KL(\cdot \,\|\, \pi)$ is

$$\partial_t \mu_t = \nabla \cdot \Big(\mu_t \nabla \log \frac{\mu_t}{\pi} \Big).$$

Expanding the logarithm yields

$$\partial_t \mu_t = \Delta \mu_t + \nabla \cdot (\mu_t \nabla V),$$

which is exactly the Fokker–Planck equation above. Thus, the Langevin diffusion is a stochastic realization of the deterministic Wasserstein gradient flow that decreases the KL divergence $KL(\mu_t \,\|\, \pi)$ over time.

When the potential V is α-convex, the KL divergence is α-geodesically convex, and the resulting gradient flow converges exponentially fast to the target distribution π:

$$KL(\mu_t \,\|\, \pi) \;\le\; e^{-2\alpha t}\, KL(\mu_0 \,\|\, \pi).$$

For additional background and a full derivation of this equivalence, see Chewi, Niles-Weed, and Rigollet [3, Chapter 6.2].

Bibliography

[1] Samuel E. Andersson, Anders Björn, and David Wiman. *An Introduction to Metric Spaces*. 2022. Course manuscript.

[2] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[3] Sinho Chewi, Jonathan Niles-Weed, and Philippe Rigollet. *Statistical Optimal Transport*. Lecture Notes in Mathematics. Springer Cham, Cham, 2025.

[4] Samuel Karlin and Howard M. Taylor. *A First Course in Stochastic Processes*. Academic Press, 2nd edition, 1975.

[5] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations (ICLR)*, 2015.

[6] Neal Parikh and Stephen Boyd. Proximal algorithms. *Foundations and Trends in Optimization*, 1(3), 2014.

[7] H. L. Royden and P. M. Fitzpatrick. *Real Analysis*. Pearson, Boston, MA, 2010.

[8] Stuart Russell and Peter Norvig. *Artificial Intelligence: A Modern Approach*. Prentice Hall Press, USA, 3rd edition, 2009.

[9] Cédric Villani. *Topics in Optimal Transportation*, volume 58 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2003.