992:Optimal Transport and Applications to Machine Learning Lecture Notes

November 27, 2023

1 Lecture 1: 09/07/2023

Scribes: Nathan Aviles and Congwei Yang

Nicolas introduced the course with an example where reframing your problem in the language of Optimal Transport may be illuminating and helpful. In his own work he was faced with the following, deliberately abstracted, situation.

Given a set of data points $\{x_1, \dots, x_n\}$ coming from a space $\mathcal{D}$, how might one compare two functions $f : \{x_1, \dots, x_n\} \to \mathbf{R}$ and $g : \mathcal{D} \to \mathbf{R}$, for instance, whether f and g are "similar"? Standard metrics are especially unsatisfactory here, something like a L_p-norm

$$||f - g||_{L_p(\mu)} = \left(\int (f(x) - g(x))^p \, \mathrm{d}\mu(x) \right)^{\frac{1}{p}}$$

would of course require that $f, g \in L_p(\mu)$, with respect to some measure μ supported on $\mathcal{D}$, which may not be the case (as stated, f is not even defined generally on $\mathcal{D}$).

We are hinted that in some way, we may see in future, the lens of Optimal Transport became fruitful for his work. He also elaborated that this problem is very similar to what is discussed in Machine Learning as Domain Adaptation; where one might like to extend a function, say f, from where it is originally defined, say $\mathcal{X} = \{x_1, \dots, x_n\}$, to a larger or different domain and environment, say $\mathcal{D}$, where it is important to deduce the properties preserved by it's extension.

1.1 Framing of Optimal Transport

What even is a transport?

Consider the set of "sources" $\mathcal{X} = \{x_1, \dots, x_n\}$ each with a "supply"/mass, $\mu = (\mu_1, \dots, \mu_n)$, and a set of "targets" $\mathcal{Y} = \{y_1, \dots, y_n\}$ with a "demand" of that supply/mass $\nu = (\nu_1, \dots \nu_m)$.

We would like to find a way to transport our "supply" from our "sources" to our "targets" in order to satisfy their "demands". For simplicity we assume that we do indeed have all the mass needed to theoretically do this, i.e.

$$\sum_{i=1}^{n} \mu_i = \sum_{j=1}^{m} \nu_i$$

and therefore without loss of generality we normalize and consider

$$\sum_{i=1}^{n} \mu_i = 1$$

These are now probability measures, more specifically μ and ν are probability mass function (pmf) over $\mathcal{X}$ and $\mathcal{Y}$ respectively.

A transport between these two measure is then a function

$$T : \{x_1, \ldots, x_n\} \to \{y_1, \ldots, y_m\}$$

which of course directs the mass at source x_i to some target y_j. We require that the amount of mass at point y_j is equal to what is demanded, i.e.

$$\nu_j = \sum_{i:T(x_i)=y_j} \mu_i$$

and that we incur a cost per unit mass $c(x, y)$ for said transport, $c : \mathcal{X} \times \mathcal{Y} \to \mathbf{R} \cup \{+\infty\}$ that is assumed to be lower-semi-continuous (lsc), i.e. f is said to be lsc in x if for a sequence

$$x_n \to x \quad \implies \quad f(x) \leq \liminf_n f(x_n)$$

1.2 Monge's Problem

This is the framing and starting for something called Monge's Optimal Transport Problem (MP) where we have the constraint that:

"All mass at a given source point must go to a single target point"

and are asked to construct a transport plan T minimizing the transport cost of moving $(\mu \to \nu)$, or rather

$$\min_{T:\nu_j=\sum_{i:T(x_i)=y_j}\mu_i} \sum_i c(x_i, T(x_i))\mu_i$$

This of course may be impossible! This says that T must be 1-to-1, and that no mass may be split so trivially transport from $\mu = (1, 0)$ to $\nu = (1/2, 1/2)$ is not feasible.

(1) commentary by Nathan Aviles: The French Baker and French Tailor

Imagine France has gone through a modern revolution, a Purge if you will. The population is down to just two people: a French Baker, principled and egalitarian as one might expect, and a French Tailor, who happens to have differing politics to the Baker.

They are both hungry and have a demand for bread ($\mathcal{Y} = \{Baker, Tailor\} = \mathcal{X}$). The Baker has just produced 1 loaf of bread ($\mu = (1, 0)$). Both of their hungers will be sated if they are willing to share the bread ($\nu = (1/2, 1/2)$), but the Baker will not have this. He cannot break bread with someone so unscrupulous. He is however French, and in all fairness, he is willing to allow them in solidarity to both starve together.

(1) End of commentary

1.3 Generalization

We will now generalize the problem. Let $\mathcal{X}, \mathcal{Y}$ be abstract spaces (for this class usually Metric Spaces), with $\mu \in \mathcal{P}(\mathcal{X})$ and $\nu \in \mathcal{P}(\mathcal{Y})$, where $\mathcal{P}(\mathcal{X})$ denotes the space of probability measures supported on $\mathcal{X}$.

For our Transport $T : \mathcal{X} \to \mathcal{Y}$, a measurable map, it is now useful to introduce the following. Define the push-forward of T as

$$T_\sharp \mu(A) = \mu(T^{-1}(A)), \quad \forall A \subseteq \mathcal{Y}$$

this is the mass transported by T coming from $\mathcal{X}$ to A. Thus we rephrase Monge's Problem as

$$\inf_{T:T_\sharp\mu=\nu} \int_{\mathcal{X}} c(x, T(x)) \mathrm{d}\mu(x)$$

where here we use an infimum as it is unclear whether a solution, a unique minimizer, exists in the search space (e.g. the search space as before may be empty, as in the example of baker and tailor).

1.4 Kantorovich's Problem

Monge's problem may sound silly in retrospect, so we may attempt to relax our constraints such that

"Mass can be split at source points"

To articulate this idea mathematically we introduce a coupling $\pi \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$, a joint probability measure which has marginals corresponding to our source and target

$$\pi(\cdot, \mathcal{Y}) = \mu, \quad \pi(\mathcal{X}, \cdot) = \nu$$

and can think of $\pi(A \times B)$ as "the amount of mass sent from source points in A to target points in B". This has the natural consequence that for bounded measurable functions ϕ, ψ

$$\int \phi(x) \mathrm{d}\pi(x, y) = \int \phi(x) \mathrm{d}\mu(x)$$

$$\int \psi(y) \mathrm{d}\pi(x, y) = \int \psi(y) \mathrm{d}\nu(y)$$

The discrete interpretation of these couplings is a matrix $\pi = \{\pi_{i,j}\}$ where $\pi_{i,j}$ = "mass sent from x_i to y_j" for which row sums and column sums "marginalize" i.e.

$$\sum_j \pi_{i,j} = \mu_i, \quad \sum_i \pi_{i,j} = \nu_j$$

with this coupling we define Kantorovich's Problem, an optimization over the space of couplings (joint probability measures) $\Gamma(\mu, \nu)$ as

$$\inf_{\pi \in \Gamma(\mu,\nu)} \int_{\mathcal{X} \times \mathcal{Y}} c(x, T(x)) \mathrm{d}\pi(x, y)$$

here we are luckier and have that $\Gamma(\mu, \nu)$ is at least always non-empty, as the product measure

$$\mu \otimes \nu(A \times B) = \mu(A)\nu(B)$$

is trivially included.

(2) commentary by Nathan Aviles: Polish Spaces

Polish Spaces may be mentioned, don't be scared, this is essentially just a "nice" abstract space where there is a metric and limits do what you'd like. Specifically it is separable, and "completely metrizable" i.e. you can give it a metric for which if a sequence converges in this metric, it's limit is contained in this space; the separable part allows you to approach any point via a subset of some countable set, think here the rationals, or for probabilistic arguments often simple functions.

(2) End of commentary

(3) commentary by Nathan Aviles: Lower-semi-continuity

These functions are nice and common, as we will see later where in useful cases they induce lower-semi-continuity (lsc) of integrals of these function (i.e. if we pass a lsc function into a functional we may end up preserving lsc in some cases). For example, in Probability it is often very useful to work with something called the Generalized Inverse of a Cumulative Distribution Function. One can show that the generalized inverse of a CDF, F, denoted

$$F^-(y) = \inf\{x : F(x) > y\}$$

is actually lower-semi-continuous. This can be seen because F is monotone increasing, always cadlag (continuous from the right, limit from the left), and therefore upper-semi-continuous; thus it's "inversion" defined here as a function on an open set is lower-semi-continuous. This is used in practice as if F is generated by probability P, and if $U \sim \mathit{Unif}[0, 1]$, we have $F^-(U) \sim P$.

(3) End of commentary

Theorem 1.4.1. *Let $(\mathcal{X}, \mu)$ and $(\mathcal{Y}, \nu)$ be two Polish probability spaces; let $a : \mathcal{X} \to \mathbb{R} \cup \{-\infty\}$ and $b : \mathcal{Y} \to \mathbb{R} \cup \{-\infty\}$ be two upper-semicontinuous functions such that $a \in L^1(\mu)$, $b \in L^1(\nu)$. Let $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R} \cup \{+\infty\}$ be a lower-semicontinuous cost function, such that $c(x, y) \geq a(x) + b(y)$ $\forall x, y$. Then there exists a optimal coupling $\Pi^* \in \Gamma(\mu, \nu)$ which minimizes the total cost $\int c(x, y)\, d\Pi(x, y)$ among all possible couplings of μ, ν.*

The lower bound assumption on cost function c guarantees the total cost $\int c(x, y)\, d\Pi(x, y)$ is well defined in $\mathbb{R} \cup \{\infty\}$. In most cases of application, one can choose $a = b = 0$.

The proof of the Theorem 1.1 relies on arguments of topology of weak convergence. We will present key ideas and supporting lemmas required but skip the detailed proof steps. One can check *Optimal transport, old and new* by Cedric Villani, p55-57 for further reference.

First let's recall Prokohorov theorem.

Theorem 1.4.2 (Prokohorov theorem)**.** *If $\mathcal{X}$ is a Polish space, then a set $\mathcal{P} \subset P(\mathcal{X})$ is precompact for the weak topology if and only if it is tight, i.e. for any $\epsilon > 0$ there is a compact set K_ϵ such that $\mu[\mathcal{X} \setminus K_\epsilon] \leq \epsilon$ for all $\mu \in \mathcal{P}$.*

Lemma 1.4.1 (Lower-semicontinuity of the cost functional)**.** *Let $\mathcal{X}$ and $\mathcal{Y}$ be two Polish spaces, and $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R} \cup \{+\infty\}$ a lower-semicontinuous cost function. Let $h : \mathcal{X} \times \mathcal{Y} \to \mathbb{R} \cup \{-\infty\}$ be an upper-semicontinuous function such that $c \geq h$. Let $(\Pi_k)_{k \in \mathbb{N}}$ be a sequence of probability measures on $\mathcal{X} \times \mathcal{Y}$, converging weakly to some $\Pi \in P(\mathcal{X} \times \mathcal{Y})$, in such a way that $h \in L^1(\Pi_k)$, $h \in L^1(\Pi)$, and*

$$\int_{\mathcal{X} \times \mathcal{Y}} h\, d\Pi_k \xrightarrow{k \to \infty} \int_{\mathcal{X} \times \mathcal{Y}} h\, d\Pi$$

Then

$$\int_{\mathcal{X} \times \mathcal{Y}} c\, d\Pi \leq \liminf_{k \to \infty} \int_{\mathcal{X} \times \mathcal{Y}} c\, d\Pi_k$$

In particular, if c is non-negative, then $F \to \int c\, d\Pi$ is lower-semicontinuous on $\mathcal{P}(\mathcal{X} \times \mathcal{Y})$, equipped with topology of weak convergence.

Lemma 1.4.2 (Tightness of transport plans)**.** *Let $\mathcal{X}$ and $\mathcal{Y}$ be two Polish spaces. Let $\mathcal{P} \subset P(\mathcal{X})$ and $\mathcal{Q} \subset P(\mathcal{Y})$ be tight subsets of $P(\mathcal{X})$ and $P(\mathcal{Y})$ respectively. Then the set $\Pi(\mathcal{P}, \mathcal{Q})$ of all transport plans whose marginals lie in $\mathcal{P}$ and $\mathcal{Q}$ respectively, is itself tight in $P(\mathcal{X} \times \mathcal{Y})$.*

The key idea of the proof of theorem 1 is to check (a) lower semicontinuity, and (b) compactness. By Polish space, we can obtain that $\{\mu\}$, $\{\nu\}$ are tight in $P(\mathcal{X})$ and $P(\mathcal{Y})$, respectively, and thus $\Gamma(\mu, \nu)$ is tight and thus precompact by Prokhorov's theorem.

Let Π_k be a sequence of probability measures such that $\int c\, d\Pi_k \to \inf_{\Pi \in \Gamma(\mu,\nu)} \int c\, d\Pi$. We may assume Π_k converges to some $\Pi \in \Gamma(\mu, \nu)$, extracting a subsequence if necessary. The function $h : (x, y) \to a(x) + b(y)$ lies in $L^1(\Pi_k)$ and $L^1(\Pi)$, and $c \geq h$ by assumption. Thus Lemma 1 implies

$$\liminf_{k \to \infty} \int c(x, y)\, d\Pi_k(x, y) \geq \int c(x, y)\, d\Pi(x, y)$$

and thus the minimizer exists.

1.5 Connection of Monge's and Kantorovich's problem

By the form of Monge's and Kantorovich's formulation of optimal transport, we can note that the Kantorovich's optimal transport is essentially a "relaxation" of Monge's optimal transport. Thus, we have

$$\inf_{T : T \# \mu = \nu} \int c(x, T(x))\, d\mu(x) \geq \inf_{\Pi \in \Gamma(\mu,\nu)} \int c(x, y)\, d\Pi(x, y)$$

When $\mathcal{X} = \{x_1, ..., x_n\}$, $\mathcal{Y} = \{y_1, ..., y_n\}$ and μ, ν are uniform measures on $\mathcal{X}$ and $\mathcal{Y}$ respectively, then Monge's formulation of optimal transport is equivalent to Kantorovich's formulation, and we can find the optimal transport map.

Also, the following Brenier theorem states another important connection between the two formulations.

Theorem 1.5.1 (Brenier). *Given μ and ν probability measures on a compact domain on $\mathbb{R}^d$ and $c(x, y) = \|x - y\|^2$. If μ has a density with respect to Lebesgue measure on $\mathbb{R}^d$, then the Monge's formulation is equivalent to Kantorovich's formulation, and there exists a unique optimal Monge map T. This map is characterized by being the unique gradient of a convex function $T = \nabla u$ such that $(\nabla u)\#\mu = \nu$.*

(4) commentary by Nathan Aviles: Non-emptiness of search space

A relatively minor, but I think interesting, observation is that by requiring a density at μ (absolute continuity with respect to the lebesgue measure) and discrete density at ν, we do not need to consider the earlier issue where "no mass-splitting" implies an empty search space of maps $T : \mathbf{X} \to \mathcal{Y}$. One can think of this continuity as allowing us to "finely" approximate any supply required by the demand at these discrete locations up to arbitrary precision.

As discussed briefly with Nicolas, I think the cleanest way to think about this is as a partition! Given a continuous density, we can generate a map which pulls it back to the uniform measure, or lebesgue on $[0, 1]$. From here it is easy to see that if I have demands $(\nu_1, \ldots, \nu_m)$ at points $(y_1, \ldots, y_m)$, all I need to do is find the "best partition" of $[0, 1]$, $(A_1, \ldots, A_m)$ of those sizes $(\nu_1, \ldots, \nu_m)$, relative to the cost $c(x, y)$. If the measurable map (inverse distribution function) inducing this transform to the uniform probability measure is continuous, then we are mapping open sets in the space of $\mathcal{X}$ to open sets in $[0, 1]$. This should for the most part provide us with "kind" partitions, where A_i are sets which are nice blobs of points "close" relative to the "distance" measured through $c(x, y) = ||x - y||_2^2$ as observed through the mapping, i.e. these are equivalent costs to $c(u, y) = ||F^-(u) - y||_2^2$ under a change of measure, $x = F^-(u)$.

This last little statement is common, as at times it may be easier to integrate over $[0, 1]^{\otimes 2}$ and our transport problem might be rephrased as

$$\int_{[0,1]^{\otimes 2}} ||F^-(u) - G^-(\tilde{u})||_2^2 \mathrm{d}\tilde{\pi}(u, \tilde{u})$$

which is nice in that these couplings are no longer as vague, but joint probability measures on marginally uniform random variables. This pops up classical Statistics literature when dealing with so called Statistical Functionals.

(4) end of commentary

1.6 The Dual of the Kantorovich's problem

As we previously introduced, the discrete Kantorovich's formulation can be written as the following linear programming problem. For $c_{ij} \geq 0$,

$$\min_{\Pi \in \Gamma(\mu,\nu)} \sum_{ij} c_{ij} \Pi_{ij}$$

under the constraints

$$\sum_j \Pi_{ij} = \mu_i, \ \forall i$$

$$\sum_i \Pi_{ij} = \nu_j, \ \forall j$$

$$\Pi_{ij} \geq 0, \ \forall i, j$$

We can further formulate this into a dual problem. For $\phi = (\phi_1, ..., \phi_n)$, $\psi = (\psi_1, ..., \psi_n)$, and $\xi = (\xi_{ij})_{ij}$, consider the following Lagrangian dual,

$$\mathcal{L}(\Pi, \phi, \psi, \xi) = \sum_{ij} c_{ij}\Pi_{ij} + \sum_i \phi_i(\mu_i - \sum_j \Pi_{ij}) + \sum_j \psi_j(\nu_j - \sum_i \Pi_{ij}) - \sum_{ij} \xi_{ij}\Pi_{ij} \quad (1)$$

$$= \sum_i \phi_i\mu_i + \sum_j \psi_j\nu_j + \sum_{ij}(c_{ij} - \mu_i - \nu_j - \xi_{ij})\Pi_{ij} \quad (2)$$

Thus,

$$\min_{\Pi} \mathcal{L}(\Pi; \phi, \psi, \xi) = \begin{cases} -\infty, & c_{ij} \neq \phi_i + \psi_j + \xi_{ij} \text{ for some } i, j. \\ \sum_i \phi_i\mu_i + \sum_j \psi_j\nu_j, & c_{ij} = \phi_i + \psi_j + \xi_{ij} \text{ for all } i, j. \end{cases}$$

Then we can write the dual problem as

$$\max_{\phi,\psi} \left\{ \sum_i \phi\mu_i + \sum_j \psi_j\nu_j \right\} \text{ such that } c_{ij} = \phi_i + \psi_j + \xi_{ij} \text{ for all } i, j,\ \xi_{ij} \geq 0 \text{ for all } i, j.$$

This is equivalent to

$$\max_{\phi,\psi} \left\{ \sum_i \phi\mu_i + \sum_j \psi_j\nu_j \right\} \text{ such that } c_{ij} \geq \phi_i + \psi_j \text{ for all } i, j.$$

For general cases, the dual formulation of $\min_{\Pi \in \Gamma(\mu,\nu)} \int c(x, y)\ d\Pi(x, y)$ is

$$\sup_{\phi,\psi} \int \phi(x)\ d\mu(x) + \int \psi(y)\ d\nu(y) \text{ such that } \phi(x) + \psi(y) \leq c(x, y)\ \forall x \in \mathcal{X}, y \in \mathcal{Y}.$$

1.7 Some quick questions related to the content of Lecture 1:

1. True or False? Suppose that $\mathcal{X} = \{x_0\}$ and let $\mathcal{Y}$ be arbitrary. Then $\mathcal{P}(\mathcal{X})$ has only one element. What about $\Gamma(\mu, \nu)$ for $\mu \in \mathcal{P}(\mathcal{X})$ and $\nu \in \mathcal{P}(\mathcal{Y})$?

2. True or False? If π_1 and π_2 are solutions to the Kantorovich problem:

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

then $\pi^* := \theta\pi_1 + (1 - \theta)\pi_2$, where $\theta \in [0, 1]$, is also a solution.

3. If $\mathcal{X} = \mathcal{Y} = \mathbb{R}$, and $\mu, \nu \in \mathcal{P}(X)$ are such that μ has a density w.r.t. Lebesgue measure, is it true that there always exists a transport **map** between μ and ν? If so, construct one such map. **Hint:** Use the composition of a suitable c.d.f. and a suitable quantile function. Why could your construction fail if μ is not absolutely continuous w.r.t. Lebesgue measure?

4. (Transport maps induce transport plans) Let $T : \mathcal{X} \to \mathcal{Y}$ be a transport **map** between μ and ν. This means that $T_\sharp\mu = \nu$, using the notation that we introduced in class. Define the map

$$Id \times T : \mathcal{X} \to \mathcal{X} \times \mathcal{Y}$$

given by

$$(Id \times T)(x) = (x, T(x)).$$

Show that $\pi_T := (Id \times T)_\sharp\mu$ is a coupling/transportation **plan** between μ and ν. Moreover, show that

$$\int c(x, T(x))d\mu(x) = \int c(x, y)d\pi_T(x, y).$$

Note: From the above one can deduce the inequality $(Kantorovich) \leq (Monge)$ that we mentioned in class.

5. Review the theory of Lagrangian duality for convex optimization problems (e.g., check the wiki page). Check also what are the KKT conditions.

2 Lecture 2: 09/14/2023

Scribes: Xinran Miao and Tinghui Xu

2.1 Recap

- OT Problems
  - $\inf_{T:T\sharp\mu=\nu} \int c(x,T(x))\mathrm{d}\mu(x)$
  - $\inf_{\pi\in\Gamma(\mu,\nu)} \int c(x,y)\mathrm{d}\pi(x,y)$, where $\pi \in \mathcal{P}$ is a proability measure.
- Probability Expression

$$\min_{\pi\in\Gamma(\mu,\nu)} \mathbb{E}_{(x,y)\in\pi}[c(x,y)]$$

  where $x \sim \mu$ and $y \sim \nu$.
- Duality

$$\sup_{(\phi,\psi),\phi:x\to\mathbb{R},\psi:y\to\mathbb{R}} \int \phi(x)\mathrm{d}\mu(x) + \int \psi(y)\mathrm{d}\nu(y)$$
$$\phi(x)+\psi(y) \le c(x,y)$$
$$\mu \text{ a.e. } x, \nu \text{ a.e. } y.$$

2.2 Wasserstein distance

Let $\mathcal{X} = \mathcal{Y} = \mathbb{R}^d$, $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$. Suppose we wish to compare $N(0,\Sigma)$ and $\mu_n = \frac{1}{n}\sum_{i=1}^n \delta_{x_i} \in \mathcal{P}(\mathbb{R}^d)$ where $x_1, \cdots, x_n \in \mathbb{R}^d$ and

$$\mathcal{P}_p(\mathbb{R}^d) \quad = \quad \left\{\mu \in \mathcal{P}(\mathcal{X}) : \int |x|^p \mathrm{d}\mu(x) < \infty\right\}.$$

Definition 2.2.1. *Define p-Wasserstein distance between μ and ν as*

$$\{W_p(\mu,\nu)^p\}^p \quad = \quad \min_{\pi\in\Gamma(\mu,\nu)} \int |x-y|^p \mathrm{d}\pi(x,y).$$

Remark 2.2.1. *$W_p(\mu,\mu_n) < \infty$ if $\mu \in \mathcal{P}_p(\mathbb{R}^d)$ and μ_n is some empirical measure.*

Theorem 2.2.1. *W_p is a distance function over $\mathcal{P}_p(\mathbb{R}^d)$.*
(Question: How can we compare two different probability measures over $\mathbb{R}^d$?)

(i) $W_p(\mu,\nu) = 0$ *iff* $\mu = \nu$.

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

(iii) $W_p(\mu_1,\mu_3) \le W_p(\mu_1,\mu_2) + W_p(\mu_2,\mu_3)$.

(iv) $\mu^k \to_{W_p} \mu$ *(*$\lim_{k\to\infty} W_p(\mu^k,\mu) = 0$*) is equivalent to*

1. $\mu^k \to \mu$ *weakly, and*

2. p-moment of μ^k converges to p-moment of μ.

proof of (iii):

Suppose $\pi_{12}^* \in \Gamma(\mu_1, \mu_2)$ is optimal for $\min_{\pi\in\Gamma(\mu_1,\mu_2)} \int |x-y|^p \mathrm{d}\pi(x,y)$ and $\pi_{23}^* \in \Gamma(\mu_2, \mu_3)$ is optimal for $\min_{\pi\in\Gamma(\mu_2,\mu_3)} \int |x-y|^p \mathrm{d}\pi(x,y)$. We have

$$\left\{\int |x_1 - x_2|^p \mathrm{d}\pi_{12}^*(x_1, x_2)\right\}^{1/p} + \left\{\int |x_2 - x_3|^p \mathrm{d}\pi_{23}^*(x_2, x_3)\right\}^{1/p} = W_p(\mu_1, \mu_2) + W_p(\mu_2, \mu_3) \quad (3)$$

We wish to find $\pi_{13} \in \Gamma(\mu_1, \mu_3)$ such that

$$\left\{\int |x_1 - x_3|^p \mathrm{d}\pi_{13}(x_1, x_3)\right\}^{1/p} \leq (3).$$

Next, we will construct $\gamma \in \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^d)$ such that $\gamma_{12} = \pi_{12}^*$, $\gamma_{23} = \pi_{23}^*$ and $\pi_{13} = \gamma_{13}$.

$$\left\{\gamma(A \times B \times C) := \int_B \left(\int_A \int_C \mathrm{d}\pi_{1|2}^*(x_1 \mid x_2)\mathrm{d}\pi_{3|2}^*(x_3 \mid x_2)\right) \mathrm{d}\mu(x_2)\right\}$$

Let $x_2 \sim \mu_2$, $x_1 \sim \pi_{1|2}^*(\cdot \mid x_2) \perp x_3 \sim \pi_{3|2}^*(\cdot \mid x_2)$.

$$\begin{aligned}
\left(\int |x_1 - x_3|^d \mathrm{d}\pi_{1|3}\right)^{1/p} &= \left(\int |x_1 - x_3|^p \mathrm{d}\gamma(x_1, x_2, x_3)\right)^{1/p} \\
&\leq \left\{\int (|x_1 - x_2| + |x_2 - x_3|)^p \mathrm{d}\gamma(x_1, x_2, x_3)\right\}^{1/p} \\
&\leq \left\{\int |x_1 - x_2|^p \mathrm{d}\gamma(x_1, x_2, x_3)\right\}^{1/p} + \left\{\int |x_2 - x_3|^p \mathrm{d}\gamma(x_1, x_2, x_3)\right\}^{1/p} \\
&= \left\{\int |x_1 - x_2|^p \mathrm{d}\pi_{12}^*(x_1, x_2)\right\}^{1/p} + \left\{\int |x_2 - x_3|^p \mathrm{d}\pi_{23}^*(x_2, x_3)\right\}^{1/p}
\end{aligned}$$

□

2.3 Domain Adaption

Question: Define two functions $f_n : \{x_1, ..., x_n\} \to \mathbb{R}$ and $f : \mathbb{R}^d \to \mathbb{R}$, where f is in the L^p space, i.e., $f \in L^p(\mu)$. Note that these two functions might not be smooth. How can we define a distance between them?

To solve it, we could assume that $\{x_i\}_n$ have a discrete distribution $\mu_n = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$, based on which we could define the following distance.

$$\left(d\left((\mu, f), (\mu_n, f_n)\right)\right)^p := \inf_{\pi\in\Gamma(\mu,\mu_n)} \int |x-y|^p d\pi(x,y) + \int |f(x) - f_n(y)|^p d\pi(x,y)$$

The first half of this distance could be treated as the **penality on y-axis** while the second half could be treated as the **penality on x-axis**. It is similar to a Wasserstein distance with one more dimension. Also, we can combine these two integrals together, in which case $|x-y|^p + |f(x) - f_n(y)|^p$ is regarded as a huge cost function.

This distance is proved to be well-defined and widely used in domain adaption. The basic problem in domain adaption is to find a function $g \in L^p(\nu)$ such that it minimizes $d\left((\mu, f), (\nu, g)\right)$ where μ, ν, and $f \in L^p(\mu)$ are given.

2.4 Relationship between Monge Problem and Kantorovich Problem

We already know that the minimal cost of the Kantorovich Problem is smaller than that of the Monge Problem, but how can we prove it?

We could convert the Monge Problem in the form of couplings. Assume an arbitrary transport map T such that $T_{\#}\mu = \nu$ and a corresponding coupling $\pi_T \in \Gamma(\mu, \nu)$. Define a function $(Id \times T) : x \to \mathcal{X} \times \mathcal{Y}$, which maps x to $(x, T(x))$. Hence, the Monge Problem could be rewritten as follows.

$$\int c(x, T(x))\, d\mu(x) = \int c(x, y)\, d\pi_T(x, y) \geq (Kantorovich)$$

Note that if the optimal transport map of the Monge Problem happens to be that the optimal solution to the Kantorovich Problem, such solution is a very sparse one. Even if no transport map is optimal, we can still guarantee the sparsity of the solution to the Kantorovich Problem by the following theorem.

Theorem 2.4.1. *(Sparsity of solutions to the Kantorovich Problem in the discrete setting) Let $\mathcal{X} = \{x_1, \ldots, x_n\}$ and $\mathcal{Y} = \{y_1, \ldots, y_m\}$. For every cost function $C : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ and for every $\mu \in \rho(\mathcal{X})$, $\nu \in \rho(\mathcal{Y})$, there exists $\pi^* \in \Gamma(\mu, \nu)$ which is a solution to the Kantorovich Problem and π^*has at least $m + n - 1$ non-zero entries.*

2.5 Brenier Theorem

Theorem 2.5.1. *(Brenier Theorem) Define spaces $\mathcal{X} = \mathcal{Y} = \mathbb{R}^d$ and measures $\mu, \nu \in \rho_2(\mathbb{R}^d)$. For computational convenience, define the cost function as $c(x, y) = \frac{1}{2}\|x - y\|^2$. Suppose μ has a density with respect to the Lebesgue measure. Then, the following results hold.*

1. *The Kantorovich Problem admits a **unique** solution π^*.*
2. *In fact, $\pi^* = (Id \times T^*)_{\#}\mu$ for some $T^* : \mathbb{R}^d \to \mathbb{R}^d$ and $T^*\mu = \nu$.*
3. *In particular, the Kantorovich Problem has the **same solution** as the Monge Problem.*
4. *Furthermore, T^* has the form $\nabla\varphi$, where $\varphi : \mathbb{R}^d \to \mathbb{R}$ is a **convex** funtion.*

Conversely, if φ is a convex function $\mathbb{R}^d \to \mathbb{R}$ such that $\nabla\varphi_{\#}\mu = \nu$, then $\nabla\varphi$ is an optimal transport map.

Example Let $\mu = N(m_1, \Sigma_1)$ and $\nu = N(m_2, \Sigma_2)$ be two normal distributions. According to the Brenier Theorem, the solution to the Kantorovich Problem and the Monge Problem must be the same and unique. We are done if we can find a convex function φ such that $\nabla\varphi_{\#}\mu = \nu$.

Define a linear transport map T.

$$T(x) = A(x - m_1) + m_2 = \nabla\varphi, \quad \text{where } \varphi \text{ is convex}$$

We need to find a symmetric matrix A such that if $Z_1 \sim N(m_1, \Sigma_1)$, then $Z_2 := T(Z_1) \sim N(m_2, \Sigma_2)$. Thus, the following two constraints for A should be satisfied.

$$\begin{cases} A^T\Sigma A = \Sigma_2 \\ A \text{ is positively semi-definite} \end{cases}$$

Many matrices satisfy $A^T\Sigma A = \Sigma_2$, but the hardest part is that such A should be positive semi-definite simultaneously. It could be proved that the unique solution to this example is

$$A = \Sigma_1^{-\frac{1}{2}} \left(\Sigma_1^{\frac{1}{2}}\Sigma_2\Sigma_1^{\frac{1}{2}}\right)^{\frac{1}{2}} \Sigma_1^{-\frac{1}{2}}.$$

Furthermore, the solution to this specific problem of normal distributions exists in a space called Bure-Wasserstein Space.

2.6 Some thoughts after Lecture 2

1. We can use two different arguments to prove Theorem 2.4.1. Notice that the conclusion of the theorem doesn't depend on the specific cost function c! As you'll see, this is simply because this theorem is really about a property of the set of extremal points of $\Gamma(\mu, \nu)$ in the discrete case.

   - Approach 1: Let π^* be an optimal coupling. We can represent π^* as a (weighted) bipartite graph over the node sets $\mathcal{X}$ and $\mathcal{Y}$ as follows: we put an edge between x_i and y_j if $\pi^*(x_i, y_j) > 0$. We will show that, if this graph has a cycle, then we can construct a new solution to the OT problem that is strictly sparser than π^*. To see this, for simplicity assume that there is a cycle consisting of nodes x_1, y_1, x_2, y_2 (this case already illustrates the main idea). This means that:

   $$\pi^*(x_1, y_1), \pi^*(x_2, y_1), \pi^*(x_2, y_2), \pi^*(x_1, y_2) > 0$$

   Let ε be the smallest of these quantities. We define a new π as follows

   $$\pi(x_1, y_1) := \pi^*(x_1, y_1) + \varepsilon, \pi(x_2, y_1) := \pi^*(x_2, y_1) - \varepsilon,$$

   $$\pi(x_2, y_2) := \pi^*(x_2, y_2) + \varepsilon, \pi(x_1, y_2) := \pi^*(x_1, y_2) - \varepsilon,$$

   and for all other pairs (x_i, y_j) we set $\pi(x_i, y_j)$ to be equal to $\pi^*(x_i, y_j)$. One can show that: 1) π is still a coupling between the original μ and ν; 2) π has the same objective value as π^*, and thus π is also a solution of the OT problem; 3) the graph associated to the new π has at least one fewer edge than π^*'s.

   Continuing in this way, we can produce solutions π to the OT problem that are sparser and sparser. The process stops the moment the graph associated to the current solution has no cycles, i.e., when it is a tree. Recalling that trees can have at most $N - 1$ edges ($N = n + m$ number of nodes), we obtain the desired result: we can always find solutions to the OT problem that have at most $m + n - 1$ non-zero entries.

   - Approach 2: we use the following result from convex geometry:

   Theorem 2.6.1 (Dubins theorem)**.** *Let K be a convex and compact subset of $\mathbb{R}^d$. Let $H_1, \ldots, H_s$ be a collection of hyperplanes in $\mathbb{R}^d$. Let*

   $$C := K \cap H_1 \cap \cdots \cap H_s.$$

   Then every extremal point of C can be represented as the convex combination of at most $s + 1$ extremal points of K.

   We use this theorem as follows:
   Take

   $$K := \{\pi \in \mathbb{R}^{n \times m} \text{ s.t. } \pi_{ij} \geq 0, \forall i, j, \text{ and } \sum_{ij} \pi_{ij} = 1\}$$

   and

   $$H_i = \{\pi \text{ s.t. } \sum_j \pi_{ij} = \mu_i\}, \quad i = 1, \ldots, n - 1,$$

   $$H^j = \{\pi \text{ s.t. } \sum_i \pi_{ij} = \nu_i\}, \quad j = 1, \ldots, m - 1,$$

   and let C be the intersection of K with all the above hyperplanes. The resulting set C is precisely $\Gamma(\mu, \nu)$. From Dubins Theorem we see that every extremal point of $\Gamma(\mu, \nu)$ can be written as the convex combination of at most $m - 1 + n - 1 + 1 = m + n - 1$ extremal points of K. However, the extremal points of K are precisely those π that are zero everywhere except at one entry. Since the OT problem is a linear optimization problem, there always exist solutions that are extremal points of $\Gamma(\mu, \nu)$. Therefore, there always exist solutions that have at most $m + n - 1$ entries.

Note 1: a proof of Dubins theorem in case K is a convex polytope can be found in the paper by Friesecke and Penka that I added to our dropbox folder. That paper proposes a numerical method for solving OT problems that is constrained to seek sparse solutions to the OT problem (contrast to Sinkhorn algorithm, which we will discuss in Lecture 3). Unfortunately, the type of theoretical guarantees that can be proved for their method is a bit weak, but I still think their paper has some interesting ideas and can be useful in practice.
Note 2: Approach 2 generalizes easily to the case of multimarginal OT, whereas Approach 1 only works for the (2 marginals) OT problem that we have studied so far. We will discuss multimarginal OT (MOT) in Lecture 4.

2. Using our discussion from the first two lectures, convince yourself that the following result is true:
   Let $\mu, \nu \in \mathcal{P}_2(\mathbb{R})$ and suppose that μ has a density w.r.t. Lebesgue measure. Let F_μ be μ's cdf and let F_ν^{-1} be ν's quantile function. Then
   $$T := F_\nu^{-1} \circ F_\mu$$
   is the unique solution to the Monge OT problem:
   $$\min_{T:\, T_\sharp \mu = \nu} \int |x - T(x)|^2 d\mu(x).$$

3 Lecture 3: 09/21/2023

Scribes: Yewei Xu and Jun Chang

3.1 Dual of Kantorovich problem revisited, C-transform, and c-concavity

In this lecture, we made extensive use of the idea of going back and forth between the primal, which we recall to be the Kantorovich problem (K) defined as

$$\min_{\pi \in \Gamma(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{Y}} c(x,y) d\pi(x,y)$$

and the dual problem (D) defined as

$$\sup_{\phi \in L^1(\mu), \psi \in L^1(\nu)} \int_{\mathcal{X}} \phi d\mu + \int_{\mathcal{Y}} \psi d\nu \;\text{ such that }\; \phi(x) + \psi(y) \le c(x,y).$$

Before proceeding any further, let's reinvestigate the dual problem (D), and try to see how we can simplify the formulation of this problem. The idea is we try to see how I can improve if I already have something feasible, or more precisely, given one of the dual variables, how I can pick the other one in the best possible way.

Definition 3.1.1. *Give ϕ, the c-transform of ϕ is defined as*

$$\phi^c(y) := \inf_{\tilde{x}} c(\tilde{x}, y) - \phi(\tilde{x}).$$

Give ψ, the c-transform of ψ is defined as

$$\psi^c(x) := \inf_{\tilde{y}} c(x, \tilde{y}) - \psi(\tilde{y}).$$

There are still a few things to do before we can really utilize this definition. We first need to check that c-transforms are well-defined, that is, (ϕ, ϕ^c) is always a feasible solution for the dual problem. This follows easily from

$$\begin{aligned}
\phi(x) + \phi^c(y) &= \phi(x) + \inf_{\tilde{x}} (c(\tilde{x}, y) - \phi(\tilde{x})) \\
&\le \phi(x) + (c(x,y) - \phi(x)) \\
&= c(x,y)\ ,
\end{aligned}$$

and a similar feasibility can be derived for $(\psi^c(x), \psi(y))$.

Another thing that we want to confirm is the advantage of the c-transform. We claim that c-transforms always give us optimal solutions to the dual problem, that is, suppose (ϕ, ψ) is feasible, then

$$\int \phi d\mu + \int \psi d\nu \leq \int \phi d\mu + \int \phi^c d\nu.$$

This follows from the fact that for any $\tilde{x}$,

$$\psi(y) \leq c(\tilde{x}, y) - \phi(\tilde{x}),$$

so by taking the infimum over $\tilde{x}$, we always have

$$\psi(y) \leq \phi^c(y).$$

Remark 3.1.1. *Using the discussion on c-transforms above, we are able to rewrite the dual problem as*

$$\sup_{\phi \in L^1(\mu)} \int \phi d\mu + \int \phi^c d\nu.$$

In this formulation, the dual problem now has only one variable to be optimized.

Finally, we define the of c-concavity and recall the notion of Fenchel duality.

Definition 3.1.2. *We say that ϕ is a c-concave function if ϕ can be written as $\phi = (\phi^c)^c$. That is, applying c-transform twice on itself.*

Definition 3.1.3. *Given a function φ, the Fenchel duality of φ is defined*

$$\varphi^*(y) := \sup_{\tilde{x}} \big\{ \langle \tilde{x}, y \rangle - \varphi(\tilde{x}) \big\}.$$

Remark 3.1.2. *The common notion of convexity, in terms of Fenchel duality, can be defined as $\varphi = \varphi^{**}$.*

3.2 Characterization of optimal solutions of the Kantorovich problem

The first character of the optimal solutions of the Kantorivich problem is described as follows.

Proposition 3.2.1. *If π is a solution to the Kantorovich problem, then*

$$\mathrm{Supp}(\pi) \subseteq \big\{ (x, y) \ : \ \phi(x) + \phi^c(y) = c(x, y) \big\}$$

for some ϕ that is c-concave. This ϕ would have to be a solution to the dual problem.

Here $\mathrm{Supp}(\pi)$ denotes the support of π, that is, the set on $\mathcal{X} \times \mathcal{Y}$ such that all its open neighbourhood has a positive π measure.

Remark 3.2.1. *The above proposition is essentially the complimentary slackness condition in the KKT (Karush–Kuhn–Tucker) conditions.*

We are not going to prove the above proposition, but we would prove its converse, which is stated below.

Proposition 3.2.2. *If there is a ϕ such that*

$$\mathrm{Supp}(\pi) \subseteq \big\{ (x, y) \ : \ \phi(x) + \phi^c(y) = c(x, y) \big\},$$

then π is a solution to the Kantorovich problem.

Proof. For our convenience, denote in this proof the solution to the Kantorovich problem (K), and that to its dual problem (D). The proof is composed of two parts. We first prove that (K) $\leq$ (D). Note that

$$\begin{aligned}
(\mathrm{K}) &\leq \int_{\mathcal{X}\times\mathcal{Y}} c(x,y)d\pi(x,y) \\
&= \int_{\mathrm{Supp}(\pi)} c(x,y)d\pi(x,y) \\
&= \int_{\mathrm{Supp}(\pi)} (\phi(x)+\phi^c(y))d\pi(x,y) \\
&= \int_{\mathcal{X}\times\mathcal{Y}} (\phi(x)+\phi^c(y))d\pi(x,y) \\
&= \int_{\mathcal{X}\times\mathcal{Y}} \phi(x)d\pi(x,y) + \int_{\mathcal{X}\times\mathcal{Y}} \phi^c(y)d\pi(x,y) \\
&= \int_{\mathcal{X}} \phi(x)d\mu(x) + \int_{\mathcal{Y}} \phi^c(y)d\nu(y) \\
&\leq (\mathrm{D})\ .
\end{aligned}$$

In the second and fourth line above, we used the property that π vanishes outside $\mathrm{Supp}(\pi)$. In the third line above, we used the assumption on $\mathrm{Supp}(\pi)$. In the fifth line, we used the fact that since $\pi \in \Gamma(\mu,\nu)$, its marginal distribution on $\mathcal{X}$ and $\mathcal{Y}$ would be μ and ν respectively.

We then prove the well-known weak-duality property, that is, (D) $\leq$ (K), using a similar strategy: Let $\pi \in \Gamma(\mu,\nu)$ be arbitrary, and (ϕ,ψ) be feasible. Then

$$\int_{\mathcal{X}} \phi(x)d\mu(x) + \int_{\mathcal{Y}} \psi(y)d\nu(y) = \int_{\mathcal{X}\times\mathcal{Y}} (\phi(x)+\psi(y))d\pi(x,y) \leq \int_{\mathcal{X}\times\mathcal{Y}} c(x,y)d\pi(x,y).$$

By taking the supremum on the left hand side and the infimum on the right hand side, we obtain (D) $\leq$ (K). □

3.3 Relationship between convexity, c-transform, and Fenchel duality

In this section, we fix $c(x,y) = \frac{1}{2}|x-y|^2$, and investigate the relationship between convexity, c-transform and Fenchel duality.

For any given ϕ, define

$$\varphi(x) := \frac{1}{2}|x|^2 - \phi(x). \tag{4}$$

Then its Fenchel duality is given by

$$\begin{aligned}
\varphi^*(y) &= \sup_{\tilde{x}} \big\{\langle \tilde{x}, y\rangle - \varphi(\tilde{x})\big\} \\
&= \sup_{\tilde{x}} \big\{\langle \tilde{x}, y\rangle - \frac{1}{2}|\tilde{x}|^2 + \phi(\tilde{x})\big\} \\
&= \frac{1}{2}|y|^2 + \sup_{\tilde{x}} \big\{ -\frac{1}{2}|t|^2 + \langle \tilde{x}, y\rangle - \frac{1}{2}|\tilde{x}|^2 + \phi(\tilde{x})\big\} \\
&= \frac{1}{2}|y|^2 + \sup_{\tilde{x}} \big\{ -c(\tilde{x}, y) + \phi(\tilde{x})\big\} \\
&= \frac{1}{2}|y|^2 - \phi^c(y)\ .
\end{aligned}$$

Its Fenchel double dual would then be

$$\varphi^{**}(x) = \frac{1}{2}|x|^2 - \phi^{cc}(y).$$

Remark 3.3.1. *The computation above implies that $\varphi = \varphi^{**}$ is equivalent to $\phi = (\phi^c)^c$, which is just the definition of ϕ being a c-concave function.*

3.4 Proof of Brenier's theorem

Recall the setting in theorem 2.5.1. Let π^* be a solution of the Kantorovich problem with cost function $c(x,y) = \|x-y\|^2/2$. Then, by proposition 3.2.1, there exists a c-concave function ϕ such that $\mathrm{Supp}(\pi^*) \subseteq \{(x,y) : \phi(x) + \phi^c(y) = c(x,y)\}$. Suppose $(x,y) \in \mathrm{Supp}(\pi^*)$. The goal is to find a transport map T^* such that $y = T^*(x)$. Immediately, we know that

$$\phi(x) + \phi^c(y) = c(x,y) \iff c(x,y) - \phi(x) = \phi^c(y) := \inf_{\tilde{x}} c(\tilde{x},y) - \phi(\tilde{x}) \implies x \in \arg\min_{\tilde{x}} c(\tilde{x},y) - \phi(\tilde{x}).$$

Since ϕ is differentiable λ^d-a.e. (see remark 3.4.1), first order conditions for optimality tell us that μ-almost everywhere,

$$0 = \nabla_{\tilde{x}} c(x,y) - \nabla_{\tilde{x}}\phi(x) = (x-y) - \nabla\phi(x) \implies y = x - \nabla\phi(x) = x - \nabla\left(\frac{\|x\|^2}{2} - \varphi(x)\right) = \nabla\varphi(x),$$

where $\varphi(x)$ is the convex function defined in (4). Namely, $\mathrm{Supp}(\pi^*) = \{(x,y) : y = T^*(x) := \nabla\varphi(x)\}$.

Remark 3.4.1. *We require that ϕ is differentiable λ^d-a.e. for this proof to work. This is why we need the condition that μ has density with respect to Lebesgue measure on $\mathbb{R}^d$ ($\mu \ll \lambda^d$). Observing*

$$\phi(x) = \|x\|^2/2 - \varphi(x),$$

the first term on the right is differentiable everywhere, and the second term is differentiable μ-a.e because φ is convex. Since $\mu \ll \lambda^d$, we have that ϕ is differentiable λ^d-a.e.

Remark 3.4.2. *The uniqueness of π^* is implied in the proof above. Suppose that π_1 and π_2 are solutions of (K), and let $T_1 := \nabla\varphi_1$ and $T_2 := \nabla\varphi_2$ be the corresponding optimal transport maps, and let $\phi_1(x) = \|x\|^2/2 - \varphi_1(x)$ and $\phi_2(x) = \|x\|^2/2 - \varphi_2(x)$ be the corresponding dual potentials. Proposition 3.2.1 tells us*

$$\phi_1(x) + \phi_1^c(y) = c(x,y) = \phi_2(x) + \phi_2^c(y).$$

Taking gradients with respect to x, which is valid due to remark 3.4.1, we have μ-a.e.,

$$x - \nabla_x\varphi_1(x) = x - y = x - \nabla_x\varphi_2(x) \implies \nabla\varphi_1(x) = \nabla\varphi_2(x).$$

We conclude that $T^* = \nabla\varphi(x)$ is unique μ-a.e., hence $\pi^* = (\mathrm{Id} \times T^*)_\sharp\mu$ is unique μ-a.e.

3.5 Entropy-regularized Optimal Transport

The Kantorovich problem requires us to solve a convex optimization problem that does not necessarily have a unique solution. It turns out that regularizing the original problem in a certain way results in a problem that has a unique solution, and allows us to use a fast alternative solver known as the Sinkhorn-Knopp algorithm.

Definition 3.5.1. *The entropy-regularized optimal transport problem is defined as*

$$\min_{\pi\in\Gamma(\mu,\nu)} \int c(x,y)\,\mathrm{d}\pi(x,y) + \varepsilon\,\mathrm{KL}(\pi\|\mu\otimes\nu), \tag{5}$$

where $\mu\otimes\nu$ is the product measure on $\mathcal{X}\times\mathcal{Y}$, and $\mathrm{KL}(\pi\|\mu\otimes\nu)$ is the Kullback–Leibler divergence between the measures π and $\mu\otimes\nu$ defined by

$$\mathrm{KL}(\pi\|\mu\otimes\nu) = \begin{cases} \int \log\left(\mathrm{d}\left(\frac{\pi}{\mu\otimes\nu}\right)\right)\mathrm{d}\pi & \text{if } \pi \ll \mu\otimes\nu, \\ \infty & \text{otherwise.} \end{cases}$$

Remark 3.5.1. *The regularization term in (5) should be small so that its optimal value is close to that of the original problem. Considering that* $\mathrm{KL}(\pi\|\mu\otimes\nu)=0 \iff \pi=\mu\otimes\nu$*, intuitively* π *will be "pushed" closer to* $\mu\otimes\nu$ *as* ε *grows, whereas* π *does not have to be close to* $\mu\otimes\nu$ *when* ε *is small.*

Proposition 3.5.1. *For* $\varepsilon>0$*, the solution* π_ε^* *of (5) is unique, and is not sparse.*

Remark 3.5.2. *The "not sparse" part is due to the intuition in remark 3.5.1 that a larger* $\varepsilon>0$ *pushes* π_ε^* *to be closer to the product measure* $\mu\otimes\nu$*, which is not sparse.*

3.6 Solving the entropy-regularized OT problem in the discrete case

We return to the discrete setting where $\mathcal{X}=\{x_1,\dots,x_n\}$ and $\mathcal{Y}=\{y_1,\dots,y_m\}$. In this setting, the entropy-regularized optimal transport problem can be re-written as

$$\min_{\pi\in\Gamma(\mu,\nu)}\sum_{i,j}c_{ij}\pi_{ij}+\varepsilon\sum_{i,j}\log\left(\frac{\pi_{ij}}{\mu_i\nu_j}\right)\pi_{ij}. \tag{6}$$

The regularization term can be further simplified as

$$\begin{aligned}\sum_{i,j}\log\left(\frac{\pi_{ij}}{\mu_i\nu_j}\right)\pi_{ij}=\sum_{i,j}\{\log(\pi_{ij})-\log(\mu_i)-\log(\nu_j)\}\pi_{ij}&=\sum_{i,j}\log(\pi_{ij})\pi_{ij}-\sum_{i,j}\log(\mu_i)\pi_{ij}-\sum_{i,j}\log(\nu_j)\pi_{ij}\\&=\sum_{i,j}\log(\pi_{ij})\pi_{ij}-\sum_{i}\log(\mu_i)\mu_i-\sum_{j}\log(\nu_j)\nu_j,\end{aligned}$$

where we have used the fact that

$$\sum_{i,j}\log(\nu_j)\pi_{ij}=\sum_j\log(\nu_j)\sum_i\pi_{ij}=\sum_j\log(\nu_j)\nu_j.$$

Disregarding the terms that do not depend on π, an equivalent simplified formation of (6) is

$$\min_{\pi\in\Gamma(\mu,\nu)}\sum_{i,j}c_{ij}\pi_{ij}+\varepsilon\sum_{i,j}\{\log(\pi_{ij})-1\}\pi_{ij}. \tag{7}$$

The Lagrangian of (7) is

$$\mathcal{L}(\pi,\phi,\psi)=\sum_{i,j}c_{ij}\pi_{ij}+\varepsilon\sum_{i,j}\{\log(\pi_{ij})-1\}\pi_{ij}+\sum_i\phi_i(\mu_i-\sum_j\pi_{ij})+\sum_j\psi_j(\nu_j-\sum_i\pi_{ij}).$$

Note that there is no Lagrange multiplier corresponding to the non-negative constraints (compare with (1): the Lagrangian for the Kantorivich problem), since any non-positive π_{ij} will make the objective function undefined due to the logarithm[1]. Now, fixing (ϕ,ψ), the dual objective becomes

$$G(\phi,\psi)=\min_{\pi\in\mathbb{R}^{n\times m}}\mathcal{L}(\pi,\phi,\psi).$$

Accordingly, the first order conditions require

$$0=c_{ij}+\varepsilon\log(\pi_{ij})-\phi_i-\psi_j \implies \pi_{ij}=\exp\left(-\frac{1}{\varepsilon}(c_{ij}-\phi_i-\psi_j)\right) \tag{8}$$

$$=\exp\left(\frac{\phi_i}{\varepsilon}\right)\exp\left(\frac{-c_{ij}}{\varepsilon}\right)\exp\left(\frac{\psi_i}{\varepsilon}\right). \tag{9}$$

[1]Furthermore, since there are no inequality constraints, there is no complementary slackness.

In other words, the unique solution π_ε^* of (7) takes the form of (9) for some ϕ and ψ. Moreover, a π of the form (9) is a solution to (7) if and only if

$$\forall j,\ \sum_i \exp\left(\frac{\phi_i}{\varepsilon}\right)\exp\left(\frac{-c_{ij}}{\varepsilon}\right)\exp\left(\frac{\psi_i}{\varepsilon}\right) = \nu_j, \quad \text{and} \quad \forall i,\ \sum_i \exp\left(\frac{\phi_i}{\varepsilon}\right)\exp\left(\frac{-c_{ij}}{\varepsilon}\right)\exp\left(\frac{\psi_i}{\varepsilon}\right) = \mu_i. \quad (10)$$

Now, the question boils down to finding ϕ and ψ that satisfy (10). It turns out that the Sinkhorn-Knopp algorithm can be used for this purpose. More on this topic to follow.

3.7 Some thoughts after Lecture 3

1. When $c(x,y) = d(x,y)$ is a distance function over a space $\mathcal{X}$, the dual of the Kantorovich problem can be written in a very particular form:

$$W_1(\mu,\nu) = \sup_{f\ s.t.\ \mathrm{Lip}(f)\le 1} \int f(x)d\mu(x) - \int f(y)d\nu(y).$$

   This identity is typically referred to as Rubinstein-Kantorovich theorem.

   In the above, Lip denotes the Lipschitz constant of a function relative to d. Try to convince yourself of the following facts:

   - If f is Lipschitz with constant less one, then $\phi = f$ and $\psi = -f$ form a feasible pair for the dual (in the general form) of the OT problem with cost $c = d$.
   - Show that if $f = f^{c\bar{c}}$, i.e., f is c-concave (for $c = d$), then f must be a Lipschitz function with constant less than one.

2. (On semi-discrete optimal transport) In the setting of Brenier theorem, imagine that μ has a density with respect to the Lebesgue measure and that ν is a discrete measure of the form $\nu = \sum_{i=1}^n b_i\delta_{x_i}$. What type of convex function φ has the property that $\nabla\varphi_\sharp\mu = \nu$? Essentially, the gradient of φ needs to be piecewise constant with exactly n different values (the x_i), and thus φ must be a (convex) piecewise linear function of the form $\max_{i=1,\dots,n}\{\langle x_i, x\rangle + c_i\}$. This suggests that the OT problem is finite dimensional.

   In fact, if you write the dual of (K) in this setting, you could write it, using the $\bar{c}$-transform, as

$$\sup_{\psi=(\psi_1,\dots,\psi_n)\in\mathbb{R}^n} \int \psi^{\bar{c}}d\mu + \sum_{i=1}^n b_i\psi_i$$

   The point of doing this is that now the OT problem is written as an optimization problem in $\mathbb{R}^n$ (even though μ has a density), and thus you can use your favorite optimization routine to try to solve it. You may want to check Santambrogio's book "Optimal Transport for Applied Mathematicians" Section 6.4.2.

3. Proposition 3.2.1 is a bit technical, but here are a few concepts that can help you understand what's going on (you can also consider the discrete case and convince yourself that in that setting the characterization follows from KKT optimality conditions).

   - It is actually easier to show first that if π is optimal, then $\mathrm{Supp}(\pi)$ must be a *c-cyclically monotone set*. Here is a definition.

     Definition 3.7.1. *A subset $\mathcal{A}$ of $\mathcal{X}\times\mathcal{Y}$ is c-cyclically monotone if the following condition holds: for any finite collection of points $(x_1,y_1),\dots,(x_n,y_n)$ in $\mathcal{A}$, and for any permutation $\sigma : \{1,\dots,n\}\to\{1,\dots,n\}$, we have*

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

To say that Supp(π) is c-cyclically monotone is to say that there is no way to rearrange π's assignments of sources with targets that produces a cheaper output. This makes sense if π is optimal! I think this is intuitive enough and a rigurous proof is actually not so difficult to work out.

- The most technical part of Proposition 3.2.1 is to show that if $\mathcal{A}$ is a c-cyclically monotone set, then it must be contained in a set of the form $\{(x,y) \ : \ \phi(x)+\phi^c(y)=c(x,y)\}$ for a c-concave function. At least for the case of quadratic cost $c(x,y)=\frac{1}{2}|x-y|^2$, this can be deduced from existing results in convex analysis:
  Step 1: A set $\mathcal{A}$ is $c-$cyclically monotone if and only if it is maximally monotone (check definition in a convex analysis textbook). **Step 2:** the set $\{(x,y) \ : \ \phi(x)+\phi^c(y)=c(x,y)\}$ for some c-concave function ϕ is the same as the subdifferential of the (convex) function $\varphi := |x|^2/2-\phi$. **Step 3:** From a result by Rockafellar, maximally monotone sets are always contained in the subdifferential of a convex function.
  How does one generalize this result to other cost functions? A strategy is to follow the proof structure in Rockafellar and adjust whatever needs to be adjusted.

4 Lecture 4: 09/28/2023

Scribes: Dibyendu Saha and Youngjoo Yun

We continue discussing the entropy-regularized optimal transport problem.

$$\min_{\pi\in\Gamma(\mu,\nu)} \sum_{i,j} c_{ij}\pi_{ij} + \epsilon \sum_{i,j} \left(\log(\pi_{ij}-1)\right)\pi_{ij} \ \text{(E-OT)} \tag{11}$$

Due to the constraints, we need to use the KKT conditions; the Lagrangian formulation of (E-OT) is given below.

$$\mathcal{L}(\pi;\phi,\psi) = \sum_{i,j} c_{ij}\pi_{ij} + \epsilon \sum_{i,j} \left(\log(\pi_{ij}-1)\right)\pi_{ij} + \sum_i \phi_i \left(\mu_i - \sum_j \pi_{ij}\right) + \sum_j \psi_j \left(\nu_j - \sum_i \pi_{ij}\right), \tag{12}$$

where ϕ keeps track of the first marginal constraint and ψ the second marginal constraint. Note that $\log(\cdot)$ in (E-OT) already implicitly enforces the non-negativity constraint. The dual objective is given below.

$$G(\phi,\psi) = \min_{\pi} \mathcal{L}(\pi;\phi,\psi). \tag{13}$$

For given (ϕ,ψ),

$$\pi_{ij} = \underbrace{e^{\frac{\phi_i}{\epsilon}}}_{=:a_i} \underbrace{e^{-\frac{c_{ij}}{\epsilon}}}_{k_{ij}} \underbrace{e^{\frac{\psi_j}{\epsilon}}}_{b_j} \forall i,j. \tag{14}$$

As a corollary, we have the following.

Corollary 4.0.1. *If π has the form in (14) for some ϕ,ψ, then π is optimal for $(E-OT)$ if and only if π satisfies the marginal constraints.*

We next derive an explicit form of $G(\phi,\psi)$.

$$G(\phi,\psi) = \mathcal{L}(\pi;\phi,\psi) \tag{15}$$

$$= \sum_{i,j} c_{ij}a_ik_{ij}b_j + \epsilon \sum_{i,j} \left(\log(a_ik_{ij}b_j)-1\right) a_ik_{ij}b_j + \sum_i \phi_i\mu_i + \sum_j \psi_j\nu_j - \sum_i\sum_j \phi_i a_i k_{ij} b_j - \sum_i\sum_j \psi_i a_i k_{ij} b_j \tag{16}$$

$$= \cdots \tag{17}$$

$$= \sum_i \phi_i\mu_i + \sum_j \psi_j\nu_j - \epsilon \sum_{ij} a_ik_{ij}b_j. \tag{18}$$

Then, we can rewrite the dual objective in (13) as the following.

$$\max_{(\phi,\psi)} \sum_i \phi_i \mu_i + \sum_j \psi_j \nu_j - \epsilon \sum_{i,j} e^{\frac{\phi_i}{\epsilon}} e^{-\frac{c_{ij}}{\epsilon}} e^{\frac{\psi_j}{\epsilon}} \tag{19}$$

Given ϕ and ψ, we have

$$\lim_{\epsilon \to 0} -\epsilon \sum_{i,j} e^{\frac{\phi_i}{\epsilon}} e^{-\frac{c_{ij}}{\epsilon}} e^{\frac{\psi_j}{\epsilon}} = \begin{cases} 0 & \text{if } \phi_i + \psi_j < c_{ij} \forall i, j \\ -\infty & \text{if } \phi_i + \psi_j > c_{ij} \exists i, j \end{cases}. \tag{20}$$

Next, we consider the computational problem of solving the optimization problem in (19).

4.1 Coordinate ascent

Note that fixing ϕ, the problem becomes

$$\operatorname*{argmax}_{\psi} \sum_i \phi_i \mu_i + \sum_j \psi_j \nu_j - \epsilon \psi_j \nu_j - \epsilon \sum_{i,j} e^{-\frac{\phi_i}{\epsilon}} e^{\frac{c_{ij}}{\epsilon}} e^{\frac{\psi_i}{\epsilon}}.$$

Taking derivatives gives

$$0 = \nu_j - \sum_i e^{\frac{\psi_i}{\epsilon}} e^{-\frac{\psi_j}{\epsilon}}.$$

Then,

$$b_j = e^{\frac{\psi_j}{\epsilon}} = \frac{\nu_j}{\sum_i e^{\frac{\phi_i}{\epsilon}} \epsilon^{-\frac{c_{ij}}{\epsilon}}} = \frac{\nu_j}{\sum_i k_{ij} a_i}.$$

Then, similarly, fixing ψ gives

$$a_i = \frac{\mu_i}{\sum_j k_{ij} b_j}. \tag{21}$$

Sinkhorn algorithm, given below, is an algorithm that naturally follows.

- Initialize: (ϕ^0, ψ^0)
- Compute (ϕ^1, ψ^1): update ϕ implicitly by updating $b_j^1 = \frac{v_j}{\sum_i k_{ij} a_i^1}$. Set $a^1 = a^0$.
- Compute (ϕ^2, ψ^2): update ϕ implicitly by updating $a_i^2 = \frac{\mu_i}{\sum_j k_{ij} b_j^2}$. Set $b^2 = b^1$.

A second interpretation of the Sinkhorn algorithm is that in each step of the algorithm, we match one of the two constraints separately.

A question that we would like to ask ourselves is whether the π is feasible or not. Recall that given ϕ and ψ, we have the constraint $\phi_{ij} = e^{\frac{\phi_i}{\epsilon}} e^{-\frac{c_{ij}}{\epsilon}} e^{\frac{\psi}{\epsilon}}$. We would like to have

$$v_j = \sum_i \pi_{ij}, \tag{22}$$

where the right hand side is equal to $e^{\frac{\psi_j}{\epsilon}} \sum_i e^{\frac{\phi_i}{\epsilon}} e^{-\frac{c_{ij}}{\epsilon}}$, i.e. we want

$$\frac{\nu_j}{\sum_i e^{\frac{\phi_i}{\epsilon}} e^{-\frac{c_{ij}}{\epsilon}}} = e^{\frac{\psi_j}{\epsilon}}. \tag{23}$$

4.2 Computational Complexity of Sinkhorn Algorithm

Note that, given (ϕ, ψ), the dual objective function can be written as,

$$G(\phi, \psi) = \sum_i \phi_i \mu_i + \sum_j \psi_j \nu_j - \varepsilon \sum_{i,j} \exp\left(\frac{\phi_i}{\varepsilon}\right) \exp\left(-\frac{c_{ij}}{\varepsilon}\right) \exp\left(\frac{\psi_j}{\varepsilon}\right) \tag{24}$$

From t to $t+1$-th interation step, we want to compare the value of the dual objective function, i.e.; we want to compare the values of $G(\phi^t, \psi^t)$ and $G(\phi^{t+1}, \psi^{t+1})$.

Without loss of generality, let us assume that, going from t to $t+1$-th step, we are updating the argument ϕ. Then the following holds:

$$\psi^t \equiv \psi^{t+1} \tag{25}$$

$$\exp\left(\frac{\phi^{t+1}}{\varepsilon}\right) = \frac{\mu_i}{\sum_j \exp\left(\frac{\psi_j^t}{\varepsilon}\right) \exp\left(-\frac{c_{ij}}{\varepsilon}\right)} \tag{26}$$

$$\sum_i \exp\left(\frac{\phi_i^t}{\varepsilon}\right) \exp\left(-\frac{c_{ij}}{\varepsilon}\right) \exp\left(\frac{\psi_j^t}{\varepsilon}\right) = \nu_j \tag{27}$$

Plugging in these above equations in the expression of dual objective G, one can show that,

$$G(\phi^{t+1}, \psi^{t+1}) - G(\phi^t, \psi^t) = \varepsilon \mathrm{KL}\left(\mu \| \pi_t^1\right) \tag{28}$$

where,

$$\pi_t^1 := \sum_j \pi_{ij}^t = \sum_j \exp\left(\frac{\phi_i^t}{\varepsilon}\right) \exp\left(-\frac{c_{ij}}{\varepsilon}\right) \exp\left(\frac{\psi_j^t}{\varepsilon}\right)$$

- **Initialization:** $(\phi^0, \psi^0) \equiv (\mathbf{0}, \mathbf{0})$
- **Total Energy Gap:**

  Lemma 4.2.1.
  $$G(\phi^*, \psi^*) - G(\phi^1, \psi^1) \leq \log\left(\frac{s}{l}\right)$$
  where, (ϕ^*, ψ^*) is a maximizer of the dual of Entropy-regularised OT and s, l are defined in the following way,
  $$s := \sum_{i,j} \exp\left(-\frac{c_{ij}}{\varepsilon}\right) \leq mn \; [\because c_{ij} \geq 0 \forall i, j]$$
  $$l := \min_{i,j} \exp\left(-\frac{c_{ij}}{\varepsilon}\right) = \exp\left(-\frac{\|C\|_\infty}{\varepsilon}\right)$$

- **Stopping Criterion:** Stop the first time,

$$\|\pi_t^1 - \mu\|_1 + \|\pi_t^2 - \nu\|_1 \leq \delta \tag{29}$$

Note,

$$\mathrm{KL}(\mu \| \pi_t^1) \geq \|\pi_t^1 - \mu\|_1^2 \tag{30}$$

- Let it be the case that the iteration has not stopped until step t^*, then for any $t \leq t^*$, due to (28) and combining (29),(30)

$$G(\phi^{t+1}, \psi^{t+1}) - G(\phi^t, \psi^t \geq \varepsilon\delta^2 \tag{31}$$

$$\implies t^* \leq \frac{\log(s/l)}{\varepsilon\delta^2} \leq \frac{\log nm + \|C\|_\infty/\varepsilon}{\varepsilon\delta^2} \tag{32}$$

If ε is not too small and we have control over $\|C\|_\infty$, then

$$t^* \approx \frac{\log nm}{\delta^2} \tag{33}$$

4.3 Preview of Bodhisattva's talk

Observe that, if $\mu = \nu$ then value of the (...) is not necessarily 0.

$$W_{\varepsilon,c}(\mu,\nu) := \sum_{i,j} c_{ij}\pi_{ij} + \varepsilon \sum_{i,j} \pi_{ij}\left(\log(\pi_{ij}) - 1\right)$$

Remark 4.3.1 (Sinkhorn Divergence)**.** *However, one can modify the definition in the following way,*

$$S_{\varepsilon,c}(\mu,\nu) := W_{\varepsilon,c}(\mu,\nu) - \frac{1}{2}W_{\varepsilon,c}(\mu,\mu) - \frac{1}{2}W_{\varepsilon,c}(\nu,\nu)$$

Note that, now $S_{\varepsilon,c}(\mu,\mu) = 0$ but we do not necessarily have the non negativity of $S_{\varepsilon,c}(\mu,\nu)$ anymore.

- If μ, ν are measures with densities, how do we compute $W_2(\mu,\nu)$ and $W_{\epsilon,c}(\mu,\nu)$?
- What if we had samples $X_1, ..., X_n \sim \mu$ and $Y_1, ..., Y_n \sim \nu$?

Let $\mu_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ and $\nu_m = \frac{1}{m}\sum_{j=1}^m \delta_{Y_j}$. Note that $W_2(\mu_n,\nu_m)$ makes perfect sense, as opposed to total variation distance. Also note that $W_2(\mu_n,\nu_m)$ recovers our quantity of interest as $n, m \to \infty$. In order to make $W_2(\mu_n,\nu_n)$ close to $W_2(\mu,\nu)$, we would need $\left(\frac{1}{\delta}\right)^d + \left(\frac{1}{\delta}\right)^d$ samples, where d is the dimension of the space we are in.

Two of the reasons why people like to consider Sinkhorn algorithm are given below.

- It is a good algorithm.
- It has statistical advantages.

Consider $W_2(\mu_n,\nu_m)$.

$$W_2(\mu_n,\nu_m) = \min_{\pi_{n,m} \in \Gamma(\mu_n,\nu_m)} \int |x-y|^2 d\pi(x,y). \tag{34}$$

Denote $\pi^{*,n,m}$ to be the minimizer to the optimization problem above—we call it the Barrycentile projection of a coupling. For a given source point, $\pi^{n,m} : x_i \to \frac{\sum_j \pi_{ij}^{*,n,m} y_j}{\sum_j \pi_{ij}^{*,m,n}}$. Would this quantity approximate Monge transport?

4.4 Additional references related to entropy-regularized optimal transport.

Here are a few references that complement what was discussed during this lecture.

1. Sinkhorn algorithm may suffer from numerical instabilities that are caused when manipulating expressions like $e^{\phi/\varepsilon}$ for small ε. In Chapter 4.4 in Cuturi's and Peyre's book you can find a discussion on some techniques to ameliorate these instabilities.

2. As was mentioned in class, one of the advantages of working with entropy-regularized optimal transport, as opposed to working with standard optimal transport, is the favorable computational complexity of Sinkhorn's algorithm. We also mentioned another advantage: the *statistical* complexity of entropy-regularized OT when compared to standard OT (e.g., recall the results discussed in Bodhi's lecture for the standard OT). The bottom line is that the sample complexity for approximating the Wasserstein distance between two measures from two empirical measures does not suffer from the curse of dimensionality in the entropy-regularized case, while it does in the standard OT setting. You can take a look at the paper by Genevay et al that I added to our Dropbox folder.

3. Although the computational complexity result discussed in class was the one of *entropy*-regularized optimal transport, one can add a rounding subroutine to the Sinkhorn algorithm to find a near-linear time algorithm (roughly $O(n^2)$ operations) to find a *feasible* coupling $\tilde{\pi} \in \Gamma(\mu, \nu)$ (μ and ν empirical measures) that satisfies
$$\sum_{ij} c_{ij}\tilde{\pi}_{ij} \leq \min_{\pi \in \Gamma(\mu,\nu)} \sum_{ij} c_{ij}\pi_{ij} + \varepsilon$$
The rounding subroutine guarantees that the output of the algorithm satisfies all marginal constraints. Check Theorem 1 in the paper by Altschuler et al (Dropbox).

5 Lecture 5: 10/03/2023 (By Bodhisattva Sen)

Scribe: Michael Harding

In this lecture we consider Barycentric Projections of optimal transport plans on finite observed datasets as approximations to optimal transport maps. Our setting is as follows:

Setting

Consider two subsets $\mathcal{X}, \mathcal{Y} \subset \mathbb{R}^d$, where $d \geq 1$, and two measures, μ on $\mathcal{X}$ and ν on $\mathcal{Y}$, where $\mu \in \mathcal{P}_{\text{ac}}(\mathcal{X})$, the set of measures that are absolutely continuous with respect to the d-dimensional Lebesgue measure. Then consider two observed datasets, $X_1, \ldots, X_m \overset{\text{i.i.d.}}{\sim} \mu$, $Y_1, \ldots, Y_n \overset{\text{i.i.d.}}{\sim} \nu$, where $X_i \perp\!\!\!\perp Y_j \ \forall\ i = 1, \ldots, m,\ j = 1, \ldots, n$.

Goal

We aim to estimate the optimal transport map $T_0 : \mathbb{R}^d \to \mathbb{R}^d$, where
$$T_0 \coloneqq \min_{T : T_{\#}\mu} \mathbb{E}_\mu \left[\|X - T(X)\|_2^2\right] \equiv W_2^2(\mu, \nu), \tag{35}$$
where the equivalence to the Wasserstein distance is given by Brenier's Theorem. Also, by Brenier's Theorem, we know that there exists a convex function $\psi_0 : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ such that $T_0 = \nabla\psi_0$ which is μ-a.e. unique.

In summary, we hope to utilize our finite observed datasets to estimate the global optimal transport plan, leveraging Brenier's Theorem to both tell us that the optimal plan exists, and also characterize it in terms of the gradient of a convex function. In designing an estimator for T_0, we will hope that our construction

has some desirable convergence and/or computational properties in order to have performance guarantees and to apply our estimator in practice. To characterize our estimators, we break our problem formulation into two cases, the first where $n = m$, and the second where $n \neq m$.

Case 1: $m = n$

Consider

$$\hat{\mu}_m = \frac{1}{m}\sum_{i=1}^{m} \delta_{X_i}, \;\; \hat{\nu}_n = \frac{1}{n}\sum_{j=1}^{n} \delta_{Y_j}, \tag{36}$$

the empirical distributions of our datasets (with δ_Z being the discrete point mass at Z). Then we can write out estimator as

$$\widehat{T} = \underset{T:T_{\#}\hat{\mu}_m = \hat{\nu}_n}{\operatorname{argmin}} \frac{1}{m}\sum_{i=1}^{m} \|X_i - T(X_i)\|_2^2 \tag{37}$$

In this case, we are estimating our global transport map simply with the "local" analogue - rather than mapping μ to ν, we find the optimal transport plan mapping $\hat{\mu}_m$ to $\hat{\nu}_n$, which is exactly solvable when $m = n$. This is known as the **assignment problem**, and it is a special case of a linear program having "out of the box" solvers with $o(n^3)$ complexity readily available.

Case 2: $m \neq n$

This is clearly the more difficult, and thus more interesting, case. We know when $m \neq n$, there does not exist any function $T : \mathbb{R}^d \to \mathbb{R}^d$ such that $T_{\#}\hat{\mu}_m = \hat{\nu}_n$, and so we have no clear "local" analogue of the Monge problem that we are hoping to solve. However, we do know that the Kantorovich problem is always solvable, and so we can consider in some way utilizing the optimal transport plan $\hat{\pi}$, where

$$\hat{\pi} = \hat{\pi}_{m,n} = \underset{\pi \in \Gamma(\hat{\mu}_m, \hat{\nu}_n)}{\operatorname{argmin}} \int \|y - x\|_2^2 \, d\pi(x, y) \tag{38}$$

From here, the question then becomes, "how do we derive a transport map from $\hat{\pi}$?" and the proposed solution we present here is the **Barycentric Projection** of $\hat{\pi}$, given as

$$\widehat{T}(x) \equiv \widehat{T}_{m,n}(x) = \mathbb{E}_{\hat{\pi}}[Y|X = x], \; \hat{\mu}_m\text{-a.e.-}x \tag{39}$$

We first note that this definition for $\widehat{T}$ passes the basic requirement of reducing to the expected solution in the case where $m = n$, as Brenier's theorem tells us that the optimal transport plan is unique and is exactly the optimal transport map composed with the identity over $\mathcal{X}$. So, in that case $\hat{\pi}$ will give $Y|X = x$ a degenerate distribution at $\widehat{T}(x)$ as given in (37), and so the two definitions for $\widehat{T}$ are in fact equivalent when $m = n$.

5.1 Barycentric Projection

Now that we have utilized the barycentric projection of our discrete optimal transport plan to define our estimator $\widehat{T}$, we hope to charaterize its convergence properties, namely providing a bound r_n in terms of n and possibly d, where

$$\mathbb{E}\left[\frac{1}{m}\sum_{i=1}^{m} \|\widehat{T}(X_i) - T_0(X_i)\|_2^2\right] \lesssim r_n \tag{40}$$

In order to determine a proper rate for r_n, we first present some helpful definitions and a proposition, before stating and (mostly) proving a theorem giving an explicit rate for r_n.

Definition 5.1.1. *Let $f : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ be a convex function. The **Legendre-Fenchel dual** of f, $f^* : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$, is defined as*

$$f^*(y) = \sup_{x \in \mathbb{R}^d} \{x^\top y - f(x)\}$$

Definition 5.1.2. *Let $f : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$. A point $\xi \in \mathbb{R}^d$ is a **subgradient** of f at $x \in \mathbb{R}^d$ if and only if*

$$f(z) \geq f(x) + \xi^\top (x - z) \; \forall \; z \in \mathbb{R}^d$$

Definition 5.1.3. *Let $f : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$. The **subdifferential** of f at $x \in \mathbb{R}^d$, $\partial f(x)$, is the set of all subgradients of f at x, i.e.*

$$\partial f(x) = \{\xi_x \in \mathbb{R}^d : f(x) + \xi_x^\top (x - z) \leq f(z) \; \forall \; z \in \mathbb{R}^d\}$$

Note: in the case where f differentiable at x, then $\partial f(x) = \{\nabla f(x)\}$, and conversely, if $\partial f(x) = \{y\}$, then f differentiable at x and $\nabla f(x) = y$.

Remark 5.1.1. $y \in \partial f(x) \iff x \in \partial f^*(y) \iff f^*(y) + f(x) = x^\top y$

Proposition 5.1.1. *Let $T_0 : \mathbb{R}^d \to \mathbb{R}^d$ be the optimal transport map in (35), and let $\psi_0 : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ be the convex function such that $T_0 = \nabla \psi_0$. Then T_0 is L-Lipschitz if and only if ψ_0^* is $\frac{1}{L}$-strongly convex, i.e.*

$$\psi_0^*(y) + (\xi_y^*)^\top (y - z) + \frac{1}{2L}\|y - z\|_2^2 \leq \psi_0^*(z) \; \forall \; z \in \mathbb{R}^d, \xi_y^* \in \partial \psi_0^*(y)$$

Definition 5.1.4. *Let $T_0 : \mathbb{R}^d \to \mathbb{R}^d$ be the optimal transport map in (35). Then define*

$$\nu_m = \frac{1}{m} \sum_{i=1}^m \delta_{T_0(X_i)},$$

the empirical measure of the transport map pushforward. Note: by construction $T_{0\#}\hat{\mu}_m = \nu_m$, so T_0 is also the optimal transport map mapping $\hat{\mu}_m$ to ν_m by Bernier's theorem.

We now present and prove the main theorem of this lecture:

Theorem 5.1.1 (Barycentric projection convergence rate)**.** *Let $\mathcal{X}, \mathcal{Y} \subset \mathbb{R}^d$, $d \geq 1$, be compact subsets. Let μ a measure over $\mathcal{X}$, ν a measure over $\mathcal{Y}$, and $\mu \in \mathcal{P}_{ac}(\mathcal{X})$. Let $X_1, \ldots, X_m \overset{i.i.d.}{\sim} \mu$, $Y_1, \ldots, Y_n \overset{i.i.d.}{\sim} \nu$, $X_i \perp\!\!\!\perp Y_j \; \forall \; i = 1, \ldots, m, \; j = 1, \ldots, n$. Let $T_0 : \mathbb{R}^d \to \mathbb{R}^d$ be the optimal transport map taking μ into ν, and $\widehat{T}$ be the barycentric projection of the optimal transport plan, $\hat{\pi}_{m,n}$, taking the empirical distributions $\hat{\mu}_m$ into $\hat{\nu}_n$. Then, if T_0 is L-Lipschitz,*

$$\mathbb{E}\left[\frac{1}{m}\sum_{i=1}^m \|T_0(X_i) - \widehat{T}(X_i)\|_2^2\right] \lesssim \begin{cases} \frac{1}{\sqrt{n}} & d = 1, 2, 3 \\ \frac{\log n}{\sqrt{n}} & d = 4 \\ n^{-2/d} & d > 4 \end{cases}$$

Proof. Let $\psi_0 : \mathbb{R}^d \to \mathbb{R}$ be the convex function such that $T_0 = \nabla \psi_0$. We begin by denoting

$$D_1 \coloneqq \int \psi_0^*(y) \, d\hat{\nu}_n(y) - \int \psi_0^*(y) \, d\nu_m(y), \tag{41}$$

which is measuring the measure estimation error in terms of the Legendre-Fenchel dual of ψ_0. To transform

these integrals so they are with respect to the same measure, consider that for the first term, we can write

$$\begin{aligned}\int \psi_0^*(y)\, d\hat{\nu}_n(y) &= \int_{\mathcal{X}} \int_{\mathcal{Y}} \psi_0^*(y)\, d\hat{\pi}(x,y) \\ &= \int_{\mathcal{X}} \int_{\mathcal{Y}} \psi_0^*(y)\, d\hat{\pi}(y|x)\, d\hat{\mu}_m(x) \\ &\geq \int_{\mathcal{X}} \psi_0^* \left(\int_{\mathcal{Y}} y\, d\hat{\pi}(y|x) \right) d\hat{\mu}_m(x) \\ &= \int_{\mathcal{X}} \psi_0^*(\widehat{T}(x))\, d\hat{\mu}_m(x),\end{aligned}$$

where the first equality comes from the fact that the second marginal of $\hat{\pi}$ is $\hat{\nu}_n$ by construction, and the inequality is due to an application of Jensen's inequality, using the fact that ψ_0^* is convex. For the second term, we can simply use the fact that $T_{0\#}\hat{\mu}_m = \nu_m$ from definition 5.1.4 to state

$$\int \psi_0^*(y)\, d\nu_m(y) = \int \psi_0^*(T_0(x))\, d\hat{\mu}_m(x)$$

Thus, this allows us to write

$$\begin{aligned}D_1 &\geq \int_{\mathcal{X}} \left[\psi_0^*(\widehat{T}(x)) - \psi_0^*(T_0(x)) \right] d\hat{\mu}_m(x) \\ &\geq \int_{\mathcal{X}} \nabla \psi_0^*(T_0(x))^\top \left(\widehat{T}(x) - T_0(x) \right) d\hat{\mu}_m(x) + \frac{1}{2L} \int_{\mathcal{X}} \|T_0(x) - \widehat{T}(x)\|_2^2\, d\hat{\mu}_m(x) \\ &= \int_{\mathcal{X}} x^\top \left(\widehat{T}(x) - T_0(x) \right) d\hat{\mu}_m(x) + \frac{1}{2L} \int_{\mathcal{X}} \|T_0(x) - \widehat{T}(x)\|_2^2\, d\hat{\mu}_m(x),\end{aligned}$$

where the second inequality is due to an application of Proposition 5.1.1, using our assumption that T_0 is L-Lipschitz, and the equality is due to the interaction between Legendre-Fenchel duals and gradients, specifically $\nabla f^* = (\nabla f)^{-1}$. Therefore, we can rearrange the terms and write

$$\frac{1}{2L} \int_{\mathcal{X}} \|T_0(x) - \widehat{T}(x)\|_2^2\, d\hat{\mu}_m(x) \leq D_1 - D_2 + D_3, \tag{42}$$

where

$$D_2 \coloneqq \int_{\mathcal{X}} x^\top \widehat{T}(x)\, d\hat{\mu}_m(x), \quad D_3 \coloneqq \int_{\mathcal{X}} x^\top T_0(x)\, d\hat{\mu}_m(x)$$

Before continuing, we note that the LHS of (42) is exactly the inner term of the expectation in our theorem statement, scaled by $\frac{1}{2L}$, so we now wish to upper bound the RHS of (42). Going term by term,

$$\begin{aligned}D_2 &= \int_{\mathcal{X}} x^\top \widehat{T}(x)\, d\hat{\mu}_m(x) \\ &= \int_{\mathcal{X}} x^\top \left(\int_{\mathcal{Y}} y\, d\hat{\pi}(y|x) \right) d\hat{\mu}_m(x) \\ &= \int_{\mathcal{X}\times\mathcal{Y}} x^\top y\, d\hat{\pi}(x,y) \\ &= \min_{\phi \text{ convex}} \int \phi\, d\hat{\mu}_m + \int \phi^*\, d\hat{\nu}_n \\ &= \int \hat{\psi}\, d\hat{\mu}_m + \int \hat{\psi}^*\, d\hat{\nu}_n,\end{aligned}$$

for some convex function $\hat{\psi}$, and where the fourth equality comes from the strong duality of the Kantorovich problem (also see remark 5.1.1). Likewise, we can write

$$D_3 = \int \psi_0 \, d\hat{\mu}_m + \int \psi_0^* \, d\nu_m \leq \int \hat{\psi} \, d\hat{\mu}_m + \int \hat{\psi}^* \, d\nu_m$$

where the inequality is due to the fact that ψ_0 is the minimizer of the sum of the integrals. Combining all of these bounds and cancelling the $\int \hat{\psi} \, d\hat{\mu}_m$ terms, we can write

$$\begin{aligned} D_1 - D_2 + D_3 &\leq \int \left(\hat{\psi}^* - \psi_0^*\right)(y) \, d(\nu_m - \hat{\nu}_m)(y) \\ &\leq \int \left(\hat{\psi}^* - \psi_0^*\right)(y) \, d(\nu_m - \nu)(y) + \int \left(\hat{\psi}^* - \psi_0^*\right)(y) \, d(\nu - \hat{\nu}_n)(y), \end{aligned}$$

and therefore

$$\mathbb{E}\left[\frac{1}{2Lm}\sum_{i=1}^{m} \|T_0(X_i) - \widehat{T}(X_i)\|_2^2\right] \leq \mathbb{E}\left|\sup_{f \in \mathcal{F}} \int (f - \psi_0) \, d(\nu_m - \nu)\right| + \mathbb{E}\left|\sup_{f \in \mathcal{F}} \int (f - \psi_0) \, d(\nu - \hat{\nu}_n)\right|, \tag{43}$$

where

$$\mathcal{F} \coloneqq \{f : \mathcal{Y} \to \mathbb{R} : f \text{ convex, bounded, Lipschitz}\}$$

By construction, we know that $\hat{\psi}^*$ is convex, but we are still left with the task of showing it is bounded and Lipschitz in order to use this bound, but these technical pieces are omitted from this proof. Instead, we will take for granted that this is the case, and then close the proof by using Dudley's inequality and results from Empirical Process Theory to bound the supremum terms on the RHS of (43) with the terms in the theorem statement. At a high level, we use the covering number for $\mathcal{F}$, which for an ε-covering is $\sim \int_0^1 \varepsilon^{-d/2}$. For the details of these final steps, see Deb, N., Ghosal, P., & Sen, B. (2021).

□

Remark 5.1.2. *Notice that the inequality in (43) replaces $\hat{\psi}^*$ with the worst case function $f \in \mathcal{F}$. This is a fairly naive bound, as we should have $\hat{\psi}^* \to \psi_0^*$ in some sense as $m, n \to \infty$, by the fact that we have convergence in $\widehat{T} \to T_0$, so it is possible to achieve a tighter bound, and a significantly tighter one as m, n grow large. In particular, it is known that $n^{-1/2}$ for $d = 1, 2, 3$ is certainly not optimal and the (non)optimality of the $\frac{\log n}{\sqrt{n}}$ for $d = 4$ is still an open question. However, contrary to this line of thinking, for $d > 4$, the "curse" of dimensionality is actually a boon for our proof technique, as the naive bound actually still achieves the optimal rate in n!*

Remark 5.1.3. *The compactness of $\mathcal{X}, \mathcal{Y}$ assumed in the statement of the theorem is helpful, but ultimately an unnecessary assumption. In fact, Deb, N., Ghosal, P., & Sen, B. (2021) states and proves this theorem only assuming $\mathcal{Y}$ compact, though there have as of yet been no results that relax this assumption for both $\mathcal{X}$ and $\mathcal{Y}$.*

Remark 5.1.4. *In the proof, we heavily leverage the fact that T_0 is L-Lipschitz through the relationship in proposition 5.1.1, where we are then given that ψ_0^* is $\frac{1}{L}$-strongly convex. In proving the bounds on the term of interest, this strong convexity assumption causes the desired $\|T_0(x) - \widehat{T}(x)\|_2^2$ term to seemingly "pop out" of nowhere. Some questions that this raises: given μ, ν, how do we know T_0 will be L-Lipschitz? What kind of conditions might we need to impose on μ, ν to induce this property? And what different assumptions might we make on T_0 to generalize a result like this to general cost functions?*

6 Lecture 6: 10/12/2023

Scribes: Jingyun Jia

This week, we extended the previous results to the multimarginal optimal transport(M-OT) problem and discussed the connection between M-OT and Wasserstein Barycenter, as well as the application in adversarial learning.

6.1 Optimality Condition for Wasserstein Barycenter

Without proof, we borrow the results from one-dimensional optimal transport problem and extend it to the M-OT problem.
We first state the M-OT problem and the dual problem.
For $\mu_1, \ldots, \mu_k \in \mathcal{P}(\mathcal{X})$ and a function $C : \mathcal{X} \times \cdots \times \mathcal{X} \to [0, \infty]$, the optimization problem of the M-OT is:

$$\min_{\gamma \in \Gamma(\mu_1, \ldots, \mu_k)} \int \mathbf{C}(x_1, \ldots, x_k) d\gamma(x_1, \ldots, x_k),$$

where $\Gamma(\mu_1, \ldots, \mu_k) = \{\gamma \in \mathcal{P}(x^* : \text{i-th marginal of } \gamma \text{ is equal to } \mu_i\}$. The corresponding dual problem is $\sup_{\phi_1, \ldots, \phi_k} \int \phi_i(x_i) d\mu_i(x_i)$, such that $\sum \phi_i(x_i) \leq \mathbf{C}(x_1, \ldots, x_k)$.
For Wasserstein Barycenter, our goal is to minimize the weighted loss $\varepsilon(\mu)$,

$$\inf_{\mu \in \mathcal{P}(\mathcal{X})} \varepsilon(\mu) = \inf_{\mu \in \mathcal{P}(\mathcal{X})} \sum_{i=1}^{k} \alpha_i C(\mu_i, \mu),$$

where $\alpha_i \in [0, 1]$, $\sum_{i=1}^{k} \alpha_i = 1$, $C(\mu_i, \mu) = min_{\pi_i \in \Gamma(\mu_i . \mu)} \int c(x_i, y) d\pi_i(x_i, y)$, Γ denotes the space of coupling and $c(\cdot, \cdot)$ is some loss function, such as $c(x, y) = |x - y|^2$.
Suppose μ^* solves the above problem, we have $\frac{d}{dt}\varepsilon((1-t)\mu^* + tv) = 0|_{t=0}$ for a random variable t between 0 and 1. By duality, we rewrite

$$C(\mu_i, \mu) = \sup_{\phi_i, \psi_i, \phi_i + \psi_i < c(\cdot, \cdot)} \int \phi_i d\mu_i, \int \psi_i d\mu = \int \phi_i^* d\mu_i, \int \psi_i^* d\mu.$$

Plug it into the derivative and get:

$$\begin{aligned}
\frac{d}{dt}\varepsilon((1-t)\mu^* + tv)|_{t=0} &= \sum_i \alpha_i \frac{d}{dt} C(\mu_i, (1-t)\mu^* + tv)|_{t=0} \\
&= \sum_i \alpha_i \frac{d}{dt} sup_{\phi, \psi, \phi + \psi < c(\cdot, \cdot)} (\int \phi d\mu_i^* + (1-t) \int \psi d\mu^* + t \int \psi) dv|_{t=0} \\
&= \sum_i \alpha_i (\int \psi_i^* dv - \int \psi_i^* d\mu^*) \\
&= \sum_i \alpha_i (\int \psi_i^*(x)(dv(x) - \mu^*(x))) = 0
\end{aligned}$$

We get the third equation by Danskin's theorem and letting t=0. Then the optimizer ψ_i must satisfy $\sum \alpha_i \psi_i^*(x) = C$, C is some constant.

Theorem 6.1.1 (Danskin's Theorem)**.** *Suppose $\sigma(x) = \max_y g(x, y)$, then $\partial_x \sigma(x) g(x, y^*(x))$, where $y^*(x) \in \operatorname{argmax}_y g(x, y)$.*
Proof idea: Chain rule.

6.2 Connection between M-OT and Wasserstein Barycenter

In this section, we show that one can recover a solution to the M-OT problem using the Wasserstein barycenter μ^* and couplings π_i realizing the costs $C(\mu_i, \mu^*)$.

Formally, let the cost in M-OT be $\mathbf{C}(x_1,\dots,x_k) := \inf_y \sum \alpha_i C(x_i,y)$, μ^* solves Wasserstein Barycenter and $\pi_i^* \in \operatorname{argmin}_{\pi_i\in\Gamma(\mu_i,\mu^*)} \int C(x_i,y)d\pi_i(x,y)$. Then we define a measure $\tilde{\beta} \in \mathcal{P}(\mathcal{X}^k \times \mathcal{X})$, such that $d\tilde{\beta}(x_1,\dots,x_k,y) = (d\pi_1^*(x_1|y)\cdots d\pi_1^*(x_k|y))d\mu^*(y)$, and $\tilde{r}$ is marginal of $\tilde{\beta}$ onto $x_1,\dots,x_k$.

$$\begin{aligned}
\text{M-OT} &\le \int \mathbf{C}(x_1,\dots,x_k)d\tilde{r}(x_1,\dots,x_k)\\
&= \int\int \mathbf{C}(x_1,\dots,x_k)d\tilde{\beta}(x_1,\dots,x_k)\\
&\le \int\int \sum_i \alpha_i C(x_i,y)d\tilde{\beta}(x_1,\dots,x_k)\\
&= \sum_i \alpha_i \int\int C(x_i,y)d\pi_i^*(x_i,y)\\
&= \sum_i \alpha_i C(\mu_i,\mu^*) = \text{Wasserstein Barycenter} \le \text{M-OT}
\end{aligned}$$

Moreover, one can solve Wasserstein Barycenter from M-OT solution π^* by defining the pushforward measure $v^* = T_{\#}\pi^*$, where $T(x_1,\dots,x_k) = \operatorname{argmin}_y \sum_i \alpha_i C(x_i,y)$. In this way, we show the equivalence of Wasserstein Barycenter and M-OT.

Remark 6.2.1. *Although the M-OT problem has a nice format and can be solved by linear programming, when the dimension k is large, it is computationally heavy to optimize.*

6.3 Generalized Barycenters and Adversarial Learning

In this section, we discuss the application of M-OT in adversarial learning. In adversarial learning, we are interested in the optimization problem:

$$\inf_{f\in\mathcal{F}} \sup_{\tilde{\mu_0}\in\mathcal{P}(\mathcal{X}\times\mathcal{Y})} \{R(f,\tilde{\mu_0}) - C(\mu,\tilde{\mu_0})\},$$

where $\mathcal{F} = \{f(x) = (f_1(x),\dots,f_k(x)) : x \to \Delta y\}$, R represents the risk function and C measures the distance between distributions. Denote $Z = \mathcal{X}\times\mathcal{Y}$ and we pick the distance function $C(\mu_0,\tilde{\mu_0}) = \min_{\pi\in\Gamma(\mu_0,\tilde{\mu_0})} \int C_Z(z,\tilde{z})d\pi(z,\tilde{z})$(here, μ_0 denotes measures in $\mathcal{P}(\mathcal{X}\times\mathcal{Y})$, with

$$C_Z(z,\tilde{z}) = \begin{cases} C(x,\tilde{x}) & \text{if } y = \tilde{y} \\ \infty & \text{otherwise} \end{cases},$$

For the risk function, we use extended 0-1 loss to weak classifier $l(f(\tilde{x}),y) = 1 - f_y(\tilde{x})$, the corresponding risk function is $R(f,\tilde{\mu}) = \mathbf{E}_{(\tilde{x},\tilde{y})\sim\mu}[l(f(\tilde{x}),\tilde{y})]$.
We then use μ_i to represent the positive measure for μ,

$$\mu_i(A) := \mu(A\times i),$$

A is any measurable set in $\mathcal{X}$, and $\mu = (\mu_1,\dots,\mu_k)$ and similarly, $\tilde{\mu} = (\tilde{\mu}_1,\dots,\tilde{\mu}_k)$. With the above distance function C and C_Z, we define $C(\mu,\tilde{\mu}) = \sum_i \mathcal{E}(\mu_i,\tilde{\mu_i})$ and

$$\mathcal{E}(\mu_i,\tilde{\mu_i}) = \inf_{\pi_i\in\Gamma(\mu_i,\tilde{\mu_i})} \int C(x_i,\tilde{x_i})d\pi_i(x_i,\tilde{x_i}).$$

Let's move back to the adversarial learning problem,

$$\begin{aligned}
\min_{f\in\mathcal{F}} \sup_{\tilde{\mu_0}\in\mathcal{P}(\mathcal{X}\times\mathcal{Y})} \{R(f,\tilde{\mu_0}) - C(\mu,\tilde{\mu_0})\} &= \min_{f\in\mathcal{F}} \sup_{\tilde{\mu_1},\ldots,\tilde{\mu_k}} \sum_i \int l(f(\tilde{x}),i) d\tilde{\mu}_i(\tilde{x}) - \sum_i \mathcal{E}(\mu_i,\tilde{\mu}_i) \\
&= \sup_{\tilde{\mu_1},\ldots,\tilde{\mu_k}} \min_{f\in\mathcal{F}} \sum_i \int l(f(\tilde{x}),i) d\tilde{\mu}_i(\tilde{x}) - \sum_i \mathcal{E}(\mu_i,\tilde{\mu}_i) \\
&= \min_{\tilde{\mu_1},\ldots,\tilde{\mu_k}} \max_{f\in\mathcal{F}} (\sum_i \int l(f(\tilde{x}),i) d\tilde{\mu}_i(\tilde{x})) + \sum_i \mathcal{E}(\mu_i,\tilde{\mu}_i) \\
&:= \min_{\tilde{\mu_1},\ldots,\tilde{\mu_k}} \int d\tilde{\mu}_{max}(\tilde{x}) + \sum_i \mathcal{E}(\mu_i,\tilde{\mu}_i) \\
&= \min_{\tilde{\mu_1},\ldots,\tilde{\mu_k},\lambda\geq\tilde{\mu_i}} \lambda(\tilde{x}) + w\sum_i \mathcal{E}(\mu_i,\tilde{\mu}_i)
\end{aligned}$$

We get the third equation by plugging the value of $l(f(\tilde{x}),i)$ and removing negative signs, $d\tilde{\mu}_{max}(\tilde{x}) = \max_i\{\frac{d\tilde{\mu}_i(\tilde{x})}{d\sum_i \tilde{\mu}_i(\tilde{x})}\} d\sum_j \mu_j(x)$.

Remark 6.3.1. *In order to have couplings $(\mu_i,\tilde{\mu}_i)$, it is sufficient that μ_i and $\tilde{\mu}_i$ have the same total mass. We do not require them to be probability measures. And when μ_i and $\tilde{\mu}_i$ have different total mass, the distance $C(\mu,\tilde{\mu}) = \infty$.*

Question: In the adversarial learning problem, the optimization problem is $\min_{\tilde{\mu_1},\ldots,\tilde{\mu_k},\lambda\geq\tilde{\mu_i}} \lambda(\tilde{x}) + w\sum_i \mathcal{E}(\mu_i,\tilde{\mu}_i)$. When will the problem coincide with the Wasserstein Barycenter(i.e. $min_v \sum_i C(\mu_i,v)$), $w\to 0$ or $w\to\infty$?

6.4 Remarks after Lecture 6

1. As with the standard OT problem, one can introduce an entropy-regularized version of M-OT as follows:

$$\min_{\gamma\in\Gamma(\mu_1,\ldots,\mu_k)} \int \mathbf{C}(x_1,\ldots,x_k) d\gamma(x_1,\ldots,x_k) + \varepsilon \mathrm{KL}(\gamma||\mu_1\otimes\cdots\otimes\mu_k),$$

   where $\mu_1\otimes\cdots\otimes\mu_k$ denotes the product measure of the measures $\mu_1,\ldots,\mu_k$. Following analogous computations as in Lecture 4, in the discrete case we can rewrite the regularized problem conveniently and deduce that its unique solution takes the form:

$$\pi(x_1,\ldots,x_k) = \exp\left(\frac{1}{\varepsilon}\sum_{l=1}^k \phi_l(x_l) - \frac{1}{\varepsilon}\mathbf{C}(x_1,\ldots,x_k)\right),$$

   where the functions ϕ_l have to be chosen so that the resulting pmf satisfies all marginal constraints.

   A Sinkhorn-like iteration is suggested by the above optimality condition: after choosing $l\in\{1,\ldots,k\}$ and keeping the potentials $\phi_1,\ldots,\phi_{l-1},\phi_{l+1},\ldots,\phi_k$ fixed, we update the potential ϕ_l as follows:

$$\mu_l(x_l)\left(\sum_{x_{-l}} \exp\left(\frac{1}{\varepsilon}\sum_{l'\neq l}\phi_{l'}(x_{l'}) - \frac{1}{\varepsilon}\mathbf{C}(x_1,\ldots,x_k)\right)\right)^{-1} = \exp(\phi_l(x_l)/\varepsilon), \quad \forall x_l.$$

   In the above we use the shorthand notation $x_{-l} = (x_1,\ldots,x_{l-1},x_{l+1},\ldots,x_k)$.

   In other words, we update ϕ_l to obtain the correct l-th marginal constraint.

   One question remains: how do we choose the potential to be updated? One approach is to select l greedily:

$$\operatorname*{argmax}_{l=1,\ldots,k} D_{\mathrm{KL}}\left(\mu_l||\pi_l\right).$$

Here we use π_l to denote the l-th marginal of π. An analysis of the resulting iterative M-OT Sinkhorn algorithm can be found in the work by Altschuler et al that can be found in our dropbox folder.

2. For the barycenter problem
$$\min_{\mu\in\mathcal{P}(\mathcal{X})} \sum_{l=1}^{k} \alpha_i C(\mu_l, \mu)$$
we derived the following optimality conditions:
$$\text{const} = \sum_{l=1}^{k} \alpha_l \psi_l,$$
where ψ_l satisfies
$$C(\mu_l, \mu) = \int \psi_l^{\bar{c}} d\mu_l + \int \psi_l d\mu,$$
i.e., ψ_l is an optimal dual potential. Assuming that μ has a density w.r.t. Lebesgue measure, and assuming that $c(x, y) = |x - y|^2$, we know by Brenier's theorem that
$$\nabla\psi_l = Id - T_l^*,$$
where T_l^* is the optimal transport map between μ and μ_l. Therefore, in this case the optimality condition can be written as
$$0 = \sum_{l=1}^{k} \alpha_l (Id - T_l^*),$$
or in other words
$$x = \sum_{l=1}^{k} \alpha_l T_l^*(x), \quad \forall x.$$
When the measures $\mu_1, \ldots, \mu_k$ are all Gaussians, one can make an "informed guess" and assume that μ is a Gaussian $N(\overline{m}, \overline{\Sigma})$. One can then plug in this guess in the optimality condition and infer from this some equations for $\overline{m}$ and $\overline{\Sigma}$. You can see details about this in the paper by Agueh and Carlier that I added to our Dropbox folder.

3. You can read more about the connections between adversarial training, generalized Wasserstein barycenters, and multimarginal OT in the two papers that I added to our Dropbox folder that my colleague Matt Jacobs and my former PhD student Jakwang Kim wrote with me.

7 Lecture 7: 10/19/2023

Scribe: Shan Leng and Bingyan Liang
This week, we first reviewed materials we have covered before; then continued to discuss dynamic OT.

7.1 Review

OT Problem: Primal & Dual

1. Primal: Kantorovich's: $\inf_{\Pi\in\Gamma(\mu,\nu)} \int c(x, y)\, d\Pi(x, y)$, where solutions exist but uniqueness is not guaranteed.

   Monge's: $\inf_{T:T\#\mu=\nu} \int c(x, T(x))\, d\mu(x)$

   When $\mathcal{X} = \{x_1, ..., x_n\}$, $\mathcal{Y} = \{y_1, ..., y_n\}$ and μ, ν are uniform measures on $\mathcal{X}$ and $\mathcal{Y}$ respectively, then Monge's formulation of optimal transport is equivalent to Kantorovich's formulation(Brenier theorem).

2. Dual: $\sup_{\phi\in L^1(\nu),\psi\in L^1(\nu)} \int_{\mathcal{X}} \phi d\mu + \int_{\mathcal{Y}} \psi d\nu$ such that $\phi(x) + \psi(y) \le c(x, y)$

3. OT: Measures μ, ν are Gaussian, $T^* = \nabla\varphi$ s.t. $T_\#^*\mu = \nu$.

Entropy Regularized OT

$$\min_{\pi\in\Gamma(\mu,\nu)} \int c(x,y)\,\mathrm{d}\pi(x,y) + \varepsilon\,\mathrm{KL}(\pi\|\mu\otimes\nu)$$

Sinkhorn is nothing but coordinate ascent of the Dual of E-OT.

Generalization: Multimarginal OT

$$\min_{\gamma\in\Gamma(\mu_1,\ldots,\mu_k)} \int \mathbf{C}(x_1,\ldots,x_k)d\gamma(x_1,\ldots,x_k)$$

7.2 Dynamic OT

Suppose $x \in \mathbf{R}^d$, $\{\vec{V}_t\}_{t\in(0,1)}$ is a vector field where $\vec{V}_t(x) \in \mathbf{R}^d$ is Lipschitz

$$\begin{cases} \dot{X}_t = \vec{V}_t(X_t) \\ x_0 = x \end{cases}$$

We set $F_t(x) := X_t$, the solution of the ODE at time t, when the initial condition for this ODE is x.

Question: What can we say about the family of probability measures:

$$t \in [0,1] \mapsto F_{t\#}\mu_0, \text{ where } \mu_0 \text{ is some fixed probability measures.}$$

Let $\phi \in \mathbf{C}_0^\infty(\mathbf{R}^d)$, then we have the continuous equation:

$$\begin{aligned} \frac{d}{dt}\int \phi(x)d\mu_t(x) &= \frac{d}{dt}\int \phi(\tilde{x})dF_{t\#}\mu_0(\tilde{x}) \\ &= \frac{d}{dt}\int \phi(F_t(x))d\mu_0(x) \\ &= \int \frac{d}{dt}\phi(F_t(x))d\mu_0(x) \\ &= \int \nabla\phi(F_t(x))\frac{d}{dt}F_t(x)d\mu_0(x) \\ &= \int \nabla\phi(F_t(x))\vec{V}_t(F_t(x))d\mu_0(x) \\ &= \int \nabla\phi(\tilde{x})\vec{V}_t(\tilde{x})d\mu_t(\tilde{x}) \end{aligned}$$

Definition 7.2.1. *We say that the path $t \mapsto (\mu_t, \vec{v_t})$ solves the continuous equation above if*

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

Question: Why this notation?

Suppose $d\mu_t = \rho_t dx$ and ρ_t is smooth in t and in x,

$$\frac{d}{dt}\int \phi(\tilde{x})\rho_t(\tilde{x})d\tilde{x} = \int \phi(\tilde{x})(\frac{\partial}{\partial t}\rho_t(\tilde{x}d\tilde{x})$$

Suppose the order of integration and derivative could be changed

$$\int \phi(\tilde{x})\vec{V}_t(x)\rho_t(\tilde{x})d\tilde{x} = -\int \phi(\tilde{x})div(\rho_t\vec{V}_t)d\tilde{x} \Rightarrow \int \phi(\tilde{x})(\frac{\partial}{\partial t}\rho_t(\tilde{x}) + div(\rho_t\vec{V}_t))d\tilde{x} = 0$$

Thus, for $\forall\phi \in \mathbf{C}_0^\infty$, $\frac{\partial}{\partial t}\rho_t(\tilde{x}) + div(\rho_t\vec{V}_t) = 0$.

1. Continuity equation $\partial_t \mu_t + \mathrm{div}(\mu_t \vec{v_t}) = 0$ for some choice of vector space $\{\vec{V_t}\}_{t\in(0,1)}$.
   - $\vec{V_t}$ is the velocity of the curve at time t.
   - $\langle \vec{v}, \vec{w} \rangle_\mu := \int \vec{v_t}(x)\vec{w_t}(x)d\mu(x)$
2. With the interpretation, it looks like $\rho_2(\mathbf{R}^d)$ can be endowed with a formal Riemann structure.
3. $t \mapsto (\mu_t, \vec{v_t})$ $\qquad \int_0^1 |\vec{v_t}(x)|^2 d\mu_t(x)dt = \int_0^1 \langle \vec{v_t}, \vec{v_t} \rangle_{\mu t} dt$

Theorem 7.2.1 (Benamou-Brenier)**.**

$$inf_{\partial_t \mu_t + div(\vec{v_t}\mu_t)=0, \mu_0=\mu_0, \mu_1=\mu_1} \int_0^1 \int |\vec{V_t}(x)|^2 d\mu_t(x)dt = \mathbf{W}_2^2(\mu_0, \mu_1)$$

proof: Assume μ_0 has a density w.r.t. Lebeague:

$$\begin{aligned}
\int_0^1 \int |\vec{V_t}(x)|^2 d\mu_t(\tilde{x})dt &= \int_0^1 \int |\vec{V_t}(F_t(x))|^2 d\mu_0(x)dt \\
&= \int \int_0^1 |\vec{V_t}(F_t(x))|^2 d\mu_0(x)dt \\
&\geqslant \int |F_1(x) - x|^2 d\mu_0(x) \\
&\geqslant \mathbf{W}_2^2(\mu_0, \mu_1)
\end{aligned}$$

Question: Let T be the optimal transparent map between μ_0 and μ_1, $\mu_1 = T_{\#}\mu_0$, $\mu_t = T_{t\#}\mu_0$, $T_t(x) = (1-t)x + tT(x)$; What is the OT map between μ_0 and μ_t?

$$T_t(x) = (1-t)x + tT(x) = (1-t)\nabla \frac{|x|^2}{2} + t\nabla\rho_1(x) = \nabla((1-t)\frac{|x|^2}{2} + t\rho_1)$$

$$\phi_t^{(x)} = (1-t)\frac{|x|^2}{2} + t\phi_1(x)$$

Question: What is the OT map between μ_s and μ_t for $s < t$?

In search of $\vec{V_t}$, It's okay to define $\vec{V_t}(T_t(x)) = T(x) - x$ as long as $T_t(x) = T_t(x') \Rightarrow x = x'$

$$\begin{aligned}
(1-t)x + tT(x) &= (1-t)x' + tT(x') \\
(1-t)(x-x') = t(T(x') - T(x)) &= t(\nabla\phi(x') - \nabla\phi(x)) \\
\Rightarrow (1-t)|x-x'|^2 = t\langle \nabla\phi(x') - \nabla\phi(x), x - x' \rangle &= -t\langle \nabla\phi(x') - \nabla\phi(x), x' - x \rangle
\end{aligned}$$

Then we consider the OT map:

$$\begin{aligned}
&\frac{d}{dt}\int \phi(\tilde{x})d\mu_t(\tilde{x}) \\
&\qquad = \frac{d}{dt}\int \phi(T_t(x))d\mu_0(x) \\
&\qquad = \int \nabla\phi(T_t(x))\frac{d}{dt}T_t(x)d\mu_0(x) \\
&\qquad = \int \nabla\phi(T_t(x))(T(x) - x)d\mu_0(x) \\
&\qquad = \int \nabla\phi(\tilde{x})\vec{V_t}(\tilde{x})d\mu_t(\tilde{x})
\end{aligned}$$

Therefore we have$\vec{V}_t(x) = T(T_t^{-1}(\tilde{x}) - T_t^{-1}(x)$. Then,

$$\begin{aligned}\int_0^1 \int |\vec{V}_t(x)|^2 d\mu_t(\tilde{x})dt &= \int_0^1 \int |T(T_t^{-1}(\tilde{x}) - T_t^{-1}(x)|^2 d\mu_t(\tilde{x})dt \\ &= \int_0^1 \int |T(T_t^{-1}(x) - T_t^{-1}(T_t(x))|^2 d\mu_0(\tilde{x})dt \\ &= \int_0^1 \int |T(x) - x|^2 d\mu_0(\tilde{x})dt \\ &= \int |T(x) - x|^2 d\mu_0(\tilde{x}) \\ &= W_2^2(\mu_0, \mu_1)\end{aligned}$$

That is, $\inf \int \langle \vec{v_t}, \vec{v_t} \rangle_{\mu t} dt = W_2^2(\mu_0, \mu_1)$.

Remark 7.2.1. *For $|x_0 - x_1|^2 = \inf_{\gamma:[0,1]\mapsto \mathbf{R}^d, \gamma(0)=x_0, \gamma(1)=x_1} \int |\gamma(t)|^2 dt$, the minimizer should have the format $\gamma^*(t) = (1-t)x_0 + tx_1$; then $\ddot{\gamma}(t) = 0$.*

Proof.

$$\begin{cases} \gamma(\cdot, s), s \in (-\epsilon, \epsilon) \\ \gamma(0, s) = x_0, \gamma(1, s) = x_1 \end{cases}$$

Then,

$$\begin{aligned} 0 = \frac{d}{ds}\Big|_{s=0} \int_0^1 |\frac{\partial}{\partial t}\gamma(t,s)|^2 dt &= \int_0^1 \frac{d}{ds}\Big|_{s=0} |\frac{\partial}{\partial t}\gamma(t,s)|^2 dt \\ &= 2\int_0^1 \frac{\partial}{\partial t}\gamma(t,0) \cdot \frac{d}{ds}\Big|_{s=0} \frac{d}{dt}\gamma(t,s)dt \\ &= 2\int_0^1 \frac{d}{dt}\gamma^*(t) \cdot \frac{d}{dt}\frac{d}{ds}\Big|_{s=0} \gamma(t,s)dt = 2\int_0^1 \frac{d}{dt}\gamma^*(t) \cdot \frac{d}{dt}\vec{\mu_t} \\ &= -2\int_0^1 \frac{d^2}{dt^2}\gamma^*(t)\vec{\mu_t}dt \end{aligned}$$

The above equation should hold for any perturbation $\vec{\mu_t}$. So $\forall t, \ddot{\gamma}^*(t) = 0$.

7.3 A remark after Lecture 7

On general Geodesics in the 2-Wasserstein space. When discussing geodesics in an arbitrary metric space $(\mathbb{M}, d)$ (notice we may not have a Riemannian structure), the following definition is given: we say that a continuous map $\gamma : [0,1] \to \mathbb{M}$ is a minimizing geodesic connecting the points x, y if $\gamma_0 = x$, $\gamma_1 = y$ and the following relation holds:

$$d(\gamma_t, \gamma_s) = |t - s| d(x, y), \quad \forall s, t \in [0,1].$$

In the 2-Wasserstein space, we discussed that if μ_0 is absolutely continuous w.r.t. Lebesgue measure and μ_1 is arbitrary (both with finite second moments), then $\mu_t := T_{t\sharp}\mu_0$ with

$$T_t(x) = (1-t)x + tT(x),$$

and T the Monge map between μ_0 and μ_1, is a minimizing geodesic between μ_0 and μ_1.

It turns out that we don't need the existence of Monge maps to prove the existence of minimizing geodesics. To see this, let π^* be an optimal coupling between μ_0 and μ_1 (i.e., a solution to the Kantorovich problem, which we know always exists). Consider the map

$$F_t(x, y) := (x, (1-t)x + ty), \quad (x, y) \in \mathbb{R}^d \times \mathbb{R}^d,$$

and define $\pi_t := F_{t\sharp}\pi^*$. Finally, let μ_t be the second marginal of π_t. One can verify that $t \in [0,1] \to \mu_t$ is a minimizing geodesic between μ_0 and μ_1.

8 Lecture 8: 10/26/2023

Scribes: Jiayang Wang and Xinyan Wang

8.1 Review

Consider $\mathbf{W}_2(\mu_0, \mu_1)$.

$$\mathbf{W}_2(\mu_0, \mu_1) = \min_{\substack{\partial_t \mu_t + div(\vec{v_t}\mu_t) = 0 \\ \mu_0 = \mu_0, \mu_1 = \mu_1}} \int_0^1 \int |\vec{V_t}(x)|^2 d\mu_t(x) dt. \tag{44}$$

When μ_0 is a.c. w.r.t Lebesgue, then $t \in [0,1]$, we have $\mu_t = T_{t\#}\mu_0$, where $T_t(x) := (1-t)x + tT(x)$ (where T is the Monge map between μ_0 and μ_1)

Given $\phi \in C_c^\infty(\mathbf{R}^d)$

$$\begin{aligned}
&\frac{d}{dt}\int \phi(\tilde{x}) d\mu_t(\tilde{x}) \\
&\qquad = \frac{d}{dt}\int \phi(T_t(x)) d\mu_0(x) \\
&\qquad = \int \nabla\phi(T_t(x)) \frac{d}{dt} T_t(x) d\mu_0(x) \\
&\qquad = \int \nabla\phi(T_t(x))(T(x) - x) d\mu_0(x) \\
&\qquad = \int \nabla\phi(\tilde{x}) \vec{V_t}(\tilde{x}) d\mu_t(\tilde{x}) \\
&\qquad = \int \nabla\phi(T_t(x)) \left(T(T_t^{-1}(T_t(x))) - T_t^{-1}(T_t(x))\right) d\mu_0(\tilde{x}) \\
&\qquad = \int \nabla\phi(\tilde{x}) \left(T(T_t^{-1}(\tilde{x})) - T_t^{-1}(\tilde{x})\right) d\mu_t(\tilde{x})
\end{aligned}$$

Claim: $\vec{V_t}(\tilde{x}) = \nabla\phi_t(\tilde{x})$.

So that, we can rewrite it as:

$$\min_{\substack{\partial_t \mu_t + div(\vec{v_t}\mu_t) = 0 \\ \mu_0 = \mu_0, \mu_1 = \mu_1}} \int_0^1 \int |\vec{V_t}(x)|^2 d\mu_t(x) dt = \min_{\substack{\partial_t \mu_t + div(\vec{v_t}\mu_t) = 0 \\ \mu_0 = \mu_0, \mu_1 = \mu_1}} \int_0^1 \int |\nabla\phi_t(\tilde{x})|^2 d\mu_t(x) dt. \tag{45}$$

8.2 Geometry of the 2-Wasserstein space

Geodesic equations:

$$\begin{cases} \partial_t \mu_t + div(\vec{v_t}\mu_t) = 0 \\ \partial_t \phi_t + \frac{1}{2}|\nabla\phi_t|^2 = 0 \end{cases} \tag{46}$$

Give ϕ_0 and μ_0: exp_{μ_0}: $\phi_0 \to \mu_1$ (Exponential map)

Opposite: log_{μ_0}: $\mu_1 \to \phi_0$ (Logarithmic map)

To solve OT problems: $\mathbf{W}_2^2(\mu_0, \mu^i)$ for i = 1, ..., N.

$$\mathbf{W}_2^2(\mu_0, \mu^i) = \int |T_i^*(x) - x|^2 d\mu_0(x)$$

We have, T_i^* is OT map between μ_0 and μ^i; ψ_i is the convex function that given $T_i^* = \nabla\psi_i$; $\phi_{0,i}(x) = \psi_i(x) - \frac{\|x\|^2}{2}$; $log_{\mu_0}(\mu^i) = \phi_{0,i}$.

Distance in $T_{\mu_0}\rho_2(\mathbb{R}^d)$:

$$\begin{aligned} dist^2(log_{\mu_0}(\mu^i), log_{\mu_0}(\mu^j)) &= dist^2(\phi_0^i, \phi_0^j) \\ &= \int \|\nabla\phi_0^i(x) - \nabla\phi_0^j\|^2 d\mu_0(x) \\ &= \left\langle \nabla\phi_0^i - \nabla\phi_0^j, \nabla\phi_0^i - \nabla\phi_0^j \right\rangle_{\mu_0} \end{aligned}$$

Remark: we have that $dist^2(log_{\mu_0}(\mu^i), 0) = \mathbf{W}_2^2(\mu_i, \mu_0)$
This map makes the computation fast. In general, there are two critical ideas of designing a map: 1."cheap" 2."Have geometric connection".

8.3 Gradient Flows

In the Euclidean case, given a function $F : \mathbb{R}^d \to \mathbb{R}$, its gradient is defined as follows:

$$\nabla F(x) = \begin{pmatrix} \frac{\partial}{\partial x_1} F(x) \\ \vdots \\ \frac{\partial}{\partial x_n} F(x) \end{pmatrix}$$

Then the first variation of F at x is defined as

Definition 8.3.1. *Given a function $F : \mathbb{R}^d \to \mathbb{R}$, its first variation is defined as*

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

where $\gamma(0) = x, \gamma : [0, 1] \to \mathbb{R}^d$. The first variation of F at x in the direction is denoted as $\dot\gamma(0)$,

Then we can define the gradient using the first variation as follows:

Definition 8.3.2. *Given a function $F : \mathbb{R}^d \to \mathbb{R}$, its gradient is defined as*

$$\frac{d}{dt} F(\gamma(t))\Big|_{t=0} = \langle \nabla F(x), \dot\gamma(0) \rangle$$

, where $\dot\gamma(0)$ we mean the velocity of the curve γ at time 0.

Given $\mathcal{E}(\mu)$, where $\mathcal{E} : \rho_2(\mathbb{R}^d) \to (-\infty, \infty]$. We provide some examples of $\mathcal{E}$.

1. $\mathcal{E}_1(\mu) = \int V(x)d\mu(x)$ (Potential Energy)
2. $\mathcal{E}_2(\mu) = \int\int K(x,y)d\mu(x)d\mu(y)$ (Interaction Energy)
3. $\mathcal{E}(\mu) = \mathbb{E}_{Z\sim V}\left[\left|\int \sigma(x,z)d\mu(x) - f(z)\right|^2\right] = \mathcal{E}_1(\mu) + \mathcal{E}_2(\mu)$ (Squared loss for shallow Neural Network training)
   where σ is RELU (non-linearity).
4. $\mathcal{E}(\mu) = KL(\mu||\mu_0) = \begin{cases} \int \log\left(\frac{d\mu}{d\mu_0}d\mu\right) & \text{if } \mu \ll \mu_0 \\ \infty & \text{otherwise.} \end{cases}$

Complete first variation:

Suppose we are given a curve in $\rho_2(\mathbb{R}^d)$
$\partial_t\mu_t + div(\mu_t\nabla\phi_t) = 0$

$$\begin{aligned}\frac{d}{dt}\mathcal{E}(\mu_t)\Big|_{t=0} = \frac{d}{dt}\int V(x)d\mu_t(x)\Big|_{t=0} &= \int \nabla V(x)\nabla\phi_t(x)d\mu_t(x)\Big|_{t=0} \\ &= \int \nabla V(x)\nabla\phi_0(x)d\mu_0(x) = \langle \nabla V, \nabla\phi_0\rangle_{\mu_0}\end{aligned}$$

, where $\nabla\phi_0$ is the velocity of any curve.

Hence we get

$$\nabla_{W_2}\mathcal{E}(\mu_0) = \nabla V$$

Question: What is $\nabla_{W_2}(\mathcal{E}_2(\mu_0))$?

$$\begin{aligned}\frac{d}{dt}\mathcal{E}_2(\mu_t)\Big|_{t=0} &= \frac{d}{dt}\int\int k(x,y)d\mu_t(x)d\mu_t(y)\Big|_{t=0} \\ &= \int\left(\frac{d}{dt}\int k(x,y)d\mu_t(x)\right)d\mu_0(y)\Big|_{t=0} + \int\left(\frac{d}{dt}\int k(x,y)d\mu_t(y)\right)d\mu_0(x)\Big|_{t=0} \\ &= \int\left(\int \nabla_1 k(x,y)\nabla\phi_0(x)d\mu_0(x)\right)d\mu_0(y) + \int\int \nabla_2 k(x,y)\nabla\phi_0(y)d\mu_0(y)d\mu_0(y) \\ &= \int\left(\int \nabla_1 k(x,y)d\mu_0(y)\right)\nabla\phi_0(x)d\mu_0(x) + \int\left(\int \nabla_2 k(y,x)d\mu_0(y)\right)\nabla\phi_0(x)d\mu_0(x) \\ &= \int\left(\int (\nabla_1 k(x,y) + \nabla_2 k(y,x))d\mu_0(y)\right)\nabla\phi_0(x)d\mu_0(x) \\ &= \left\langle \int (\nabla_1 k(x,y) + \nabla_2 k(y,x))d\mu_0(y), \nabla\phi_0 \right\rangle_{\mu_0}\end{aligned}$$

, where $\nabla\phi_0$ is the velocity of any curve.

Hence we get

$$\nabla_{W_2}\mathcal{E}_2(\mu_0) = \int (\nabla_1 k(x,y) + \nabla_2 k(y,x))d\mu_0(y)$$

8.4 A remark after Lecture 8

At the end of section 8.1 we made the following claim: $\vec{V}_t(\tilde{x}) = \nabla\phi_t(\tilde{x})$ for some scalar function ϕ_t. Recall that $\vec{V}_t$ is the vector field in the continuity equation associated to the geodesic connecting μ_0 and μ_1 via $T_{t\sharp}\mu_0$, where

$$T_t(x) = (1-t)x + tT(x),$$

for T the Monge map between μ_0 and μ_1.

Let's prove this claim. First, by what we discussed in class we can take

$$\vec{V}_t(\tilde{x}) = T(T_t^{-1}(\tilde{x})) - T_t^{-1}(\tilde{x}).$$

From the definition of T_t we see that

$$\tilde{x} = (1-t)T_t^{-1}(\tilde{x}) + tT(T_t^{-1}(\tilde{x})),$$

and thus we can write

$$\vec{V}_t(\tilde{x}) = \frac{1}{t}(\tilde{x} - T_t^{-1}(\tilde{x})).$$

Now, by Brenier's theorem we know that $T = \nabla\varphi$ for φ a convex function. If we define

$$\varphi_t(x) := (1-t)\frac{|x|^2}{2} + t\varphi(x),$$

we see that

$$\nabla\varphi_t(x) = T_t(x).$$

We have the following equivalences thanks to convexity:

$$x = T_t^{-1}(\tilde{x}) \iff \tilde{x} = T_t(x) \iff \tilde{x} = \nabla\varphi_t(x) \iff x = \nabla\varphi_t^*(\tilde{x}), \tag{47}$$

where φ_t^* is the Fenchel dual of φ_t. In particular $T_t^{-1}(\tilde{x}) = \nabla\varphi_t^*(\tilde{x})$. This means that

$$\vec{V}_t(\tilde{x}) = \nabla\phi_t(\tilde{x}),$$

where

$$\phi_t(\tilde{x}) := \frac{1}{t}\left(\frac{1}{2}|\tilde{x}|^2 - \varphi_t^*(\tilde{x})\right).$$

9 Lecture 9: 11/02/2023

Scribes: Nursultan Azhimuratov and Alejandro Calle-Saldarriaga

At the beginning of the class, Nicolas made these two comments which are related to the logistics of the course.

Comment 1: was about the final project. He said that if any student has questions, they can email him or meet with him to discuss the final project and get some advice on an interesting topic related to them.

Comment 2: was about the class syllabus and where we are in it. He mentioned that everything is going according to plan, following the syllabus, and that this class will cover all the topics listed in the syllabus

9.1 Convexity

First $\mathcal{E} : \rho_2(\mathbb{R}^d) \to \mathbb{R} \cup \{\infty\}$, where the output is just Real numbers or potentially infinity. That is, $\mathcal{E}$ takes a measure with finite second moment and outputs the 'energy' of that measure. Also, $\nabla_{W_2}\mathcal{E}(\mu)(\cdot)$ is a vector field such that for any $x \in \mathbb{R}^n$, and

$$\frac{d}{dt}\mathcal{E}(\mu_t)\bigg|_{t=0} = \left\langle \nabla_{W_2}\mathcal{E}(\mu_0), \dot{\vec{v_0}} \right\rangle_{\mu_0}$$

We consider the mapping $t \mapsto \mu_t$ where $t \in (0,1)$, which is the change of Energy along a certain trajectory that crosses a given measure of μ_0 at the time t_0, given by the continuity equation

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

We like to consider vector fields $\overrightarrow{V} = \nabla\phi$, that is, gradients of scalar functions.

Some common energies we might consider are

1. $\mathcal{E}_1(\mu) = \int V(x)d\mu(x)$ usually called potential energy. $\nabla_{W_2}\mathcal{E}_1(\mu) = \nabla V$. The main idea is that, regardless of what your μ is, the vector field remains the same, that is, like in usual calculus, we have constant 'derivatives' for linear functions.

2. $\mathcal{E}_2 = \int\int K(x,y)d\mu(x)d\mu(y)$ usually called the interaction energy, where K is a Kernel. A choice of Kernel could be a such that $K(x,y) = \phi(|x-y|)$

$$\begin{aligned}\nabla_{W_2}\mathcal{E}_2(\mu)(\cdot) &= \int \nabla_1 K(\cdot,y)d\mu(y) + \int \nabla_2 K(y,\cdot)d\mu(y)\\ &= 2\int \nabla_1 K(\cdot,y)d\mu(y)\end{aligned}$$

   when Kernel is symmetric, like $K(x,y) = K(y,x)$ for any $x, y \in \mathbb{R}^d$

3. $\mathcal{E}_3(\mu) = E_{Z\sim V}\left[\left(\int \sigma(\langle z, 0\rangle)d\mu(\theta) - f(z)\right)^2\right]$ Squared loss for shallow Neural Network training

$$\nabla_{W_2}\mathcal{E}_3(\mu) = \nabla_{W_2}\mathcal{E}_2(\mu) + \nabla_{W_2}\mathcal{E}_1(\mu)$$

4. $\mathcal{E}_4(\mu) = KL(\mu||\tilde{\mu})$ (Internal Energy), here we suppose $\tilde{\mu} \propto e^{-V(x)}dx$ where a possible choice is $V(x) = \frac{|x|^2}{2}$ (Gaussian measure).

$$\mathcal{E}_4(\mu) = \int V(x)d\mu(x) + \int \log\left(\frac{d\mu}{dx}\right)\frac{d\mu}{dx}dx$$

   (here $\int \log\left(\frac{d\mu}{dx}\right)\frac{d\mu}{dx}dx =:$F) Assuming that μ has a density w.r.t Lebesque measure

$$\nabla_{W_2}\mathcal{E}_4(\mu) = \nabla_{W_2}\mathcal{E}_1(\mu) + \nabla_{W_2}F(\mu)$$

Let's return to our focus to the continuity equation

$$\partial_t\mu_t + div(\mu_t\nabla\phi_t) = 0$$

If we suppose $d\mu_t = \rho_t dx$ (μ_t is absolutely continuous with relation to the Lebesegue measure), we have

$$\frac{\partial}{\partial t}\rho_t + div(\rho_t\nabla\phi_t) = 0$$

and we can write

$$\frac{d}{dt}F(\mu_t)|_{t=0} \quad = \quad \frac{d}{dt}\int \log(\rho_t)\rho_t dx = \int \frac{d}{dt}\log(\rho_t)\rho_t dx$$

here $a = a_t$ and $\frac{d}{dt}\log(a_t)a_t = \frac{d}{da}(\log(a)a)\frac{da}{dt} = (1+\log(a))\frac{da}{dt}$ so

$$\begin{aligned}\frac{d}{dt}F(\mu_t)|_{t=0} &= \int (1+\log(\rho_t))\frac{d}{dt}\rho_t(x)dx\\ &= \int \frac{d}{dt}\rho_t(x)dx + \int \log(\rho_t)\frac{d}{dt}\rho_t dx\\ &= -\int \log(\rho_t)div(\rho_t\nabla\phi_t)dx\\ &= \int (\nabla\log(\rho_t)\nabla\phi_t)\rho_t dx\\ &= \int \nabla\log(\rho_t)\nabla\phi_t d\mu_t(x) = \langle\log(\rho_t), \nabla\phi_t\rangle_{\mu_t}\end{aligned}$$

that is, we have the relationship $\nabla_{W_2} F(\mu) = \nabla log(\rho)$ where $d\mu = \rho dx$

Coming back to the example 3

$$\begin{aligned}
\mathcal{E}_3(\mu) &= E_{Z\sim V_0}\left[\left(\int \sigma(\langle\theta, z\rangle)d\mu(\theta) - f(z)\right)^2\right] \\
&= E_{Z\sim V_0}\left[\int\int \sigma(\langle\theta, z\rangle)\sigma(\langle\tilde{\theta}, z\rangle)d\mu(\theta)d\mu(\tilde{\theta}) - 2f(z)\int \sigma(\langle\theta, z\rangle)d\mu(\theta) - f(z)^2\right] \\
&= \int\int E_{Z\sim V_0}\left[\sigma(\langle\theta, z\rangle)\sigma(\langle\tilde{\theta}, z\rangle)\right]d\mu(\theta)d\mu(\tilde{\theta}) - 2\int E_{Z\sim V_0}\left[f(z)\sigma(\langle\theta, z\rangle)d\mu(\theta) + (E_{Z\sim V_0} f(z))^2\right]
\end{aligned}$$

Question: Given $\mathcal{E}(\mu)$ and how would we compute $\nabla_{W_2}\mathcal{E}(\mu)$?

Step 1 Compute the **First Variation** of $\mathcal{E}$

Definition 9.1.1. *we say that function $\mu(x) : \mathbb{R}^d \to \mathbb{R}$ is the **First Variation** of $\mathcal{E}$ at μ if the following holds:*

$$\frac{d}{dt}\Big|_{t=0} \mathcal{E}(\mu + t(\tilde{\mu} - \mu)) = \int \mu(x)d(\tilde{\mu} - \mu)(x)$$

for any other $\tilde{\mu}$

Example: $\mathcal{E}(\mu) = \mathcal{E}_1(\mu)$

$$\frac{d}{dt}\Big|_{t=0} \mathcal{E}_1(\mu + t(\tilde{\mu} - \mu)) = \frac{d}{dt}\Big|_{t=0}\int V(x)d(\mu + t(\tilde{\mu} - \mu)) = \int V(x)d(\tilde{\mu} - \mu)(x)$$

Step 2 $\nabla_{W_2}\mathcal{E}(\mu) = \nabla\frac{\partial\mathcal{E}}{\partial\mu}(\mu)$. That is, you pick an energy and compute Wasserstein gradients for that energy. Note that we are using some sort of calculus-like computations. Calculus in this space, interpreting Wasserstein distances as Riemannian distances in the space of probability measures with finite second moment, is called Otto Calculus.

Definition 9.1.2. *We denote by $\frac{\partial\mathcal{E}}{\partial\mu}(\mu)$ the first variation of $\mathcal{E}$ at μ*

9.2 Wasserstein Gradient flows

In this section we are going to introduce Wasserstein Gradient flows. It is useful to build up from our intuitions in Euclidean space. Let's write out the equations that describe euclidean gradient flows. Let $f : \mathbb{R}^d \to \mathbb{R}$. Then a gradient flow is

$$\begin{cases}\dot{\gamma}(t) = -\nabla f(\gamma(t)) \\ \gamma(0) = x_0\end{cases}$$

That is, a curve $\gamma : [0, \infty) \to \mathbb{R}$ such that $\gamma(0) = x_0$ and at later times, it's velocity is $\nabla f(\gamma_t)$. We also have that

$$\frac{d}{dt}f(\gamma_t) = \langle\nabla f(\gamma_t), \gamma_t'\rangle = -\langle\nabla f(\gamma_t), \nabla f(\gamma_t)'\rangle$$

i.e., energy dissipates while moving along the curve.

Now, keeping this brief detour in mind, let's think about the gradient flow of $\mathcal{E}$ in the space W_2. Akin to the euclidean case, it is gonna be a curve, but here the curve maps $t \mapsto \mu_t$, and at all times after the start, our velocity is going to be $\nabla W_2\mathcal{E}(\mu_t)$. Also, since it is a curve in W_2, it must satisfy the continuity equation we have discussed on earlier lectures. Therefore the Wasserstein gradient flow is described by

$$\begin{cases}\partial_t\mu_t - div\left((\mu_t + \nabla\frac{\delta\mathcal{E}}{\delta\mu}(\mu_t)\right) = 0 \\ \mu_0 = \mu_0\end{cases}$$

and we have an analogous "energy dissipation" behaviour

$$\begin{aligned}\frac{d}{dt}\mathcal{E}(\mu_t) &= \langle \nabla_{W_2}\mathcal{E}(\mu_t), -\nabla_{W_2}\mathcal{E}(\mu_t)\rangle_{\mu_t} \\ &= -\langle \nabla_{W_2}\mathcal{E}(\mu_t), \nabla_{W_2}\mathcal{E}(\mu_t)\rangle_{\mu_t} \\ &= -\int \left|\nabla \frac{\delta}{\delta\mu}\mathcal{E}(\mu_t)\right|^2 d\mu_t\end{aligned}$$

Now, let's see how some of the gradient flows of the previously considered energies look like

Example. Let $\mathcal{E}_1(\mu) = \int V(x)d\mu(x)$. Then the gradient flows are governed by

$$\begin{cases}\partial_t\mu_t - div\,(\mu_t\nabla V) = 0 \\ \mu_0 = \mu_0\end{cases}$$

Let $\mu_0 = \frac{1}{n}\sum_{i=1}^n \delta_{X_0^i}$, the empirical measure. So how does the gradient flow make our μ_t look like? We would think it looks like $\mu_t = \frac{1}{n}\sum_{i=1}\delta_{X_t^i}$ where

$$\begin{cases}\dot{X}_t^i = -\nabla V(X_t^i) \\ X_0^i = X_0^i\end{cases}$$

We can think about this problem using a particle analogy. We start with some particles arranged in a certain way (described by our initial empirical measure), and we want to track our particles via μ_t, that is, at any time t we will have another empirical measure μ_t which describe the particle configuration at time t. In general, you can consider the random variable $\mathcal{X}_0 \sim \mu_0$ and

$$\begin{cases}\dot{X}_t = -\nabla V(X_t) \\ X_0 = x_0\end{cases}$$

and we have that the law of X_t is μ_t. Note that in this example, every particle is on its own and don't care about interactions with other particles. Another more exciting example would then be:

Example. Let $\mathcal{E}_2(\mu) = \int\int K(x,y)d\mu(x)d\mu(y)$, with $K(x,y) = K(y,x) \qquad \forall x, \forall y$ (symmetric), and it is relatively smooth such that we don't loose mass (analogy: a flock of birds. K relates to how the birds interact with each other (they want closeness to other birds, but a bit of space too). Smoothness here will mean that birds do not collide and get on top each other so that we loose some birds in that process). Let's write the gradient flow

$$\begin{cases}\partial_t\mu_t - 2div\left(\mu_t\nabla_X \int K(x,y)d\mu(y)\right) = 0 \\ \mu_0 = \mu_0\end{cases}$$

Once again, let's consider the empirical measure $\mu_0 = \frac{1}{n}\sum_{i=1}^n \delta_{X_0^i}$, and we write

$$\dot{X}_t^i = -\frac{2}{n}\sum_{j=1}^n \nabla_1 K(X_t^i, X_t^j)$$

where ∇_1 means that we compute the gradient with relation to the first argument in the expression, and we claim then that $\mu_t = \frac{1}{n}\sum_{i=1}^n \delta_{X_t^i}$.

Let's take a step back and forget about gradient flows for a minute. Let's say, if we were to do gradient descent in Wasserstein space, would we arrive at something similar to the gradient flows we have considered?

Example. Consider a shallow Neural Network with n neurons

$$z \mapsto \frac{1}{n}\sum_{i=1}^n \sigma(\langle\theta^i, t\rangle)$$

where σ is some activation function. Also, consider the loss function

$$l(\theta^1, \dots, \theta^n) = \mathbb{E}_{z\sim\nu_0}\left[\left(\frac{1}{n}\sum_{i=1}\sigma(\langle\theta^i, z\rangle) - f(z)\right)^2\right]$$

then we can write the gradient descent equations:

$$\begin{cases}\dot{\overrightarrow{\theta}} = -\nabla l(\overrightarrow{\theta}) \\ \overrightarrow{\theta}_0 = \overrightarrow{\theta}_0\end{cases}$$

This is just plain old gradient descent, what we would have done if we did not know about Wasserstein Gradient flows. We have that $\overrightarrow{\theta}_t = (\theta_t^1, \dots, \theta_t^n)$. We could ask ourselves: how does $\overrightarrow{\theta}_t$ evolves in t? Does this have something to do with the gradient flows we have considered earlier? The answer is yes (at least for this shallow network): this $\overrightarrow{\theta}_t$ evolves over time exactly as described by the Wasserstein gradient flows. So what we have learned provides us with new tools and new perspectives to study and reason about usual problems.

Remark 9.2.1. *It is interesting to think which $\mathcal{E}$ we could consider are convex. Intuitively, the more interesting $\mathcal{E}$ are not convex. Returning to the flock analogy, it is likely that a flock configuration will stay at local minima. You would have to do something extreme for the flock to rearrange and reconfigure itself in a "better" configuration they are also comfortable on (for example, screaming at them).*

9.3 Revisiting convexity

The natural direction we might take here is to ask about convexity in Wasserstein spaces. As before, we will first take a detour and consider the Euclidean case and try to see how to connect those ideas to our Wasserstein spaces. Remember that in the Euclidean cases, convexity is related to Hessians. Let $f : \mathbb{R}^d \to \mathbb{R}$.

$$\begin{aligned}\frac{d^2}{dt^2}f(\gamma_t) &= \frac{d}{dt}\left(\langle\nabla f(\gamma_t), \dot{\gamma}_t\right) \\ &= \langle\frac{d}{dt}\nabla f(\gamma_t), \dot{\gamma}_t\rangle + \langle\nabla f(\gamma_t), \ddot{\gamma}_t\rangle \\ &= \langle D^2 f(\gamma_t)\dot{\gamma}_t, \dot{\gamma}_t\rangle + \langle\nabla f(\gamma_t), \ddot{\gamma}_t\rangle \\ &= \langle D^2 f(\gamma_t)\dot{\gamma}_t, \dot{\gamma}_t\rangle\end{aligned}$$

Assuming $\ddot{\gamma}_t = 0$. When we want to characterize Hessians, we are looking at some sort of geodesic trajectory. Based on this, we can define

Definition 9.3.1. *(λ-convexity) We say that f is λ-convex if*

$$\frac{d^2}{dt^2}f(\gamma_t) \geq \lambda$$

for all geodesics γ_t with $|\dot{\gamma}(t)| = 1$.

Definition 9.3.2. *(λ-convexity, alternative definition) We say that f is λ-convex everywhere if $\forall x, y \in \mathbb{R}^d$ we have that*

$$\frac{d^2}{dt^2}f(\gamma_t) \geq d^2(x, y)\lambda$$

where $\gamma_t : [0, 1] \to \mathbb{R}^d$ is the minmizing geodesic connecting x and y (in Euclidean space it is $\gamma_t = (1 - t)x + ty$).

This formulation is nice because it is clear how we can generalize it to Wasserstein space. Then we can define

Definition 9.3.3. *We say $\mathcal{E} : \rho_2(\mathbb{R}^d) \to \mathbb{R} \cup \{\infty\}$ is λ-geodesically (or displacement) convex if for any $\mu_0, \mu_1 \in \rho_2(\mathbb{R}^d)$ we have*

$$\frac{d^2}{dt^2}\mathcal{E}(\mu_t) \geq W_2^2(\mu_0, \mu_1)\lambda$$

where μ_t is a minimizing geodesic between μ_0 and μ_1.

Example. Lets consider $\mathcal{E}_1(\mu) = \int V(x)d\mu(x)$. If the potential V is λ-convex in the euclidean space, then $\mathcal{E}_1$ is λ-convex in the Wasserstein space.

Example. Now, consider $\mathcal{E}_2 = \int\int K(x,y)d\mu(x)d\mu(y)$. This case is a bit more complicated. If $K(x,y) = \phi(|x-y|)$, that is, K is a function of some notion of distance between x and y, then ϕ is λ-convex in the Euclidean sense if and only if $\mathcal{E}_2$ is λ-convex in the Wasserstein sense.

Remark 9.3.1. *In general, we don't expect convexity for neural networks. In our trivial, shallow case we have convexity, and there are also some cases in which you can prove convexity but it is not trivial.*

Example. $\int log(\rho)\rho dx$ is 0-convex (i.e. just convex).

10 Lecture 10: 11/09/2023

Scribes: Gokcan Tatli and Joe Shenouda

We start with recalling geodesic convexity definition. λ-geodesic convexity of $\mathcal{E}$ (Definition 9.3.3) is also equivalent to say that $t \in [0,1] \mapsto \underbrace{\mathcal{E}(\mu_t)}_{g(t)}$ is $(\lambda W_2^2(\mu_0,\mu_1))$ - convex, which is also equivalent to say $g(t) - \frac{\lambda W_2^2(\mu_0,\mu_1)t^2}{2}$ is convex.

Remark 10.0.1. *(Displacement Convexity vs Convexity in the linear interpolation sense)*
$(\mu_0, \mu_1) \longrightarrow (1-t)\mu_0 + t\mu_1$ (Convex combination in the linear sense)
$\Rightarrow t \in [0,1] \mapsto \mathcal{E}((1-t)\mu_0 + t\mu_1)$
$T_{t\#}\mu_0$ where $T_t^* = (1-t)x + tT^\star(x)$ where T^* is the Brenier map between μ_0 and μ_1 is convex
$\Rightarrow t \in [0,1] \mapsto \mathcal{E}(T_{t\#}, \mu_0)$ is convex

Example 1 $\mathcal{E}_2(\mu) = \int\int K(x,y)d\mu(x)d\mu(y)$
Suppose $K(.,.)$ is a kernel $\Rightarrow \mathcal{E}_2$ is convex in the linear interpolation sense
$K(\theta, \hat{\theta}) = \mathbb{E}_{z\sim\nu_0}\left[\sigma(\langle\theta, z\rangle)\sigma(\langle\hat{\theta}, z\rangle)\right]$

$\mathcal{E}_1(\mu) = \int V(x)d\mu(x)$ is convex in the linear interpolation sense.
Unfortunately, $\mathcal{E}_2$ is in general not convex in the Wasserstein sense.
Cases where we know we have displacement convexity
$K(x,y) = \phi(x-y)$ and $\phi : \mathbb{R} \to \mathbb{R}$ is convex.
x and y should not be too close or too far.
Neural network loss is like the summation of $\mathcal{E}_1 + \mathcal{E}_2$

10.1 Other Examples of Displacement Convexity

Examples. $\mathcal{E}_1(\mu) = \int V(x)d\mu(x)$ (where V is λ-convex e.g. $V(x) = \frac{|x|^2}{2}$)

Proof.

$$\frac{d^2}{dt^2}\mathcal{E}_1(\mu_t) = \frac{d}{dt}\left(\underbrace{\frac{d}{dt}\mathcal{E}_1(\mu_t)}_{\langle \nabla_{W_2}\mathcal{E}_1(\mu_t), \vec{V_t}\rangle_{\mu_t}}\right)$$

$$\mu_t = T_{t\#}\mu_0 = \frac{d}{dt}\left(\int \nabla_{W_2}\mathcal{E}_1(\mu_t).\vec{V_t}d\mu_t\right)$$

$$= \frac{d}{dt}\left(\frac{d}{dt}\int V(x)d\mu_t(x)\right)$$

$$= \frac{d}{dt}\left(\frac{d}{dt}\int V(T_t(x))d\mu_0(x)\right)$$

$$= \frac{d}{dt}\left(\frac{d}{dt}\int \nabla V(T_t(x)).\frac{d}{dt}T_t(x)d\mu_0(x)\right)$$

where $T_t(x) = (1-t)x + tT^*(x)$ and $\frac{d}{dt}T_t(x) = T^*(x) - x$

$$= \frac{d}{dt}\left(\frac{d}{dt}\int \nabla V(T_t(x)).(T'(x) - x)d\mu_0(x)\right)$$

$$= \int \langle \nabla^2 V(T_t(x))\frac{d}{dt}T_t(x), T^*(x) - x\rangle d\mu_0(x)$$

$$= \int \langle \nabla^2 V(T_t(x))(T'(x) - x), (T'(x) - x)\rangle d\mu_0(x)$$

$$\geq \lambda \int |T'(x) - x|^2 d\mu_0(x) = \lambda W_2^2(\mu_0, \mu_1)$$

□

Example 2 $\mathcal{E}_4(\mu) = \int \log(\rho)\rho dx$, $d\mu = \rho dx$

$$\frac{d^2}{dt^2}\mathcal{E}_4(\mu_t) = \frac{d}{dt}\left(\frac{d}{dt}\int \log(\rho_t)\rho_t dx\right)$$

$d\mu_t = \rho_t dx$

$$\begin{cases} \underbrace{\partial_t \mu_t}_{\frac{\partial}{\partial t}\rho_t} + div\left(\underbrace{\mu_t}_{\rho_t}\nabla\phi_t\right) = 0 \\ \partial_t\phi_t + \frac{1}{2}|\nabla\phi_t|^2 = 0 \end{cases}$$

$$\frac{d}{dt}\left(\int (1 + \log(\rho_t))\frac{\partial}{\partial t}\rho_t dx\right) = \frac{d}{dt}\left(\int \log(\rho_t)\frac{\partial}{\partial t}\rho_t dx\right)$$

$\int \frac{\partial}{\partial t}\rho_t(x)dx = \frac{\partial}{\partial t}\underbrace{\int \rho_t(x)dx}_{1}$

$$\begin{aligned}
\frac{d}{dt}\left(-\int \log(\rho_t) div(\rho_t \nabla \phi_t) dx\right) &= \frac{d}{dt}\left(\int \nabla \log(\rho_t).\nabla\phi_t \rho_t dx\right) \\
&= \frac{d}{dt}\int \nabla\rho_t.\nabla\phi_t dx \\
&= \int \left(\frac{d}{dt}\nabla\rho_t.\nabla\phi_t + \nabla\rho_t.\frac{d}{dt}\nabla\phi_t\right) dx \\
&= \int \left(\nabla\frac{d}{dt}\rho_t.\nabla\phi_t + \nabla\rho_t.\nabla\frac{d}{dt}\phi_t\right) dx \\
&= \int -\nabla(div(\rho_t\nabla\phi_t)).\nabla\phi_t - \frac{1}{2}\nabla\rho_t.\nabla(|\nabla\rho_t|^2) dx \\
&= \int -\rho_t\nabla\phi_t.\nabla div(\nabla\phi_t) + \frac{1}{2}\rho_t div(\nabla|\nabla\phi_t|^2) \\
&= \int \left(\frac{1}{2}div(\nabla|\nabla\phi_t|^2) - \nabla\phi_t.\nabla div(\nabla\phi_t)\right)\rho_t dx \\
&= \int \left(\frac{1}{2}\Delta(|\nabla\phi_t|^2) - \nabla\phi_t.\nabla div(\nabla\phi_t)\right)\rho_t dx
\end{aligned}$$

$\Delta = div \circ \nabla$
By Bochner's formula, $\|D^2\phi^{(*)}\|_F^2 = \frac{1}{2}\Delta(|\nabla\phi|^2) - \nabla\phi.\nabla\phi(\nabla\phi)$

Corollary 10.1.1. *$\mathcal{E}(\mu) = KL(\mu\|\tilde{\mu})$ if $\tilde{\mu} = e^{-V}dx$*
$\mathcal{E}(\mu) = \mathcal{E}_1(\mu) + \mathcal{E}_4(\mu)$ then $KL(.\|\tilde{\mu})$ is λ - geodesically convex if V is λ - geodesically convex

10.2 Functional Inequalities

Functional inequalities include the log-Sobolev inequalities and Talagrand's transport inequality. Consider the functional $\mathcal{E}(\mu)$ and let

$$\mu^* \in \operatorname*{argmin}_{\mu} \mathcal{E}(\mu) \tag{48}$$

Theorem 10.2.1. *If $\mathcal{E}$ is λ-geodesically convex with $\lambda > 0$ (e.g. $\mathcal{E}(\mu) = KL(\cdot||\tilde{\mu})$) then*

$$\mathcal{E}(\mu) - \mathcal{E}(\mu^*) \leq C\|\nabla_{W_2}\mathcal{E}(\mu)\|_\mu^2 \tag{49}$$
$$W_2^2(\mu,\mu^*) \leq C(\mathcal{E}(\mu) - \mathcal{E}(\mu^*)) \tag{50}$$

We can use this fact to now prove a rate of convergence for Wasserstein gradient flows on geodesically convex functionals. Let μ_t follow a gradient flow on a geodesically convex functional $\mathcal{E}$ starting at μ_0. Then it must satisfy the following continuity equations.

$$\begin{cases} \partial_t\mu_t - \operatorname{div}(\mu_t\nabla_{W_2}\mathcal{E}(\mu_t)) &= 0 \\ \nabla_{W_2}\mathcal{E}(\mu_t) = \nabla\frac{\delta\mathcal{E}}{\delta\mu}(\mu_t) \end{cases}$$

Now at any time t we have

$$\mathcal{E}(\mu_t) = \mathcal{E}(\mu_0) + \int_0^t \frac{d}{ds}\mathcal{E}(\mu_s) ds \tag{51}$$
$$= \mathcal{E}(\mu_0) + \int_0^t \langle \nabla_{W_2}\mathcal{E}(\mu_s), -\nabla_{W_2}\mathcal{E}(\mu_s)\rangle_{\mu_s} \tag{52}$$

This follows from the discussion in Section 9.2. Now by the inequalities described in the theorem we can conclude the following

$$\begin{aligned}\mathcal{E}(\mu_0) + \int_0^t \langle \nabla_{W_2}\mathcal{E}(\mu_s), -\nabla_{W_2}\mathcal{E}(\mu_s)\rangle &= \mathcal{E}(\mu_0) - \int_0^t \|\nabla_{W_2}\mathcal{E}(\mu_s)\|^2_{\mu_s} \\ &\leq \mathcal{E}(\mu_0) - \frac{1}{c}\int_0^t (\mathcal{E}(\mu_s) - \mathcal{E}(\mu^*))ds\end{aligned}$$

Thus we have

$$\mathcal{E}(\mu_t) - \mathcal{E}(\mu^*) \leq \mathcal{E}(\mu_0) - \mathcal{E}(\mu^*) - \frac{1}{c}\int_0^t (\mathcal{E}(\mu_s) - \mathcal{E}(\mu^*))ds$$

If we define

$$\begin{aligned}g_t &= \mathcal{E}(\mu_t) - \mathcal{E}(\mu^*) \\ g_0 &= \mathcal{E}(\mu_0) - \mathcal{E}(\mu^*) \\ g_s &= \mathcal{E}(\mu_s) - \mathcal{E}(\mu^*)\end{aligned}$$

we can rewritten the inequality above as

$$g_t \leq g_0 - \frac{1}{c}\int_0^t g_s ds$$

Now by Grönwall's inequality we have

$$g_t \leq g_0 e^{-\frac{1}{c}t} \tag{53}$$

10.3 Langevin Dynamics and Diffusion Models

We can interpret these as an interacting particle system or stochastic particles. Consider the stochastic differential equation (SDE)

$$dX_t = -\nabla V(X_t)dt + \sqrt{2}dB_t \tag{54}$$

where $X_0 \sim \mu_0$. We are interested in understanding the law of μ_t in order to sample from it. This is the Langevin SDE and can be interpreted as gradient descent + noise on $\mathcal{E}(\mu) = \mathrm{KL}(\mu||e^{-V})$ with continuity equations

$$\begin{cases}\partial_t \mu_t - \mathrm{div}(\mu_t \nabla_{W_2}\mathcal{E}(\mu_t)) &= 0 \\ \mu_0 = \mu_0\end{cases}$$

10.3.1 One Version of Diffusion Models

Recall we are interested in μ_t. If $V(x) = \frac{|x|^2}{2}$ then the Langevin SDE reduces to

$$dX_t = -X_t dt + \sqrt{2}dB_t$$

Also define

$$Y_t := X_{T-t} \quad t \in [0, T] \tag{55}$$

Thus the law of X_t is the Wasserstein gradient of $\mathrm{KL}(\cdot||\mathcal{N}(0, I))$. Then by the continuity equations

$$\begin{cases}\partial_t \rho_t - \Delta\rho_t - \mathrm{div}(\rho_t(X)) &= 0 \\ \frac{d}{ds}\int \phi(x)d\mu_s &= -\int \nabla\phi(x)(-\nabla V(x) - \nabla\log(\rho_s))\rho_s dx\end{cases}$$

If we let V_t be the distribution of Y_t then

$$\begin{aligned}\frac{d}{dt}\int \phi(y)dV_t(y) &= \frac{d}{dt}\int \phi(x)d\mu_{T-t}(x)\\ &= \int \nabla\phi(x)\cdot(\nabla V(x)+\nabla\log(\rho_{T-t}))\rho_{T-t}dx\\ &= \int \nabla\phi(x)\cdot(\nabla V(x)+2\nabla\log(\rho_{T-t})-\nabla\log(\rho_{T-t}))dx\end{aligned}$$

Therefore

$$dV_t = (\nabla V(V_t)+2\nabla\log(\rho_{T-t}(V_t)))dt+\sqrt{2}dB_t \quad t\in[0,T]$$

Problem: We do not have μ_0!! The next lecture discusses how to overcome this challenge.

11 Lecture 11: 11/16/2023

Scribes: Daniel Ye, Yaling Hong and Shuangyu Wang

Firstly, the lecture focuses on a diffusion process aimed at transforming an initial distribution μ_0 into a standard normal distribution $N(0,1)$. This process is described through a continuous-time stochastic differential equation (SDE) and involves concepts from optimal transport theory.

11.1 Diffusion Model

11.1.1 Forward Process

The initial data x_0 follows the distribution μ_0. We define a continuous-time diffusion process $\{X_t\}_{t=0}^T$, where X_t gradually approaches $N(0,1)$. This process can be described by the following SDE:

$$dX_t = -\frac{1}{2}\nabla\log\mu_t(X_t)\,dt + dB_t, \quad X_0\sim\mu_0 \tag{56}$$

Here, B_t is a standard Brownian motion, and μ_t is the marginal distribution of X_t.

11.1.2 Gradient Flow in Optimal Transport

This process can be viewed as the gradient flow of the KL divergence $D_{KL}(\mu_t||N(0,1))$. The "energy" E represents the KL divergence between μ_t and $N(0,1)$.

11.1.3 Reverse Process

To recover μ_0 from $N(0,1)$, we consider the time-reversed process. Let $Y_t = X_{T-t}$ and consider $Y_0\sim N(0,1)$. The SDE for the reverse process is described as:

$$dY_t = (Y_t+\nabla\log\mu_{T-t}(Y_t))\,dt+\sqrt{2}dB_t, \quad Y_0\sim N(0,1) \tag{57}$$

Here, $\nabla\log\mu_{T-t}$ is the key "denoising" term, helping us recover the original data from noise.

11.1.4 Assessment of the Process

Ultimately, we wish to assess how close ν_t (the distribution in the reverse process) is to the original distribution μ_0. This can be done by comparing the Wasserstein distance or KL divergence between the two distributions.

11.2 Schrödinger Bridge

11.2.1 Setup

In the context of optimal transport theory, the Schrödinger Bridge Problem seeks a probability measure within the space of all measures defined on $\mathbb{R}^d \times \mathbb{R}^d$:

$$p \in \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d)$$

The objective is to minimize the Kullback-Leibler divergence $KL(p||q)$ between measure p and a reference measure q:

$$KL(p||q) = \mathbb{E}_p \left[\log \left(\frac{dp}{dq} \right) \right] = \int \log \left(\frac{dp}{dq} \right) dp$$

Given the probability measures at initial and final times T_0 and T_1, denoted by p_0 and p_1 respectively:

$$p_0 = p(\cdot|T_0), \quad p_1 = p(\cdot|T_1)$$

We consider T_0 and T_1 as random variables with distributions p_0 and p_1:

$$T_0 \sim p_0, \quad T_1 \sim p_1$$

11.2.2 Optimization Problem Definition

We seek a probability measure p that minimizes the Kullback-Leibler divergence with respect to a reference measure q:

$$\min_{p \in \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d)} KL(p||q)$$

where $\mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d)$ denotes the set of all probability measures on $\mathbb{R}^d \times \mathbb{R}^d$.

11.2.3 Kullback-Leibler Divergence

The Kullback-Leibler divergence $KL(p||q)$ for measure p relative to measure q is given by:

$$KL(p||q) = \mathbb{E}_{P_{0t}} \left[\log \frac{dP_{0t}}{dQ_{0t}} \right]$$

It can be computed via the integral:

$$= \int \log \frac{dP_{0t}}{dQ_{0t}} dP_{0t}$$

11.2.4 Constraints

The optimization problem is subject to the marginal distributions constraints:

$$\text{s.t. } (P_{0t})_0 = \mu_0, \quad (P_{0t})_t = \mu_t$$

where $(P_{0t})_0$ and $(P_{0t})_t$ represent the marginal distributions of measure P_{0t} at times 0 and t, respectively.

11.2.5 Choice of Reference Measure

To facilitate the problem, a specific form of the reference measure Q_{0t} is chosen:

$$\text{Choose } Q_{0t} = Q_{t0}(Z_0, Q_{Z_0})$$

In this context, Z_0 and Q_{Z_0} pertain to related stochastic variables or processes.

11.3 Back to linearization

Given a sequence of probability measures $(\mu_n)_{n\in\mathbb{N}}$ and a target measure μ, we consider the Brenier map between μ_0 and μ_n, denoted T_{μ_n}, where $T_{\mu_n} = \nabla\phi_n$ and ϕ_n is a convex potential. We explore the convergence properties associated with these maps.

11.3.1 Questions of Convergence

Two main questions are considered:

1. If $T_{\mu_n} \to T_\mu$, does it follow that $\mu_n \to \mu$ in some sense?
2. If $\mu_n^{(2)} \to \mu^{(2)}$, is it true that $T_{\mu_n} \to T_\mu$?

We define a map F as follows:

$$F : X \in \mathbb{R}^d \to (T_{\mu_0}(X), T_\mu(X))$$

and consider the pushforward measure $\pi = F_\#\mu_0$ in the space of couplings $P(\mu_0, \mu)$.

The Wasserstein-2 distance W_2 between μ_0 and μ satisfies the inequality:

$$W_2(\mu_0, \mu) \leq \|T_{\mu_n} - T_\mu\|$$

Finally, we relate the squared Wasserstein distance to the squared difference of the Brenier maps integrated against the measure μ_0:

$$\int (y - z)^2 d\pi(y, z) \leq \int (T(x) - T_\mu(x))^2 d\mu_0(x)$$

11.3.2 Implications of Convergence

In this section, we discuss the convergence properties within the context of optimal transport. Specifically, we consider the sequence of measures $(\mu_n)_{n\in\mathbb{N}}$ and their corresponding Brenier maps $(T_{\mu_n})_{n\in\mathbb{N}}$.

The following implications are considered:

- Does convergence $T_{\mu_n} \to T_\mu$ imply $\mu_n \to \mu$?
- If $W_2(\mu_n, \mu) \to 0$, does this imply $\mu_n \to \mu$?

11.3.3 Tightness and Convergence

We note that the first term $\int |x - T_{\mu_n}(x)|^2 d\mu_n(x)$ is tight, suggesting that the sequence is controlled and does not diverge as $n \to \infty$.

11.3.4 Convergence Inequality

A convergence inequality is introduced, relating the squared difference of the Brenier maps to the convergence of measures:

$$\lim_{n\to\infty} \int |T_{\mu_n}(x) - T_\mu(x)|^2 d\mu_n(x) \leq \lim_{n\to\infty} W_2(\mu_n, \mu)^2$$

11.4 Gromov-Wasserstein Distances in Optimal Transport

The final part of our discussion focuses on the concept of Gromov-Wasserstein (GW) distances which serve as a measure of discrepancy between probability measures in different metric spaces.

11.4.1 Definition of Gromov-Wasserstein Distance

The GW distance generalizes the notion of Wasserstein distances to the setting of different metric spaces. Given two metric spaces $(\mathcal{X}, d_{\mathcal{X}})$ and $(\mathcal{Y}, d_{\mathcal{Y}})$, and probability measures μ on $\mathcal{X}$ and ν on $\mathcal{Y}$, the GW distance is defined as:

$$\begin{aligned} GW(\mu, \nu) &= \inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{Y}} \int_{\mathcal{X}\times\mathcal{Y}} |d_{\mathcal{X}}(x, x') - d_{\mathcal{Y}}(y, y')|^2 d\pi(x, y) d\pi(x', y') \\ &= \inf_{\pi \in \Pi(\mu,\nu)} \mathcal{E}(\pi) \end{aligned}$$

where $\Pi(\mu, \nu)$ denotes the set of all couplings between μ and ν, and $\mathcal{E}(\pi)$ represents the cost associated with a particular coupling π.

12 Lecture 12

Scribes: