Space of Probability Measures and Optimal Transport (확률측도 공간 및 최적 수송)

김지수 (Jisu KIM)

확률론 2 (Probability Theory 2), 2025 2nd semester (fall)

The lecture note is a combination of the lecture notes from Prof. Larry Wasserman's "Statistical Machine Learning" and other references.

For the optimal transport, two good references for this topic are:

Kolouri, Soheil, et al. Optimal Mass Transport: Signal processing and machine-learning applications. *IEEE Signal Processing Magazine* 34.4 (2017): 43-59.

Villani, Cedric. *Topics in optimal transportation*. No. 58. American Mathematical Soc., 2003.

As usual, you can find a wealth of information on the web.

1 Space of Probability Measures

A core topic of statistics is to compare two probability measures, or compare on a sequence of probability measures. Suppose when $X_1, \dots, X_n$ are i.i.d. from a probability measure P, then we want to see how the empirical measure $P_n := \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ is close to P, where δ_x is the Dirac measure at x, i.e., $\delta_x(A) = I(x \in A)$. Recall the Glivenko-Cantellli theorem (글리벤코-칸텔리 정리):

Theorem ([2, Theorem 2.4.9])**.** *Glivenko-Cantellli theorem (글리벤코-칸텔리 정리) Let P be a probability measure on $(\mathbb{R}, \mathcal{R})$, and $P_n := \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ be the empirical measure. Let F and F_n be distribution function for P and P_n, respectively, then*

$$\sup_x |F_n(x) - F(x)| \to 0 \text{ a.s. as } n \to \infty.$$

Hence we would like to think of a the space of probability measures, with some kind of distance d on probabilty measures so that $d(P_n, P) \to 0$ a.s. as $n \to \infty$.

The space of probability measures, or the space of measures, is not only for comparing two probability measures; it is a very useful space to represent your data, or some objects. For example, your data is a point cloud in $\mathbb{R}^d$; i.e.,

$$\mathcal{X} = \{x_1, \dots, x_n\} \subset \mathbb{R}^d.$$

Maybe an elementary way to represent this point cloud is to represent everything as a vector in $\mathbb{R}^{nd}$, or a matrix in $\mathbb{R}^{n\times d}$:

$$(x_{11}, \dots, x_{1d}, x_{21}, \dots, x_{2d}, \dots, x_{n1}, \dots, x_{nd}) \in \mathbb{R}^{nd} \qquad \text{or} \qquad \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1d} \\ x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nd} \end{pmatrix} \in \mathbb{R}^{n\times d}. \tag{1}$$

This can be treated not as an ordered list of vectors but as an empirical distribution that assigns equal mass to each point. Formally, the point cloud is embedded into the space of measures through

$$P_{\mathcal{X}} = \frac{1}{n}\sum_{i=1}^n \delta_{x_i} \qquad \text{or} \qquad \mu_{\mathcal{X}} = \sum_{i=1}^n \delta_{x_i}. \tag{2}$$

In contrast to the vector or the matrix representations in (1), this measure representation in (2) removes any dependence on the ordering of points, since the empirical measure remains unchanged under any permutation of the indices. Also, (2) enables to compare point clouds of different sizes n. Also, in many cases, point cloud is some

(possibly random) sample from a probability distribution P; the probability measure representation enables to naturally compare the point cloud and the probability distribution via $P_{\mathcal{X}}$ and P.

If you are familiar with Topological Data Analysis, probability measure space, in particular equipped with Wasserstein metric (which will be introduced later); see for example [1] for Čech complex or Vietoris-Rips complex, and [4] for persistent homology.

Given two metric spaces $\mathcal{A}$ and $\mathcal{B}$, an embedding of $\mathcal{A}$ into $\mathcal{B}$ is a map $\phi : \mathcal{A} \to \mathcal{B}$ that approximately preserves distances, in the sense that the distortion is small:

$$Ld_{\mathcal{A}}(u, v) \leq d_{\mathcal{B}}(\phi(u), \phi(v)) \leq CLd_{\mathcal{A}}(u, v), \qquad \text{for all } u, v \in \mathcal{A}. \tag{3}$$

for some uniform constants $L > 0$ and $C \geq 1$. The distortion of the embedding ϕ is the smallest C such that (3) holds.

One can characterize how "large" a space is (its representational capacity) by the spaces that embed into it with low distortion. In practical terms, this capacity determines the types of data (and relationships between them) that can be well-represented in the embedding space. $\mathbb{R}^n$ with the Euclidean metric, for example, embeds into the L^1 metric with low distortion, while the reverse is not true (Deza & Laurent, 2009).

Wasserstein spaces are very large: Many spaces can embed into Wasserstein spaces with low distortion, even when the converse is not true. $W_p(\mathcal{A})$, for $\mathcal{A}$ an arbitrary metric space, embeds any product space $\mathcal{A}^n$, for example (Kloeckner, 2010), via discrete distributions supported at n points. Even more generally, certain Wasserstein spaces are universal, in the sense that they can embed arbitrary metrics on finite spaces. $W_1(\ell^1)$ is one such space (Bourgain, 1986), and it is still an open problem to determine if $W_1(\mathbb{R}^k)$ is universal for any $k < \infty$. Recently it has been shown that every finite metric space embeds the $\frac{1}{p}$ power of its metric into $W_p(\mathbb{R}^3)$, $p > 1$, with vanishing distortion (Andoni et al., 2015). A hopeful interpretation suggests that $W_1(\mathbb{R}^3)$ may be a plausible target space for arbitrary metrics on symbolic data, with a finite set of symbols; we are unaware of similar universality results for L^p or hyperbolic spaces, for example. See [3, Section 2.3] for more discussions.

2 Distances based on densities

There are several distances for probability measures based on densities. Specifically,

Total Variation	$TV(P,Q)$	$=$	$\sup_A \lvert P(A) - Q(A)\rvert$
L_1	$\lVert P - Q\rVert_1$	$=$	$\int \lvert p - q\rvert$
Kullback-Leibler	$KL(P,Q)$	$=$	$\int p \log(p/q)$
χ^2	$\chi^2(P,Q)$	$=$	$\int \left(\frac{p}{q} - 1\right)^2 dQ = \int \frac{p^2}{q} - 1$
Hellinger	$H(P,Q)$	$=$	$\sqrt{\int (\sqrt{p} - \sqrt{q})^2}$.

Here, Kullback-Leibler and χ^2 are not really distances, in particular since they are not symmetric: $KL(P,Q) \neq KL(Q,P)$ and $\chi^2(P,Q) \neq \chi^2(Q,P)$.

We also define the *affinity* between P and Q by

$$a(P,Q) = \int (p \wedge q).$$

There are many relationships between these quantities. These are summarized in the next two theorems. We leave the proofs as exercises.

Theorem. *The following relationships hold:*

1. $\mathsf{TV}(P,Q) = \frac{1}{2}\lVert P - Q\rVert_1 = 1 - a(P,Q)$. *(Scheffés Theorem.)*
2. $\mathsf{TV}(P,Q) = P(A) - Q(A)$ *where* $A = \{x : \ p(x) > q(x)\}$.
3. $0 \leq H(P,Q) \leq \sqrt{2}$.
4. $H^2(P,Q) = 2(1 - a(P,Q))$.
5. $a(P,Q) \geq \frac{1}{2}a^2(P,Q) = \frac{1}{2}\left(1 - \frac{H^2(P,Q)}{2}\right)^2$. *(Le Cam's inequalities.)*

6. $\frac{1}{2}H^2(P,Q) \leq \mathsf{TV}(P,Q) = \frac{1}{2}\|P-Q\|_1 \leq H(P,Q)\sqrt{1-\frac{H^2(P,Q)}{4}}$.
7. $\mathsf{TV}(P,Q) \leq \sqrt{\mathsf{KL}(P,Q)/2}$. *(Pinsker's inequality.)*
8. $\int(\log dP/dQ)_+ dP \leq \mathsf{KL}(P,Q) + \sqrt{\mathsf{KL}(P,Q)/2}$.
9. $a(P,Q) \geq \frac{1}{2}e^{-\mathsf{KL}(P,Q)}$.
10. $\mathsf{TV}(P,Q) \leq H(P,Q) \leq \sqrt{\mathsf{KL}(P,Q)} \leq \sqrt{\chi^2(P,Q)}$.

Let P^n denote the product measure based on n independent samples from P.

Theorem. *The following relationships hold:*

1. $H^2(P^n,Q^n) = 2\left(1-\left(1-\frac{H^2(P,Q)}{2}\right)^n\right)$.
2. $a(P^n,Q^n) \geq \frac{1}{2}a^2(P^n,Q^n) = \frac{1}{2}\left(1-\frac{1}{2}H^2(P,Q)\right)^{2n}$.
3. $a(P^n,Q^n) \geq \left(1-\frac{1}{2}\|P-Q\|_1\right)^n$.
4. $\mathsf{KL}(P^n,Q^n) = n\mathsf{KL}(P,Q)$.

These distances are all useful, in particular in the context of minimax theory. But they have some drawbacks:

1. We cannot use them to compare P and Q when one is discrete and the other is continuous. For example, suppose that P is uniform on $[0,1]$ and that Q is uniform on the finite set $\{0,1/N,2/N,\ldots,1\}$. Practically speaking, there is little difference between these distributions. But the total variation distance is 1 (which is the largest the distance can be). The Wasserstein distance is $1/N$ which seems quite reasonable.

2. These distances ignore the underlying geometry of the space. To see this consider Figure 1. In this figure we see three densities p_1, p_2, p_3. It is easy to see that $\int|p_1-p_2| = \int|p_1-p_3| = \int|p_2-p_3|$ and similarly for the other distances. But our intuition tells us that p_1 and p_2 are close together. We shall see that this is captured by Wasserstein distance.

3. When we average different objects — such as distributions or images — we would like to make sure that we get back a similar object. The top plot in Figure 2 shows some distributions, each of which is uniform on a circle. The bottom left plot shows the Euclidean average of the distributions which is just a gray mess. The bottom right shows the Wasserstein barycenter (which we will define later) which is a much better summary of the set of images.

4. When we compute the usual distance between two distributions, we get a number but we don't get any qualitative information about why the distributions differ. But with the Wasserstein distance we also get a map that shows us how we have to move the mass of P to morph it into Q.

5. Suppose we want to create a path of distributions (a geodesic) P_t that interpolates between two distributions P_0 and P_1. We would like the distributions P_t to preserve the basic structure of the distributions. Figure 5 shows an example. The top row shows the path between P_0 and P_1 using Wasserstein distance. The bottom row shows the path using L_2 distance. We see that the Wasserstein path does a better job of preserving the structure.

6. Some of these distances are sensitive to small wiggles in the distribution. But we shall see that the Wasserstein distance is insensitive to small wiggles. For example if P is uniform on $[0,1]$ and Q has density $1+\sin(2\pi kx)$ on $[0,1]$ then the Wasserstein distance is $O(1/k)$.

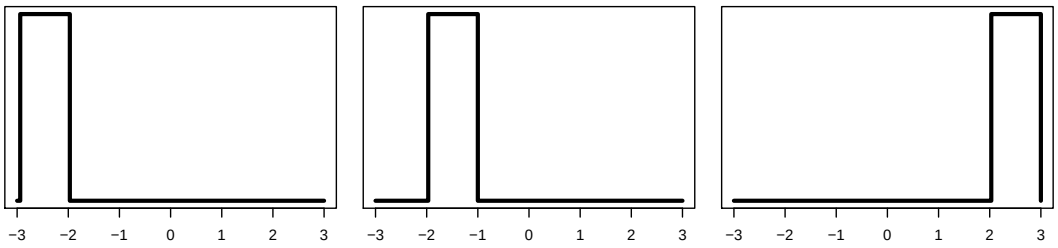

Figure 1: *Three densities p_1, p_2, p_3. Each pair has the same distance in L_1, L_2, Hellinger etc. But in Wasserstein distance, p_1 and p_2 are close.*

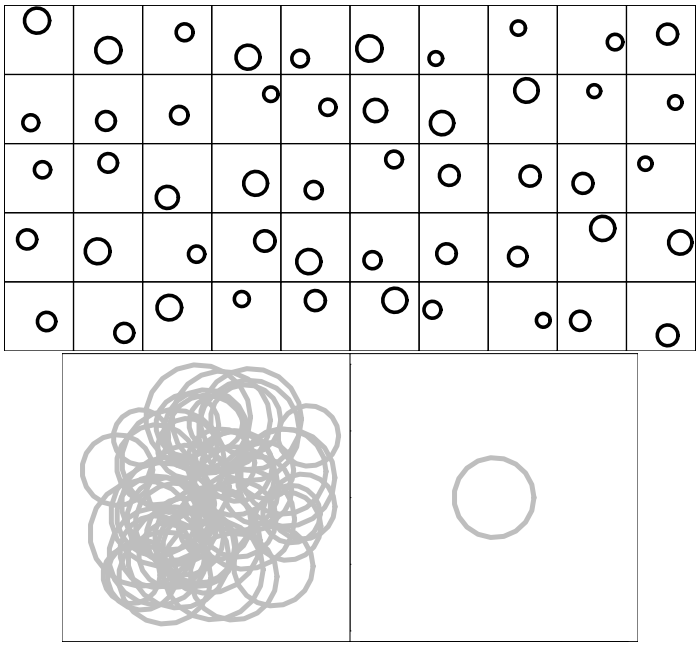

Figure 2: *Top: Some random circles. Bottom left: Euclidean average of the circles. Bottom right: Wasserstein barycenter.*

3 Optimal Transport

If $T : \mathbb{R}^d \to \mathbb{R}^d$ then the the distribution of $T(X)$ is called the push-forward of P, denoted by $T_\# P$. In other words,

$$T_\# P(A) = P\Big(\{x : \ T(x) \in A\Big) = P(T^{-1}(A)).$$

The *Monge* version of the optimal transport distance is

$$\inf_T \int ||x - T(x)||^p dP(x)$$

where the infimum is over all T such that $T_\# P = Q$. Intuitively, this measures how far you have to move the mass of P to turn it into Q. A minimizer T^*, if one exists, is called the *optimal transport map*.

If P and Q both have densities than T^* exists. The map $T_t(x) = (1-t)x + tT^*(x)$ gives the path of a particle of mass at x. Also, $P_t = T_{t\#} P$ is the geodesic connecting P to Q.

But, the minimizer might not exist. Consider $P = \delta_0$ and $Q = (1/2)\delta_{-1} + (1/2)\delta_1$ where δ_a. In this case, there is no map T such that $T_\# P = Q$. This leads us to the Kantorovich formulation where we allow the mass at x to be split and move to more than one location.

Let $\mathcal{J}(P, Q)$ denote all joint distributions J for (X, Y) that have marginals P and Q. In other words, $T_{X\#} J = P$ and $T_{Y\#} J = Q$ where $T_X(x, y) = x$ and $T_Y(x, y) = y$. Figure 4 shows an example of a joint distribution with two

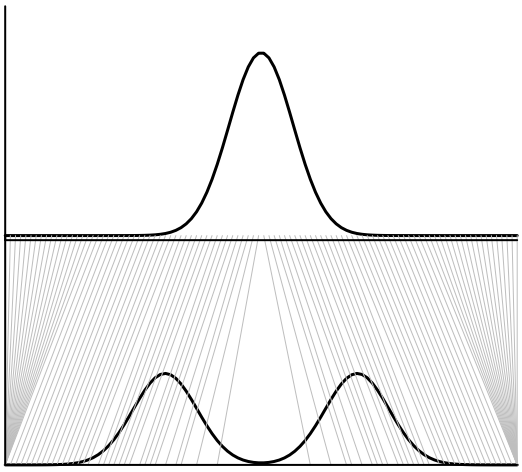

Figure 3: *Two densities p and q and the optimal transport map to that morphs p into q.*

given marginal distributions. Then the Kantorovich, or Wasserstein, distance is

$$W_p(P,Q) = \left(\inf_{J \in \mathcal{J}(P,Q)} \int ||x-y||^p dJ(x,y) \right)^{1/p}$$

where $p \geq 1$. When $p = 1$ this is also called the *Earth Mover distance*. The minimizer J^* (which does exist) is called the *optimal transport plan* or *the optimal coupling*. In case there is an optimal transport map T then J is a singular measure with all its mass on the set $\{(x, T(x))\}$.

It can be shown that

$$W_p^p(P,Q) = \sup_{\psi,\phi} \int \psi(y) dQ(y) - \int \phi(x) dP(x)$$

where $\psi(y) - \phi(x) \leq ||x-y||^p$. This is called the dual formulation. In special case where $p = 1$ we have the very simple representation

$$W_1(P,Q) = \sup \left\{ \int f(x) dP(x) - \int f(x) dQ(x) : \ f \in \mathcal{F} \right\}$$

where $\mathcal{F}$ denotes all maps from $\mathbb{R}^d$ to $\mathbb{R}$ such that $|f(y) - f(x)| \leq ||x-y||$ for all x, y.

When $d = 1$, the distance has a closed form:

$$W_p(P,Q) = \left(\int_0^1 |F^{-1}(z) - G^{-1}(z)|^p \right)^{1/p}$$

and F and G are the cdf's of P and Q. If P is the empirical distribution of a dataset $X_1, \ldots, X_n$ and Q is the empirical distribution of another dataset $Y_1, \ldots, Y_n$ of the same size, then the distance takes a very simple function of the order statistics:

$$W_p(P,Q) = \left(\sum_{i=1}^n ||X_{(i)} - Y_{(i)}||^p \right)^{1/p}.$$

An interesting special case occurs for Normal distributions. If $P = N(\mu_1, \Sigma_1)$ and $Q = N(\mu_2, \Sigma_2)$ then

$$W^2(P,Q) = ||\mu_1 - \mu_2||^2 + B^2(\Sigma_1, \Sigma 2)$$

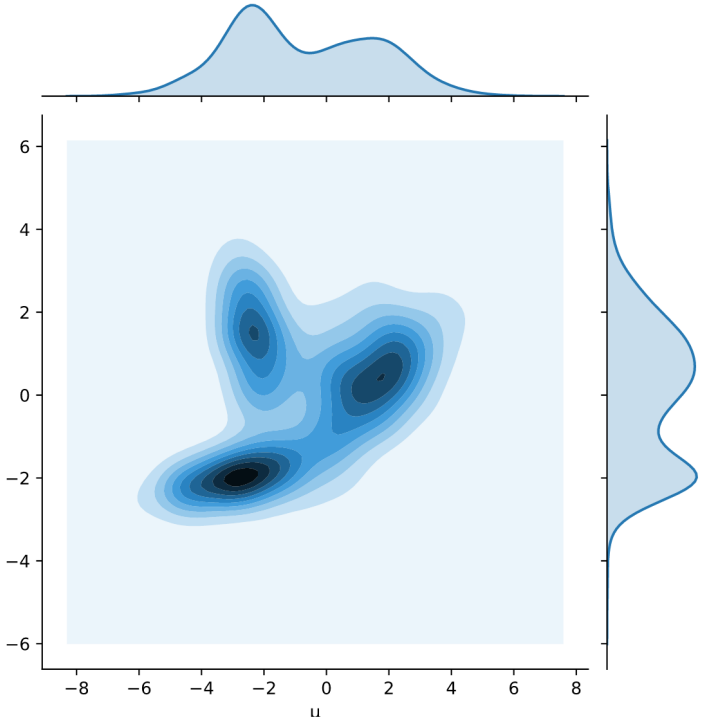

Figure 4: *This plot shows one joint distribution J with a given X marginal and a given Y marginal. Generally, there are many such joint distributions. Image credit: Wikipedia.*

$$W_2^2(P,Q) = ||\mu_1 - \mu_2||^2 + B^2(\Sigma_1, \Sigma_2)$$

where

$$B^2(\Sigma_1, \Sigma 2) = \text{tr}(\Sigma_1) + \text{tr}(\Sigma_2) - 2\text{tr}\left[(\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2})^{1/2}\right].$$

$$B^2(\Sigma_1, \Sigma_2) = \text{tr}(\Sigma_1) + \text{tr}(\Sigma_2) - 2\text{tr}\left[(\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2})^{1/2}\right].$$

There is a connection between Wasserstein distance and L_1 distance (Indyk and Thaper 2003). Suppose that P and Q are supported on $[0,1]^d$. Let $G_1, G_2, \ldots$ be a dyadic sequence of cubic partitions where each cube in G_i has side length $1/2^i$. Let $p^{(i)}$ and $q^{(i)}$ be the multinomials from P and Q one grid G_i. Fix $\epsilon > 0$ and let $k = \log(2d/\epsilon)$. Then

$$W_1(P,Q) \leq 2d \sum_{i=1}^{m} \frac{1}{2^i} ||p^{(i)} - q^{(i)}||_1 + \frac{\epsilon}{2}. \tag{4}$$

There is an almost matching lower bound (but it actually requires using a random grid).

More generally, as discussed in Weed and Bach (2017), for any sequence of dyadic partitions $\mathcal{A}_1, \mathcal{A}_2, \ldots, \mathcal{A}_m$ we have

$$W_p^p(P,Q) \leq \delta^{mp} + \sum_{j=1}^{m} \delta^{(j-1)p} \sum_{A \in \mathcal{A}_j} |P(A) - Q(A)|$$

where $\text{diam}(A) \leq \delta^j$ for every $A \in \mathcal{A}_j$.

These results show that, in some sense, Wasserstein distance is like a multiresolution L_1 distance.

4 Geodesics

Let P_0 and P_1 be two distributions. Consider a map c taking $[0,1]$ to the set of distributions, such that $c(0) = P_0$ and $c(1) = P_1$. Thus $(P_t : \ 0 \leq t \leq 1)$ is a path connecting P_0 and P_1, where $P_t = c(t)$. The length of c — denoted by $L(c)$ — is the supremum of $\sum_{i=1}^m W_p(c(t_{i-1}), c(t_i))$ over all m and all $0 = t_1 < \cdots < t_m = 1$. There exists such a path c such that $L(c) = W(P_0, P_1)$. In other words, $(P_t : 0 \leq t \leq 1)$ is the geodesic connecting P_0 and P_1. It can be shown that

$$P_t = F_{t\#}J$$

where J is the optimal coupling and $F_t(x,y) = (1-t)x + ty$. Examples are shown in Figures 5 and 6.

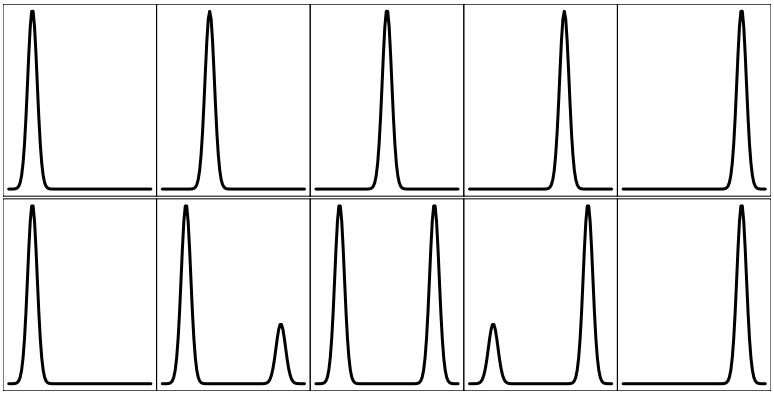

Figure 5: Top row: Geodesic path from P_0 to P_1. Bottom row: Euclidean path from P_0 to P_1.

Figure 6: Morphing one image into another using the Wasserstein geodesic. Image credit: Bauer, Joshi and Modin 2015.

5 Barycenters and PCA

Suppose we have a set of distributions $P_1, \ldots, P_N$. How do we summarize these distributions with one "typical" distribution? We could take the average $\frac{1}{N}\sum_{j=1}^n P_j$. But the resulting average won't look like any of the P_j's. See Figure 7.

Instead we can use the Wasserstein barycenter which is the distribution P that minimizes

$$\sum_{j=1}^{N} W(P, P_j).$$

The bottom right plot of Figure 7 shows an example. You can see that this does a much better job.

We can do the same thing for data sets. See Figure 8. Here we simple regard a dataset as an empirical distribution. The average (red dots) $N^{-1}\sum_j \hat{P}_j$ of these empirical distributions $\hat{P}_j$ is useless. But the Wasserstein barycenter (blue dots) gives us a better sense of what a typical dataset looks like.

Let's pursue this last example a bit more since it will give us some intuition. Suppose we have N datasets $\mathcal{X}_1, \ldots, \mathcal{X}_N$ where $\mathcal{X}_j = \{X_{j1}, \ldots, X_{jn}\}$. For simplicity, suppose that each is of the same size n. In this case, we can describe the Wasserstein barycenter in a simple way. First we find the order statistics for each data set:

$$X_{(j1)} \leq X_{(j2)} \leq \cdots \leq X_{(jn)}.$$

Now for each $1 \leq r \leq n$, we find the average r^{th} average order statistic:

$$Y_{(r)} = \frac{1}{N}\sum_{j=1}^{N} X_{(jr)}.$$

Then $\mathcal{Y} = \{Y_{(1)}, \ldots, Y_{(n)}\}$ is the Wasserstein barycenter. In a sense, all we are really doing is converting to quantiles and averaging.

If $P_j = N(\mu_j, \Sigma_j)$ for $j = 1, \ldots, N$ then the Barycenter is $N(\mu, \Sigma)$ where $\mu = N^{-1}\sum_j \mu_j$ and Σ satisfies

$$\frac{1}{N}\sum_j (\Sigma^{1/2}\Sigma_j\Sigma^{1/2})^{1/2}.$$

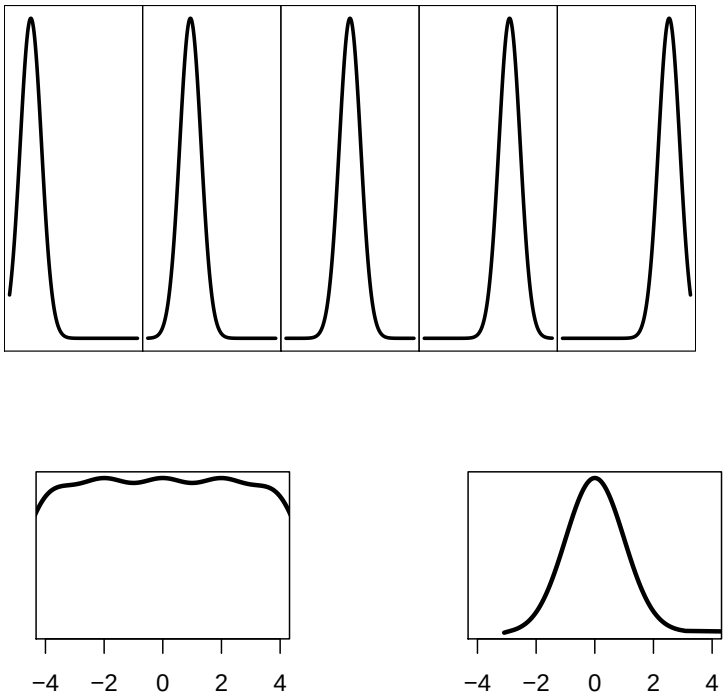

Figure 7: *Top: Five distributions. Bottom left: Euclidean average of the distributions. Bottom right: Wasserstein barycenter.*

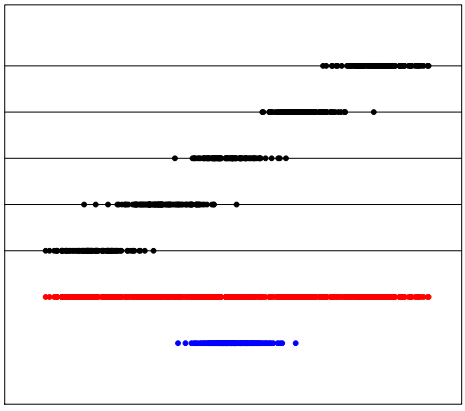

Figure 8: *The top five lines show five, one-dimensional datasets. The red points the what happens if we simple average the give empirical distributions. The blue dots show the Wasserstein barycenter which, in this case, can be obtained simply by averaging the order statistics.*

Now that we have a notion of average, it is possible to define a Wasserstein version of PCA. There are several approaches; see, for example Seguy and Cuturi (2015), Boissard et al (2013), Bigot (2014), Wang, Wei and Slepcev (2013). The idea, as with the barycenters, is to find orthogonal directions of variation in the space of measures (or images). Here I'll briefly describe the method from Wang, Wei and Slepcev (2013).

Let $P_1, \ldots, P_N$ be distributions with densities. Let R be a reference distribution with density r. Define $\psi_j(x) = (T_j(x) - x)\sqrt{r(x)}$. The set of distributions endowed with the W^2 distance is a manifold and $\int(\psi_j(x) - \psi_k(x))^2 dx$ is the distance between the projections onto the tangent space at R. In other words, ψ_j defines an approximate embedding of the set of distributions and L^2. We can now perform PCA on the functions $\psi_1, \ldots, \psi_N$.

6 Minimax Rates

Equation (4) can be used to compute rates of convergence. Suppose that the sample space is $[0,1]^d$. P_n converges to P in Wasserstein distance as

$$\mathbb{E}W_p(P_n, P) \lesssim \begin{cases} n^{-1/2} & p \geq d/2 \\ n^{-1/2}\log(1+n) & p = d/2 \\ n^{-p/d} & p < d/2. \end{cases}$$

The lower bound for minimax rate is

$$\inf_{\hat{P}} \sup_{P \in \mathcal{P}} \mathbb{E}W_p(\hat{P}, P) \gtrsim \begin{cases} n^{-1/2} & p > d/2 \\ n^{-p/d} & p < d/2. \end{cases}$$

The optimal estimator is the empirical distribution. This is a nice property about Wasserstein: there is no need to smooth.

Now suppose we observe $X_1, \ldots, X_n \sim P$ supported on $[0, \Delta]^d$. We want to test $H_0 : P = P_0$ versus $H_1 : W_1(P, P_0) > \epsilon$. Ba et al (2013) and Deng, Li and Wu (2017) showed that the minimax separation rate is (ignoring some log terms)

$$\epsilon_n \asymp \frac{2\Delta d}{n^{\frac{3}{2d}}}.$$

In the special case were P and P_0 are concentrated in k small clusters, the rate becomes

$$\epsilon_n \asymp d\Delta \left(\frac{k}{n}\right)^{1/4}.$$

7 Confidence Intervals

How do we get hypothesis tests and confidence intervals for the Wasserstein distance? Usually, we would use some sort of central limit theorem. Such results are available when $d = 1$ but are elusive in general.

del Barrio and Loubes (2017) show that

$$\sqrt{n}(W_2^2(P, P_n) - \mathbb{E}[W_2^2(P, P_n)]) \rightsquigarrow N(0, \sigma^2(P))$$

for some $\sigma^2(P)$. And, in the two sample case

$$\sqrt{\frac{nm}{n+m}}\left(W_2^2(P_n, Q_m) - \mathbb{E}[W_2^2(P_n, Q_m)]\right) \rightsquigarrow N(0, \sigma^2(P, Q))$$

for some $\sigma^2(P, Q)$. Unfortunately, these results do not give a confidence interval for $W(P, Q)$ since the limit is centered around $\mathbb{E}[W_2^2(P_n, Q_m)]$ instead of $W_2^2(P, Q)$. However, del Barrio, Gordaliza and Loubes (2018) show that if some smoothness assumptions holds, then the distribution centers around $W_2^2(P, Q)$. More generally, Tudor, Siva and Larry have a finite sample confidence interval for $W(P, Q)$ without any conditions.

All this is for $d = 1$. The case $d > 1$ seems to be unsolved.

Another interesting case is when the support $\mathcal{X} = \{x_1, \ldots, x_k\}$ is a finite metric space. In this case, Sommerfeld and Munk (2017) obtained some precise results. First, they showed that

$$\left(\frac{nm}{n+m}\right)^{\frac{1}{2p}} W_p(P_n, Q_m) \rightsquigarrow \left(\max_u \langle G, u\rangle\right)^{1/p}$$

G is a mean 0 Gaussian random vector and u varies over a convex set. By itself, this does not yield a confidence set. But they showed that the distribution can be approximated by subsampling, where the subsamples of size m with $m \to \infty$ and $m = o(n)$.

You might wonder why the usual bootstrap does not work. The reason is that the map $(P, Q) \mapsto W_p^p(P, Q)$ is not Hadamard differentiable. This means that the map does not have smooth derivatives. In general, the problem of constructing confidence intervals for Wasserstein distance is unsolved.

8 Robustness

One problem with the Wasserstein distance is that it is not robust. To see this, note that $W(P, (1-\epsilon)P + \epsilon\delta_x) \to \infty$ as $x \to \infty$.

However, a partial solution to the robustness problem is available due to Alvarez-Esteban, del Barrio, Cuesta Albertos and Matran (2008). They define the α-trimmed Wasserstein distance

$$\tau(P, Q) = \inf_A W_2(P_A, Q_A)$$

where $P_A(\cdot) = P(A \bigcap \cdot)/P(A)$, $Q_A(\cdot) = Q(A \bigcap \cdot)/Q(A)$ and A varies over all sets such that $P(A) \geq 1 - \alpha$ and $Q(A) \geq 1 - \alpha$. When $d = 1$, they show that

$$\tau(P, Q) = \inf_A \left(\frac{1}{1-\alpha} \int_A (F^{-1}(t) - G^{-1}(t))^2 dt \right)^{1/2}$$

where A varies over all sets with Lebesgue measure $1 - \alpha$.

9 Inference From Simulations

Suppose we have a parametric model $(P_\theta : \ \theta \in \Theta)$. We can estimate θ using the likelihood function $\prod_i p_\theta(X_i)$. But in some cases we cannot actually evaluate p_θ. Instead, we can simulate from P_θ. This happens quite often, for example, in astronomy and climate science. Berntom et al (2017) suggest replacing maximum likelihood with minimum Wasserstein distance. That is, given data $X_1, \ldots, X_n$ we use

$$\hat{\theta} = \text{argmin}_\theta W(P_\theta, P_n)$$

where P_n is the ampirical measure. We estimate $W(P_\theta, P_n)$ by $W(Q_N, P_n)$ where Q_N is the empirical measure based on a sample $Z_1, \ldots, Z_N \sim P_\theta$.

10 Computing the Distance

We saw that, when $d = 1$,

$$W_p(P, Q) = \left(\int_0^1 |F^{-1}(z) - G^{-1}(z)|^p \right)^{1/p}$$

and F and G are the cdf's of P and Q. If P is the empirical distribution of a dataset $X_1, \ldots, X_n$ and Q is the empirical distribution of another dataset $Y_1, \ldots, Y_n$ of the same size, then the distance takes a very simple function of the order statistics:

$$W_p(P, Q) = \left(\sum_{i=1}^n ||X_{(i)} - Y_{(i)}||^p \right)^{1/p}.$$

The one dimensional case is, perhaps, the only case where computing W is easy.

For any d, if P and Q are empirical distributions — each based on n observations — then

$$W_p(P, Q) = \inf_\pi \left(\sum_i ||X_i - Y_{\pi(i)}||^p \right)^{1/p}$$

where the infimum is over all permutations π. This may be solved in $O(n^3)$ time using the Hungarian algorithm.

Suppose that P has density p and that $Q = \sum_{j=1}^m q_j \delta_{y_j}$ is discrete. Given weights $w = (w_1, \ldots, w_m)$ define the power diagram $V_1, \ldots, V_m$ where $y \in V_j$ if y is closer to the ball $B(y_j, w_j)$ and any other ball $B(y_s, w_s)$. Define the map $T(x) = y_j$ when $x \in V_j$. According to a result known as Bernier's theorem, if have that $P(V_j) = q_j$ then

$$W_2(P,Q) = \left(\sum_j \int_{V_j} ||x - y_j||^2 dP(x) \right)^{1/2}.$$

The problem is: how do we choose w is that we end up with $P(V_j) = q_j$? It was shown by Aurenhammer, Hoffmann, Aronov (1998) that this corresponds to minimizing

$$F(w) = \sum_j \left(q_j w_j - \int_{V_j} [||x - y_j||^2 - w_j] dP(x) \right).$$

Merigot (2011) gives a multiscale method to minimize $F(w)$.

There are a few papers (Merigot 2011 and Gerber and Maggioni 2017) use multiscale methods for computing the distance. These approaches make use of decompositions like those used for the minimax theory.

Cuturi (2013) showed that if we replace $\inf \mathbb{E}||x-y||^p dJ(x,y)$ with the regularized version $\inf \mathbb{E}||x-y||^p dJ(x,y) + \int j(x,y) \log j(x,y)$ then a minimizer can be found using a fast, iterative algorithm called the Sinkhorn algorithm. However, this requires discretizing the space and it changes the metric.

Finally, recall that, if $P = N(\mu_1, \Sigma_1)$ and $Q = N(\mu_2, \Sigma_2)$ then

$$W^2(P,Q) = ||\mu_1 - \mu_2||^2 + B^2(\Sigma_1, \Sigma 2)$$

where

$$B^2(\Sigma_1, \Sigma 2) = \text{tr}(\Sigma_1) + \text{tr}(\Sigma_2) - 2\text{tr}\left[(\Sigma_1^{1/2} \Sigma_2 \Sigma_1^{1/2})^{1/2} \right].$$

Clearly computing the distance is easy in this case.

11 Estimating Optimal Transport Maps from Samples

In many modern applications of optimal transport, the object of primary interest is not merely the Wasserstein distance between two probability distributions, but the optimal transport map that pushes one distribution onto another in a cost-minimizing manner. When the underlying distributions P and Q are known and sufficiently regular—for instance, when both are absolutely continuous with respect to Lebesgue measure on $\mathbb{R}^d$—the classical Monge formulation defines an optimal transport map

$$T_0 \;=\; \arg\min_{T : T_\# P = Q} \int \|T(x) - x\|^2 \, dP(x),$$

where $T_\# P = Q$ means the pushforward of P under T equals Q. Then Brenier's theorem guarantees that, under suitable convexity assumptions, this map is unique and equals the gradient of a convex potential.

In practice, however, the distributions P and Q are almost never known analytically. Instead, we observe independent samples

$$X_1, \ldots, X_n \sim P, \qquad Y_1, \ldots, Y_m \sim Q,$$

and the central question becomes whether it is possible to estimate T_0 from these empirical observations, and, if so, at what rate and with what kind of inferential guarantees.

The recent survey by Balakrishnan, Manole, and Wasserman (2025) provides a comprehensive overview of modern advances on this problem and places the estimation of transport maps into a broader statistical-inference framework.

To understand what it means to estimate the optimal transport map, one treats T_0 as a functional of the pair (P, Q). An estimator $\widehat{T}$ is typically defined by solving an optimal transport problem between empirical (or smoothed empirical) distributions $\widehat{P}_n$ and $\widehat{Q}_m$, possibly followed by a regularization step to stabilize the resulting discrete map. The statistical objective is to analyze how well $\widehat{T}$ approximates T_0, either pointwise or through integrated risk such as

$$\int \|\widehat{T}(x) - T_0(x)\|^p \, dP(x).$$

A second, increasingly important aim is to develop limit theorems—for example, central limit theorems or bootstrap approximations—that permit the construction of confidence sets and uncertainty quantification for the estimator $\widehat{T}$. Such tasks are technically challenging because the map $(P, Q) \mapsto T_0$ is highly nonlinear and may be unstable under even small perturbations of the underlying distributions unless strong regularity assumptions are imposed.
There are many ways to construct estimates of the OT and entropic maps from the observed samples. One natural strategy to construct an estimate of the transport map is to first construct an estimate of the optimal coupling between the empirical measures

$$P_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}, \quad Q_m = \frac{1}{m}\sum_{j=1}^{m} \delta_{Y_j}.$$

This amounts to solving Kantorovich's linear program with P and Q replaced by P_n and Q_m, which can be done by the Hungarian algorithm or the Auction algorithm. If $m = n$ then this amounts to finding the permutation π that minimizes

$$n^{-1}\sum_{i} \|X_i - Y_{\pi(i)}\|^2.$$

This does not however yield an estimate of the transport map defined outside the sample points $X_1, \ldots, X_n$. We can obtain a transport map defined on the entire domain using ideas from nonparametric regression (for instance by nearest neighbors or local smoothing).
An alternative, closely related, class of estimators can be constructed via the plugin principle. Here we first construct estimates $\hat{P}_n$ and $\hat{Q}_m$ of the measures P and Q, and then define $\hat{T}_{nm}$ to be the optimal transport map between $\hat{P}_n$ and $\hat{Q}_m$. For $\hat{T}_{nm}$ to be well-defined, the estimators $\hat{P}_n$ and $\hat{Q}_m$ need to be proper, in the sense that they define probability measures in their own right, and are related by an optimal transport map (which is, for instance, the case if $m = n$ for the empirical distributions, or if $\hat{P}_n$ is absolutely continuous). In cases where $\hat{P}_n$ and $\hat{Q}_m$ define probability measures but are not related by an OT map, one can instead compute an OT coupling Π_{nm} and project this coupling onto the space of maps.
This projection, known as the barycentric projection, is simply the conditional expectation of the OT coupling:

$$\hat{T}_{nm}(x) = \mathbb{E}_{\Pi_{nm}}[Y \mid X = x].$$

Theorem. *Theorem 3. Suppose that $Q = (\nabla\varphi_0)_{\#}P$, for a β-smooth and α-strongly convex function φ_0. Then for any P and Q the barycentric projection $\hat{T}_{nm}$ above satisfies:*

$$\left\|\hat{T}_{nm} - T_0\right\|^2_{L^2(P)} \leq \left((1+\beta)W_2(\hat{P}, P) + \sqrt{\frac{\beta}{\alpha}}W_2(\hat{Q}, Q)\right)^2.$$

The following result appears in the work of Balakrishnan and Manole [2025] who build on previous results of Manole et al. [2024a].

Theorem. *Theorem 4 (Informal). For any $s > 0$, suppose that p, q are s-smooth and strictly positive on a known regular domain Ω. Then there exist proper and absolutely continuous estimators $\hat{P}_n$ and $\hat{Q}_n$ such that the unique optimal transport map $\hat{T}_n$ pushing $\hat{P}_n$ forward onto $\hat{Q}_m$ satisfies*

$$\mathbb{E}\left\|\hat{T}_{nm} - T_0\right\|^2_{L^2(P)} \leq C\epsilon_{n\wedge m}, \qquad \text{where } \epsilon_n = \begin{cases} 1/n, & d = 1, \\ \log n/n, & d = 2, \\ n^{-\frac{2(s+1)}{2s+d}}, & d \geq 3. \end{cases}$$

12 Applications

The Wasserstein distance is now being used for many tasks in statistical machine learning including:

- Two-sample testing without smoothness
- goodness-of-fit
- analysis of mixture models
- image processing

- dimension reduction
- generative adversarial networks
- domain adaptation
- signal processing

The domain adaptation application is very intriguing. Suppose we have two data sets $\mathcal{D}_1 = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ and $\mathcal{D}_2 = \{(X_1', Y_1'), \ldots, (X_N', Y_N')\}$ from two related problems. We want to construct a predictor for the first problem. We could use just $\mathcal{D}_1$. But if we can find a transport map T that makes $\mathcal{D}_2$ similar to $\mathcal{D}_1$, then we can apply the map to $\mathcal{D}_2$ and effectively increase the sample size for problem 1. This kind of reasoning can be used for many statistical tasks.

13 Summary

Wasserstein distance has many nice properties and has become popular in statistics and machine learning. Recently, for example, it has been used for Generative Adversarial Networks (GANs).

But the distance does have problems. First, it is hard to compute. Second, as we have seen, we do not have a way to do inference for the distance. This reflects the fact that the distance is not a smooth functional which is, itself not a good thing. We have also seen that the distance is not robust although, the trimmed version may fix this.

References

[1] Henry Adams, Florian Frick, and Žiga Virk. Vietoris thickenings and complexes have isomorphic homotopy groups. *J. Appl. Comput. Topol.*, 7(2):221–241, 2023.

[2] Rick Durrett. *Probability—theory and examples*, volume 49 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 2019. Fifth edition.

[3] Charlie Frogner, Farzaneh Mirzazadeh, and Justin Solomon. Learning embeddings into entropic wasserstein spaces. *CoRR*, abs/1905.03329, 2019.

[4] Weichen Wu, Jisu Kim, and Alessandro Rinaldo. On the estimation of persistence intensity functions and linear representations of persistence diagrams. In Sanjoy Dasgupta, Stephan Mandt, and Yingzhen Li, editors, *International Conference on Artificial Intelligence and Statistics, 2-4 May 2024, Palau de Congressos, Valencia, Spain*, volume 238 of *Proceedings of Machine Learning Research*, pages 3610–3618. PMLR, 2024.