Optimal Transport
10716, Spring 2020
Pradeep Ravikumar

1 Introduction

In the "world of atoms," many important problems involving logistics, production planning, and routing, among others, can be cast as moving (typically discrete) distributions. This "transport" problem was also a key motivational driver behind the development of theory of linear programming itself. In statistical machine learning, almost all tasks involve distributions, and many modern tasks such as domain adaption, as well as classical tasks such as clustering, classification and statistical estimation could be cast as involving an optimal transport of distributions. Even where data do not take the form of distributions per se, e.g. images, a classical approach has been to consider local features (e.g. statistics of image patches), and to consider an image as a histogram (i.e. discrete distribution) of such local features. This is called a bag of features approach, but there could be more general approaches to "lift" inputs to a space of distributions.

Even without such lifting, ML tasks crucially involve distances between distributions (e.g. KL divergences specify the classical MLE estimator), however classical divergences such as f-divergences are not always suited to more complex data such as images. As we will see, optimal transport based distances leverage the underlying geometry of the input space, are more suited to discrete distributions, and more generally to "singular" distributions which do not have support everywhere (e.g. restricted to a manifold), and to settings where the modeling distribution does not have a closed form density: all of which are not just nice to have, but form crucial characteristics of modern ML settings.

2 Monge Assignments

Consider an input space $\mathcal{X}$, and suppose we have a pairwise cost function (think of this as a distance function) $c : \mathcal{X} \mapsto \mathcal{X} \mapsto \mathbb{R}$. Thus, given two points x_i and y_j from $\mathcal{X}$, the "transport cost" to move x_i to y_j is $c(x_i, y_j)$. Now, suppose we have two sets of n points $\{x_i\}_{i=1}^n$, and $\{y_j\}_{j=1}^n$.

What is the cost of transporting one set of points to the other?

A natural extension of the simple one point case is the following:

$$\inf_{\sigma \in \mathrm{Perm}(n)} \sum_{i=1}^{n} c(x_i, y_{\sigma(i)}),$$

where Perm(n) denotes the set of all permutations $\sigma : [n] \mapsto [n]$ over $[n] = \{1, \ldots, n\}$.

We could generalize this development above to general discrete measures. Let

$$\alpha = \sum_{i=1}^{n} a_i \delta_{x_i}$$
$$\beta = \sum_{j=1}^{n} b_j \delta_{y_n}, \tag{1}$$

be two discrete measures, where $\sum_{i=1}^{n} a_i = \sum_{j=1}^{n} b_j = 1$, and δ_z is the Dirac measure on $\{z\}$. Then consider the so-called "Monge" problem that seeks a map $T : \{x_1, \ldots, x_n\} \mapsto \{y_1, \ldots, y_m\}$ from one set to the other that satisfies the following mass conservation constraint:

$$b_j = \sum_{i \in [n] : T(x_i) = y_j} a_i,$$

which we can write compactly as $T_{\#}\alpha = \beta$. We then seek a map that minimizes the overall transportation cost while satisfying the mass conservation constraint:

$$\min_{T} \left\{ \sum_{i=1}^{n} c(x_i, T(x_i)) \ : \ T_{\#}\alpha = \beta \right\}.$$

We can generalize this further to general measures.

Pushforward Operator. For any continuous map $T : \mathcal{X} \mapsto \mathcal{Y}$, we can define its corresponding pushforward operator $T_{\#} : \mathcal{M}(\mathcal{X}) \mapsto \mathcal{M}(\mathcal{Y})$ as follows. For discrete measures $\alpha = \sum_{i=1}^{n} a_i \delta_{x_i}$, the push-forward operator merely moves the positions of the points in the support, so that:

$$T_{\#}\alpha = \sum_{i=1}^{n} a_i \delta_{T(x_i)}.$$

For a general measure $\alpha \in \mathcal{M}(\mathcal{X})$, $\beta = T_{\#}\alpha$ satisfies: for all measurable sets $B \subseteq \mathcal{Y}$,

$$\beta(B) = \alpha(\{x \in \mathcal{X} \ : \ T(x) \in B\}) = \alpha(T^{-1}(B)),$$

where $T^{-1}(B)$ is the pre-image of B under T. Thus, while T moves points in $\mathcal{X}$ to points in $\mathcal{Y}$, $T_{\#}$ pushes forward probability mass of measure $\alpha \in \mathcal{M}(\mathcal{X})$ to obtain a measure $\beta \in \mathcal{M}(\mathcal{Y})$.

When α, β have densities ρ_α, ρ_β, we know that by the change of variables formula:

$$\rho(\alpha)(x) = |\det J\, T(x)|\, \rho_\beta(T(x)),$$

where $JT(x)$ is the Jacobian of T at x. Thus, if α has density ρ_α, then $T_\#\alpha$ does not simply have the density $\rho_\alpha \circ T$, due to the presence of the Jacobian. Though, as discussed in the "reparametrization trick" in the previous lecture, we do have that:

$$\mathbb{E}_{Y\sim T_\#\alpha}[g(Y)] = \mathbb{E}_{X\sim\alpha}[g\circ T(X)].$$

Given this notation of a pushforward operator, we can define the general Monge problem between arbitrary measures as follows. Given two arbitrary probability measures $\alpha \in \mathcal{M}(\mathcal{X})$, and $\beta \in \mathcal{M}(\mathcal{Y})$, supported on two spaces $\mathcal{X}, \mathcal{Y}$, we wish to solve:

$$\min_T \left\{ \int_{\mathcal{X}} c(x, T(x)) d\alpha(x) \ : \ T_\#\alpha = \beta. \right\}$$

The main difficulty with the Monge assignment problem is that the Monge map $T(\cdot)$ that solves the Monge assignment problem need not always exist, particularly between two discrete measures with differing support sizes $m \neq n$. Since in such a case, it might not always be possible to get maps T that satisfy the mass conservation constraint. Even if such maps exists, the mass convservation constraint is highly non-convex, and is thus difficult to solve or approximate.

3 Kantorovich Relaxation

A relaxation of the Monge assignment problem, which also makes the problem much simpler, is to relax the deterministic transport map T to a probabilistic transport map instead. For the case with discrete measures in Eqn.(1), consider the following set of probabilistic maps:

$$U(a,b) = \left\{ P \in \mathbb{R}^{m\times n} \ : \ \forall i \in [n], j \in [m], \ \sum_{j=1}^{m} P_{ij} = a_i, \ \sum_{i=1}^{n} P_{ij} = b_j \right\}.$$

The Monge assignment problem can then be relaxed to Kantorovich's optimal transport problem:

$$L_C(a,b) = \min_{P\in U(a,b)} \sum_{i,j} C_{ij} P_{ij}, \tag{2}$$

where $C \in \mathbb{R}^{n\times m}$ is the cost matrix with entries $C_{ij} = c(x_i, x_j)$. This can be seen to be a linear program, which is much more tractable than the non-convex Monge assignment problem.

Example: Factories and Warehouses. Such optimal transport arises frequently in the context of resource allocation. Suppose there are n warehouses, indexed by $i \in [n]$, and where the i-th warehouse has a_i units of raw material. While there are m factories, indexed by $j \in [m]$, and where the j-th factory needs b_j units of raw material. The transportation or logistics company charges $c(i, j)$ to transport a unit of raw material from the location of warehouse i to the location of factory j (perhaps as a function of the locations x_i and y_j). Then, the overall cost to transport raw material from warehouses to factories is precisely the optimal transport objective in Eqn. (2).

For more general measures α, β, the relaxed mass conservation constraint can be written as a constraint on marginal distributions of a joint distribution:

$$U(\alpha, \beta) = \{\pi \in \mathcal{M}(\mathcal{X} \times \mathcal{Y}) \,:\, P_{\mathcal{X}\#}\pi = \alpha,\ P_{\mathcal{Y}\#}\pi = \beta\}\,,$$

where $P_{\mathcal{X}}(x, y) = x$, and $P_{\mathcal{Y}}(x, y) = y$ are coordinate projection maps. The Kantorovich optimal transport problem then becomes:

$$\mathcal{L}_c(\alpha, \beta) = \min_{\pi \in U(\alpha,\beta)} \int_{\mathcal{X}\times\mathcal{Y}} c(x, y) d\pi(x, y). \tag{3}$$

This is also a linear program, albeit an infinite-dimensional one. It always has a solution when $\mathcal{X}, \mathcal{Y}$ are compact spaces, and c is continuous. The minimizer π^* is called the *optimal transport plan* or *the optimal coupling*.

We can also write the optimal transport problem just in terms of random variables as:

$$\mathcal{L}_c(\alpha, \beta) = \min_{(X,Y)} \left\{\mathbb{E}_{(X,Y)} c(X, Y) \,:\, X \sim \alpha,\, Y \sim \beta\right\}.$$

Figure 1 shows an example of a joint distribution with two given marginal distributions.

4 Kantorovich Dual

Consider the discrete optimal transport problem in (2), which was a finite-dimensional linear program. Its corresponding dual program can then be written as:

$$L_C(a, b) = \max_{(f,g)\in R(C)} \langle f, a\rangle + \langle g, b\rangle,$$

where $R(C) = \{(f, g) \in \mathbb{R}^n \times \mathbb{R}^m \,:\, \forall (i, j) \in [n] \times [m],\ f_i + g_j \le C_{ij}\}$ is the set of admissible dual variables.

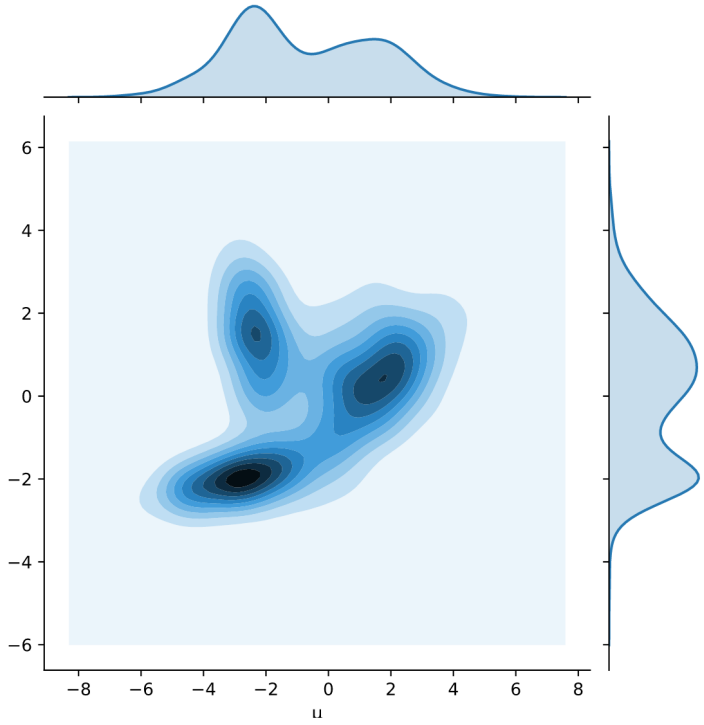

Figure 1: *This plot shows one joint distribution J with a given X marginal and a given Y marginal. Generally, there are many such joint distributions. Image credit: Wikipedia.*

Example: Factories and Warehouses (Contd.) Continuing with the example of n warehouses, and m factories, suppose a logistics vendor charges price f_i per unit raw material to move from warehouse i, and price g_j per unit raw material to move to factory j. Then, we would consider this vendor only when they satisfy the constraint that $f_i + g_j \leq C_{ij}$, since otherwise we would just go with the transportation company. On the other hand, the vendor would aim to ask for the highest possible prices, so that they would aim for the highest overall price $\langle f, a\rangle + \langle g, b\rangle$ subject to the cost constraint above. By duality of linear programs, the two objectives are exactly the same. But note that this allows us to pass the onus of the potentially difficult task of ensuring optimality to the logistics vendor: we just need to ensure the individual constraints that $f_i + g_j \leq C_{ij}$.

For more general measures, we have that:

$$\mathcal{L}_c(\alpha, \beta) = \sup_{(f,g)\in\mathcal{R}(c)} \int_{\mathcal{X}} f(x)d\alpha(x) + \int_{\mathcal{Y}} g(y)d\beta(y),$$

where, denoting $\mathcal{C}(\mathcal{X})$ as the set of real-valued continuous functions with domain $\mathcal{X}$, the set of admissible dual "potentials" is:

$$\mathcal{R}(c) = \{(f, g) \in \mathcal{C}(\mathcal{X}) \times \mathcal{C}(\mathcal{Y}) \,:\, \forall (x, y),\, f(x) + g(y) \leq c(x, y)\}.$$

Here f, g are continuous functions, that are also called "Kantorovich" potentials.

5 Wasserstein Distance

There are two ingredients to the optimal transport problem: the cost function $c(x, y)$, and the two measures α, β. An important setting is where $\mathcal{X} = \mathcal{Y}$, and the cost function corresponds to some metric $d(x, y)$ over $\mathcal{X}$. In such cases, optimal transport specifies a very useful class of distances called Wasserstein distances.

Given a distance metric $d : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}_+$, the p-Wasserstein distance on $\mathcal{X}$ is given as:

$$\mathcal{W}_p(\alpha, \beta) = \mathcal{L}_{d^p}(\alpha, \beta)^{1/p}.$$

The p-Wasserstein distance is actually a distance metric over $\mathcal{M}(\mathcal{X})$: it is symmetric, non-negative, is zero iff the two measures are equal, and satisfies the triangle inequality. When not specified, we typically mean the 2-Wasserstein distance.

When $p = 1$ this is also called the *Earth Mover distance.*

Crucial advantages of Wasserstein distances is that it uses the underlying geometry of the input space (via the metrix d), and is also more suited to discrete distributions. To see this, let us compare Wasserstein with some classical divergence measures.

5.1 Comparative Vignettes of Wasserstein Distances

Suppose $X \sim P$ and $Y \sim Q$ and let the densities be p and q. We assume that $X, Y \in \mathbb{R}^d$. Some popular divergence measures between P and Q include:

$$\begin{aligned}
\text{Total Variation} : & \sup_A |P(A) - Q(A)| = \frac{1}{2} \int |p - q| \\
\text{Hellinger} : & \sqrt{\int (\sqrt{p} - \sqrt{q})^2} \\
L_2 : & \int (p - q)^2 \\
\chi^2 : & \int \frac{(p - q)^2}{q}.
\end{aligned}$$

These distances are all useful, but they have some drawbacks.

Applicability to Singular Discrete Distributions. We cannot use classical distances 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

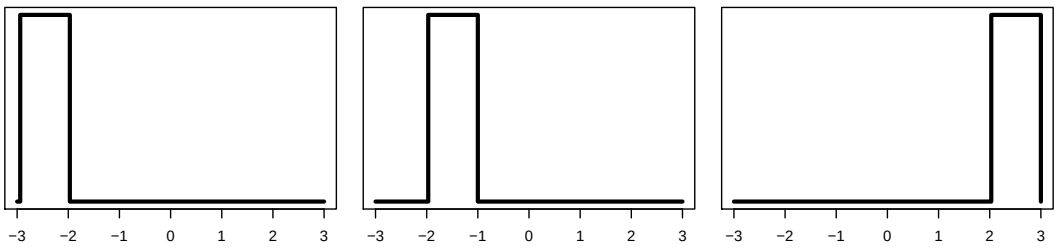

Figure 2: *Three densities p_1, p_2, p_3. Each pair has the same distance in L_1, L_2, Hellinger etc. But in Wasserstein distance, p_1 and p_2 are close.*

variation distance is 1 (which is the largest the distance can be). The Wasserstein distance is $1/N$ which seems quite reasonable.

More generally, classical divergences are not always suited to "singular" distributions which do not have support everywhere (e.g. restricted to a manifold), whereas Wasserstein distances are still well-defined.

Using Underlying Geometry of Space. Classical distances ignore the underlying geometry of the space. To see this consider Figure 2. 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. This is captured by the Wasserstein distance.

As another consequence, some of the classical distances are sensitive to small wiggles in the distribution. But 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 $O(1/k)$.

Geometry over Space of Distributions. 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 an "optimal transport" map that shows us how we have to move the mass of P to morph it into Q.

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 4 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.

When we average different objects — such as distributions or images — we would like to make

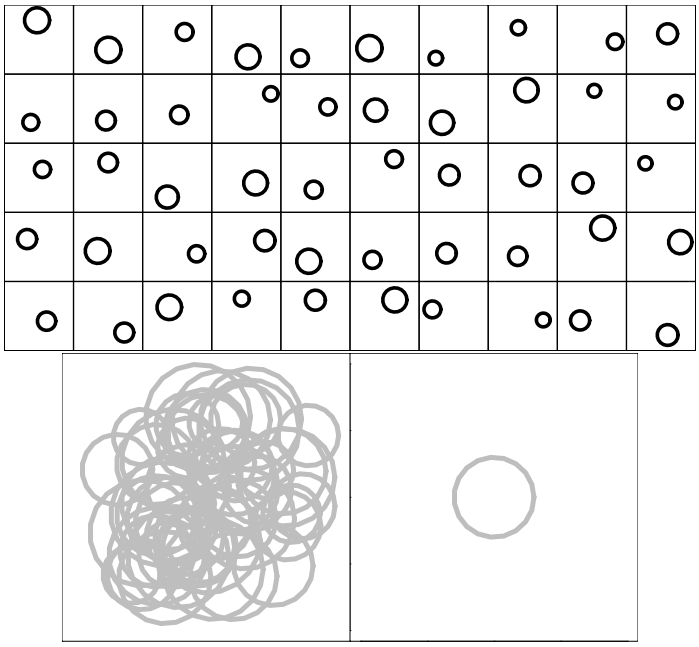

Figure 3: *Top: Some random circles. Bottom left: Euclidean average of the circles. Bottom right: Wasserstein barycenter.*

sure that we get back a similar object. The top plot in Figure 3 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.

Weak Convergence. On a compact domain $\mathcal{X}$, a sequence of measures $\{\alpha_n\}$ is said to weakly converge to a measure $\alpha \in \mathcal{M}(\mathcal{X})$ iff for all continuous functions $f : \mathcal{X} \mapsto \mathbb{R}$, we have that $\int_{\mathcal{X}} f d\alpha_n \to \int_{\mathcal{X}} f d\alpha$, as $n \to \infty$. This convergence can be shown to be equivalent to $\mathcal{W}_p(\alpha_n, \alpha) \to 0$.

5.2 Computing Wasserstein Distances

Total Variation. In the discrete case, where $C_{ij} = \mathbb{I}[i \neq j]$, then the 1-Wasserstein distance between a and b is given by $\|a-b\|_1$. And for arbitrary measures, with $c(x, y) = \mathbb{I}(x \neq y)$, the 1-Wasserstein distance between α and β over $\mathcal{X}$ is equal to their TV distance $\sup_A |\alpha(A) - \beta(A)| = \frac{1}{2} \int_{\mathcal{X}} |d\alpha(x) - d\beta(x)|$.

1D case: Empirical Measures. Suppose $\mathcal{X} = \mathbb{R}$. Then, for any two empirical measures $\alpha = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$, and $\beta = \frac{1}{n}\sum_{j=1}^n \delta_{y_j}$, then we have the following simple formula:

$$\mathcal{W}_p(\alpha, \beta)^p = \frac{1}{n}\sum_{i=1}^n |x_{(i)} - y_{(i)}|^p,$$

where $x_{(i)}$ is the i-th order statistic of $\{x_i\}_{i=1}^n$, so that $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$. Thus, $\mathcal{W}_p(\alpha, \beta)$ is simply the ℓ_p norm between ordered values of α and β.

1D Case: General Measures. For any measure α on $\mathbb{R}$, let $F_\alpha : \mathbb{R} \mapsto [0, 1]$ be its CDF, so that $F_\alpha(x) = \int_{-\infty}^x d\alpha$. Let $F_\alpha^{-1} : [0, 1] \mapsto \mathbb{R} \cup \{-\infty\}$ denote its pseudo-inverse (note that the CDF can be flat even if non-decreasing):

$$F_\alpha^{-1}(t) = \arg\min_x \{x \in \mathbb{R} \cup \{-\infty\} \,:\, F_\alpha(x) \geq t\}.$$

F_α^{-1} is also called the generalized quantile function of the measure α. Then,

$$\mathcal{W}_p(\alpha, \beta)^p = \int_0^1 |F_\alpha^{-1}(t) - F_\beta^{-1}(t)|^p dt.$$

General Empirical Measures. Suppose $\mathcal{X} = \mathbb{R}^d$. Then, for any two empirical measures $\alpha = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$, and $\beta = \frac{1}{n}\sum_{j=1}^n \delta_{y_j}$, we have that:

$$\mathcal{W}_p(\alpha, \beta)^p = \min_{\sigma \in \mathrm{Perm}(n)} \sum_i \|x_i - y_{\sigma(i)}\|^p.$$

Though the minimization is over all permutations $\sigma \in \mathrm{Perm}(n)$, this may be solved in $O(n^3)$ time using the Hungarian algorithm.

Connection to L_1. 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 on grid G_i. Fix $\epsilon > 0$ and let $m = \log(2d/\epsilon)$. 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).

This result shows that, in some sense, Wasserstein distance is like a multiresolution L_1 distance.

Gaussians. Suppose $\alpha = \mathcal{N}(\mu_\alpha, \Sigma_\alpha)$, and $\beta = \mathcal{N}(\mu_\beta, \Sigma_\beta)$. Then, it can be shown that:

$$\mathcal{W}_2^2(\alpha, \beta) = \|\mu_\alpha - \mu_\beta\|_2^2 + \mathcal{B}(\Sigma_\alpha, \Sigma_\beta)^2,$$

where $\mathcal{B}$ is the so-called Bures metric between positive-definite matrices defined as:

$$\mathcal{B}(\Sigma_\alpha, \Sigma_\beta)^2 = \operatorname{tr}\left(\Sigma_\alpha + \Sigma_\beta - 2\left(\Sigma_\alpha^{1/2}\Sigma_\beta\Sigma_\alpha^{1/2}\right)^{1/2}\right),$$

where $\Sigma^{1/2}$ is the matrix square-root. $\mathcal{B}$ can be shown to be a distance on covariance matrices, and moreover $\mathcal{B}^2$ is convex with respect to both its arguments. As an instructive simple case, when $\Sigma_\alpha = \operatorname{diag}(\mathbf{r})$, and $\Sigma_\beta = \operatorname{diag}(\mathbf{s})$ are diagonal matrices, then

$$\mathcal{B}(\Sigma_\alpha, \Sigma_\beta) = \|\mathbf{r} - \mathbf{s}\|_2.$$

For 1D Gaussians, $\mathcal{W}_2$ is thus simply the 2D Euclidean distance between the means and standard-deviations of the two distributions.

Empirical and General Measures; $\mathcal{W}_2$. Suppose $\mathcal{X} = \mathbb{R}^d$. Suppose that one of the distributions P has density p and the other $Q = \sum_{j=1}^m q_j \delta_{y_j}$ 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

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

Entropic Regularization. Cuturi (2013) showed that if we replace $\inf_{(X,Y)} \mathbb{E}_{(X,Y)}\|X - Y\|^p$ with the regularized version $\inf_{(X,Y)} \mathbb{E}_{(X,Y)}\|X - Y\|^p + \epsilon I(X, Y)$, where $I(X, Y) = \mathrm{KL}(\pi, \alpha \times \beta)$, 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.

6 Dual Characterization of Wasserstein Distance

The Kantorovich dual for Wasserstein distances can be further simplified for $p \in (0, 1]$ where $d(x, y)^p$ is itself a distance, and satisfies the triangle inequality. Under such a setting, it can

be shown that the set

$$\{(f, -f) \,:\, \mathrm{Lip}_p(f) \leq 1\}$$

suffice as a set of Kantorovich potentials, where

$$\mathrm{Lip}_p(f) = \sup_{x,y \in \mathcal{X},\, x \neq y} \left\{ \frac{|f(x) - f(y)|}{d(x,y)^p} \right\}.$$

This thus entails that the Kantorovich dual for Wasserstein distances can be written as:

$$\mathcal{W}_p(\alpha, \beta)^p = \max_{f \,:\, \mathrm{Lip}_p(f) \leq 1} \int_{\mathcal{X}} f(x)(d\alpha(x) - d\beta(x)).$$

It can thus be seen that $\mathcal{W}_p(\alpha, \beta)$ can be written as a dual norm of $(\alpha - \beta)$. We will discuss more about such dual norm based distances over probability measures, known as integral probability metrics in the next section.

6.1 Examples

Note that in all the examples below, we assume that $p \in (0, 1]$. The most common setting of p in this setting is likely $p = 1$, with $\mathcal{W}_1$ as the corresponding Wasserstein distance.

Discrete Measures. Suppose $\alpha - \beta = \sum_k \mathbf{m}_k \delta_{z_k}$, for $z_k \in \mathcal{X}$. When α and β are probability measures, note that we have that $\sum_k \mathbf{m}_k = 0$. We then have that:

$$\mathcal{W}_p(\alpha, \beta)^p = \max_{\mathbf{f}} \left\{ \sum_k \mathbf{f}_k \mathbf{m}_k \,:\, \forall (k, \ell),\, |\mathbf{f}_k - \mathbf{f}_\ell| \leq d(z_k, z_\ell)^p \right\}.$$

When $\mathcal{X} = \mathbb{R}$, we can further reduce the number of constraints by ordering the support points $\{z_k\}$ as $z_1 \leq z_2 \leq \ldots$, and obtain that:

$$\mathcal{W}_p(\alpha, \beta)^p = \max_{\mathbf{f}} \left\{ \sum_k \mathbf{f}_k \mathbf{m}_k \,:\, \forall k,\, |\mathbf{f}_{k+1} - \mathbf{f}_k| \leq d(z_{k+1}, z_k)^p \right\}.$$

Euclidean Spaces. When $\mathcal{X} = \mathcal{Y} = \mathbb{R}^d$, and $d(x, y) = \|x - y\|_2$, and for $p = 1$, the global Lipshitz constraint on the Kantorovich potential can be reduced to a uniform bound on the gradient, so that:

$$\mathcal{W}_1(\alpha, \beta) = \max_{f : \|\nabla f\|_\infty \leq 1} \int_{\mathbb{R}^d} f(x)(d\alpha(x) - d\beta(x)).$$

7 Other Distribution Divergences

We now briefly contrast the Wasserstein distance with two other classical classes of divergences over distributions.

7.1 f-divergence

Let $f : \mathbb{R} \mapsto \mathbb{R} \cup \{\infty\}$ be convex, with some additional regularity properties that it be: lower semi-continuous, with domain $\mathrm{dom}(f) \subset [0, \infty)$ that is also non-trivial so that $\mathrm{dom}(f) \cap (0, \infty) \neq \emptyset$. Such a function is also called an entropy function. Given two measures $\alpha, \beta \in \mathcal{M}(\mathcal{X})$ such that α is absolutely continuous with respect to β, the f-divergence D_f between them is defined as:

$$D_f(\alpha, \beta) = \int_{\mathcal{X}} f\left(\frac{d\alpha}{d\beta}\right) d\beta.$$

When α and β are both absolutely continuous with respect to a base measure μ, with corresponding densities ρ_α and ρ_β, we can write this as:

$$D_f(\alpha, \beta) = \int_{\mathcal{X}} f\left(\frac{\rho_\alpha(x)}{\rho_\beta(x)}\right) \rho_\beta(x) d\mu(x).$$

Most divergences used in machine learning are instances of the class of f-divergences, including KL, Hellinger, total variation, and chi-squared divergences.

There are three main caveats to this class of divergences. The first is the requirement of absolute continuity of α with respect to β, without which these divergences are not bounded (or even well-defined). This precludes their use with singular distributions with support restricted to some manifold, or even discrete measures. The second caveat, though this is subjective, is that they induce a geometry over divergences that is much more complex and might not be exactly what we want. We discussed some examples in the introductory section; we can also look at the simple example of a pair of 1D Gaussian distributions, $\alpha = \mathcal{N}(\mu_\alpha, \sigma_\alpha^2)$, and $\beta = \mathcal{N}(\mu_\beta, \sigma_\beta^2)$. Then f-divergences such as KL induce a hyperbolic geometry over the space of Gaussian parameters $(m, \sqrt{\sigma})$. For instance, it can be shown that

$$\mathrm{KL}(\alpha, \beta) = \frac{1}{\sigma_\beta^2}\left(\frac{1}{2}(\mu_\alpha - \mu_\beta)^2 + (\sqrt{\sigma_\beta} - \sqrt{\sigma_\alpha})^2\right) + o((\mu_\alpha - \mu_\beta)^2, (\mu_\alpha - \mu_\beta)^2).$$

While the Wasserstein distance is simply associated with the Euclidean geometry over $(m, \sqrt{\sigma})$, since

$$\mathcal{W}_2(\alpha, \beta)^2 = (\mu_\alpha - \mu_\beta)^2 + (\sqrt{\sigma_\beta} - \sqrt{\sigma_\alpha})^2.$$

Lastly, these divergences are not always easy to approximate given samples. We will discuss more about statistical rates in a following section.

7.2 Integral Probability Metrics

The Kantorovich dual form of $\mathcal{W}_1$ is a special instance of a larger class of dual norm based divergences, called integral probability metrics.

Suppose B is a symmetric, convex set of measurable functions. Then, we can define the following dual norm:

$$\|\alpha\|_B = \max_{f \in B} \int_{\mathcal{X}} f(x) d\alpha(x),$$

which can then be used to define a metric over measures as $\|\alpha - \beta\|_B$.

The total variation norm is an instance of this dual norm class, with

$$B = \{f \in \mathcal{C}(\mathcal{X}) \,:\, \|f\|_\infty \leq 1\}.$$

As discussed earlier, $\mathcal{W}_1$ is an instance of such a dual norm with

$$B = \{f \,:\, \mathrm{Lip}(f) \leq 1\}.$$

A caveat with the $\mathcal{W}_1$ norm is that it requires $\int d\alpha = 0$, otherwise $\|\alpha\|_B = \infty$. To fix this, we can bound the value of the Kantorovich potential f, similar to TV.

One approach towards this, the flat norm, sometimes also called the Kantorovich-Rubinstein norm, uses:

$$B = \{f \,:\, \|\nabla f\|_\infty \leq 1,\ \|f\|_\infty \leq 1\}.$$

A related norm, corresponding to the Dudley metric, uses:

$$B = \{f \,:\, \|\nabla f\|_\infty + \|f\|_\infty \leq 1\}.$$

7.2.1 Maximum Mean Discrepancy, Dual RKHS Norms

Given an RKHS $\mathcal{H}$, a natural class of functions to specify the dual norm is simply:

$$B = \{f \in \mathcal{H} \,:\, \|f\|_{\mathcal{H}} \leq 1\}.$$

Given the kernel $k : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ corresponding to the RKHS $\mathcal{H}$, this can be simplified to:

$$\|\alpha\|_{\mathcal{H}_k} = \int_{\mathcal{X} \times \mathcal{X}} k(x, y) d\alpha(x) d\alpha(y),$$

which have been called "Maximum Mean Discrepancy" or "kernel norms". The above expression can also be written more compactly as:

$$\|\alpha\|_{\mathcal{H}_k} = \mathbb{E}_{X, X' \sim \alpha}[k(X, X')].$$

8 Empirical Estimators, Convergence Rates

Suppose we are given n samples $\{x_i\}_{i=1}^n$ drawn iid from α, and m samples $\{y_j\}_{j=1}^m$ drawn iid from β. How do we estimate some specified divergence $D(\alpha, \beta)$ given samples?

Let $\widehat{\alpha}_n = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$, and $\widehat{\beta}_n = \frac{1}{m}\sum_{j=1}^m \delta_{y_j}$. Then, a natural estimator of $D(\alpha, \beta)$ is simply $D(\widehat{\alpha}_n, \widehat{\beta}_m)$.

However, for TV distance, and indeed most f-divergences, $D(\widehat{\alpha}_n, \widehat{\beta}_m)$ does not converge to $D(\alpha, \beta)$. For instance, $\|\widehat{\alpha}_n - \widehat{\beta}_n\|_{TV} = 2$ with probability 1 since the supports of the two discrete measures will likely not overlap. One needs to devise careful smoothing of these empirical measures, which forms a large body of work on non-parametric and parametric estimation of distributions.

But it turns out for Wasserstein distances, just using the empirical distributions suffices.

For $\mathcal{X} = \mathbb{R}^d$, and where the measures α, β have support in a bounded set, (Dudley, 1969) showed that for $d \geq 2$, and $1 \leq p < \infty$,

$$\mathbb{E}|\mathcal{W}_p(\widehat{\alpha}_n, \widehat{\beta}_n) - \mathcal{W}_p(\alpha, \beta)| = O(n^{-1/d}).$$

It is also possible to prove concentration of $\mathcal{W}_p(\widehat{\alpha}_n, \widehat{\beta}_n)$ around its mean $\mathcal{W}_p(\alpha, \beta)$ (Weed and Bach, 2017).

For MMD distances, there is the interesting result that the rate does not depend on the ambient dimension (Sriperumbudur et al., 2012):

$$\mathbb{E}[\|\widehat{\alpha}_n - \widehat{\beta}_n\|_k] - \|\alpha - \beta\|_k = O(n^{-1/2}).$$

9 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 $0 = t_1 < \cdots < t_m = 1$.

A geodesic connecting P_0 and P_1 is a path $(P_t : 0 \leq t \leq 1)$ as above that also satisfies $L(c) = \mathcal{W}_p(P_0, P_1)$. It can be shown that such a path always exists, and moreover, that for this geodesic path

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

where π is the optimal coupling between P_0 and P_1, and $F_t(x, y) = (1 - t)x + ty$. Examples are shown in Figures 4 and 5.

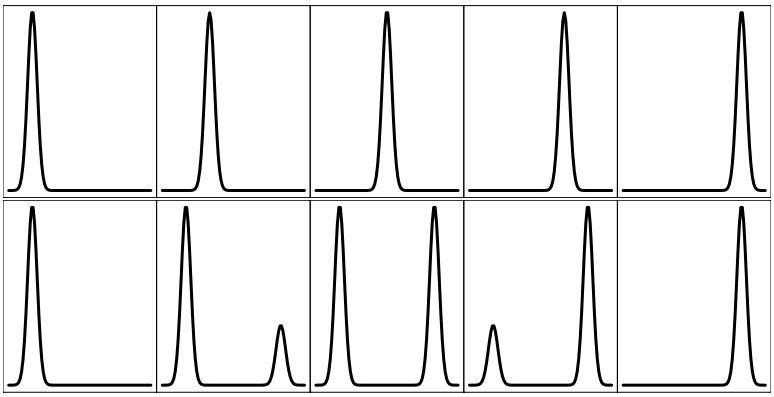

Figure 4: *Top row: Geodesic path from P_0 to P_1. Bottom row: Euclidean path from P_0 to P_1.*

Figure 5: *Morphing one image into another using the Wasserstein geodesic. Image credit: Bauer, Joshi and Modin 2015.*

10 Robustness

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

$$\tau(P, Q) = \inf_A \mathcal{W}_2(P_A, Q_A)$$

where $P_A(\cdot) = P(A \cap \cdot)/P(A)$, $Q_A(\cdot) = Q(A \cap \cdot)/Q(A)$ 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 $1 - \alpha$.

11 Optimal Transport for Classification

Consider the task of binary classification, with response $Y \in \{0, 1\}$, and input $X \in \mathcal{X}$, and where $(X, Y) \sim P$. The Bayes risk, that is the risk of the Bayes optimal classifier, with respect to the zero-one loss is given by:

$$\inf_f P(f(X) \neq Y).$$

Suppose $P(Y = 1) = P(Y = 0) = 1/2$. Note that:

$$\begin{aligned} P(f(X) = Y) &= P(f(X) = 1|Y = 1)P(Y = 1) + P(f(X) = 0|Y = 0)P(Y = 0) \\ &= \frac{1}{2} + \frac{1}{2}(P(f(X) = 1|Y = 1) - P(f(X) = 1|Y = 0)) \\ &= \frac{1}{2} + \frac{1}{2}(P_{X|Y=1}(A_f) - P_{X|Y=0}(A_f)), \end{aligned}$$

where $A_f = \{x \in \mathcal{X} : f(x) = 1\}$. It can thus be seen that:

$$\begin{aligned} \sup_f P(f(X) = Y) &= \frac{1}{2} + \frac{1}{2} \sup_{A \subseteq \mathcal{X}} (P_{X|Y=1}(A) - P_{X|Y=0}(A)) \\ &= \frac{1}{2} + \frac{1}{2} \mathrm{TV}(P_{X|Y=1}, P_{X|Y=0}), \end{aligned}$$

so that

$$\begin{aligned}\inf_f P(f(X) \neq Y) &= \frac{1}{2} - \frac{1}{2}\mathrm{TV}(P_{X|Y=1}, P_{X|Y=0}) \\ &= \frac{1}{2} - \frac{1}{2}\mathcal{L}_{c_{0/1}}(P_{X|Y=1}, P_{X|Y=0})\end{aligned}$$

where $c_{0/1}(x, x') = \mathbb{I}(x \neq x')$. Thus, the misclassification error of the Bayes classifier has a closed form expression in terms of the TV distance, which is the optimal transport distance under the zero-one cost function, between $P_{X|Y=1}$ and $P_{X|Y=0}$. Note that when the TV distance is one, the supports are disjoint, and hence are perfectly separable, and consequently the Bayes risk is zero. And when the TV distance is zero, so that the two conditional distributions are indistinguishable, then the Bayes risk is $1/2$. The identity above is sometimes called the Lecam method identity for binary classification, which Lecam used in turn provide information-theoretic lower bounds for statistical estimation (by reducing those to binary and multi-class classification).

This can be extended to adversarial ML settings. Here, at test time, the input x could be adversarially moved to a nearby $x' \in B(x)$, for some ball $B(x)$ around x, where the aim of the adversary is to get the classifier to change its decision even with such a small perturbation. The aim of robust classification then is to learn classifiers that cannot be manipulated by such adversaries.

A common adversarial risk is given by:

$$\mathbb{E}_{(X,Y)\sim P} \sup_{X' \in B(X)} P(f(X') \neq Y).$$

What is the Bayes optimal classifier with respect to the robust classification objective above? (Bhagoji et al, 2019) showed that Lecam method identity could be extended to this robust setting with the slight modification of the cost from the zero-one cost to the robust cost:

$$c_r(x, x') = \mathbb{I}(B(x) \cap B(x') = \emptyset),$$

which reduces to the standard zero-one loss when $B(x) = \{x\}$. They then showed that:

$$\inf_f \mathbb{E}_{(X,Y)\sim P} \sup_{X' \in B(X)} P(f(X') \neq Y) = \frac{1}{2} - \frac{1}{2}\mathcal{L}_{c_r}(P_{X|Y=1}, P_{X|Y=0}),$$

which they then used to provide lower bounds on robust misclassification error for simple parametric classes such as Gaussian distributions for $P_{X|Y=1}$ and $P_{X|Y=0}$.

12 Optimal Transport for Clustering

A critical subtask in clustering, and also an important problem in its own right, is to compute the "mean" or "barycenter" of a set of data points. Given $\{x_j\}_{j=1}^m$ each lying in some space

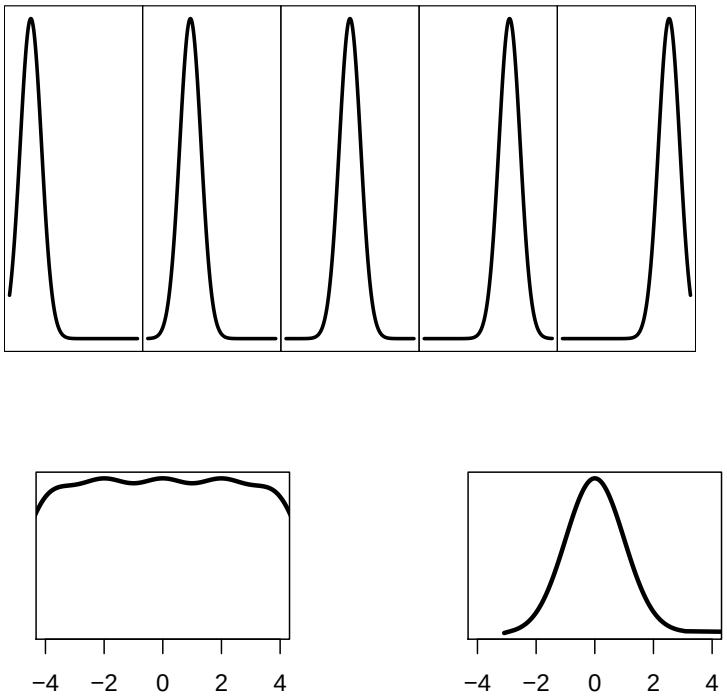

Figure 6: *Top: Five distributions. Bottom left: Euclidean average of the distributions. Bottom right: Wasserstein barycenter.*

$\mathcal{X}$ with metric d, a general weighted barycenter is defined as:

$$\min_{x \in \mathcal{X}} \sum_j \lambda_j \, d(x, x_j)^p,$$

for a given set of weights $\{\lambda_j\}_{j=1}^n$, which are typically just set to 1, and where p is typically set to 2. When $\mathcal{X} = \mathbb{R}^d$, and $d(x, y) = \|x - y\|_2$, this simply leads to the sample mean $\sum_j \lambda_j x_j / (\sum_j \lambda_j)$ for $p = 2$, and the sample median for $p = 1$.

How could we extend this concept to a set of measures? Suppose $\{\alpha_j\}_{j=1}^n$ are a set of measures defined on some input space $\mathcal{X}$. We could take the average $\frac{1}{n} \sum_{j=1}^n \alpha_j$. But the resulting average won't look like any of the α_j's. See Figure 6.

The optimal transport based barycenter is on the other hand defined as:

$$\min_{\alpha \in \mathcal{M}(\mathcal{X})} \sum_j \lambda_j \, \mathcal{L}_c(\alpha, \beta_j).$$

A natural choice for the cost function is $c = d^2$, so that we would obtain the $\mathcal{W}_2$ Wasserstein barycenter. The bottom right plot of Figure 6 shows an example. You can see that this does a much better job.

The same holds for empirical measures (regarding datasets as empirical measures). See Figure 7. The average (red dots) $n^{-1} \sum_j \widehat{\alpha}_j$ of these empirical distributions $\widehat{\alpha}_j$ is useless.

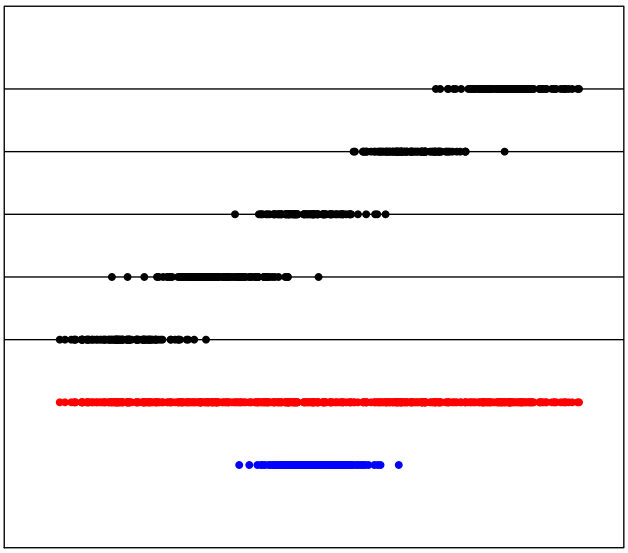

Figure 7: *The top five lines show five, one-dimensional datasets. The red points the what happens if we simple average the give empirical distributions. The blue dots show the Wasserstein barycenter which, in this case, can be obtained simply by averaging the order statistics.*

But the Wasserstein barycenter (blue dots) gives us a better sense of what a typical dataset looks like.

12.1 KMeans via Optimal Transport

Suppose we are given a single empirical measure $\beta = \sum_{i=1}^{n} \delta_{x_i}$, where the input space $\mathcal{X} = \mathbb{R}^d$ endowed the Euclidean distance metric. Then, letting $\mathcal{M}_k(\mathcal{X})$ denote distributions with finite support of size upto k, the k-means problem can be seen to be solving:

$$\min_{\alpha \in \mathcal{M}_k(\mathcal{X})} \mathcal{W}_2(\alpha, \beta).$$

The support of the solution α are the centroids of k-means, and its weights correspond to the fraction of points assigned to the corresponding centroid.

13 Optimal Transport for Statistical Estimation

In the general setup of statistical estimation, we are given n samples $\{x_i\}_{i=1}^{n} \subset \mathcal{X}$ drawn from some unknown distribution β, and the goal is to fit a parametric model $\alpha_\theta \in \mathcal{M}(\mathcal{X})$ to

the observed empirical measure $\beta_n = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$:

$$\min_{\theta\in\Theta} \mathcal{L}(\alpha_\theta, \beta_n),$$

where $\mathcal{L}$ is some suitable loss function. When α_θ has density ρ_θ with respect to the Lebesgue measure, a classical approach is to use the MLE which optimizes the negative log-likelihood:

$$\mathcal{L}_{\text{MLE}}(\alpha_\theta, \beta_n) = -\sum_{i=1}^n \log(\rho_\theta(x_i)),$$

which as we saw in the decision theory lecture is a sample estimate of the KL divergence $\text{KL}(\beta, \alpha_\theta)$ (modulo a constant).

But KL is not suited for settings where the true distribution β is singular, or when the statistical model α_θ either does not have a density, or the density is not available in closed form or is otherwise inaccessible (e.g. due to a difficult to estimate log-partition function/normalization constant).

In such settings, it is useful to use dual norm based distances instead:

$$\mathcal{L}(\alpha_\theta, \beta) = \max_{f\in B} \left\{ \int_X f(x) d\alpha_\theta(x) - \int_{\mathcal{X}} f(x) d\beta(x) \right\},$$

where B is some set of functions. A natural choice is to use $B = \{f : \text{Lip}(f) \leq 1\}$, which as we saw earlier yields the Wasserstein divergence $\mathcal{W}_1$. Unlike KL and similar divergences, these are suited to singular true distributions, and moreover do not require that α_θ have a readily available density so long as we can optimize the variational problem above.

Thus, we get the following class of estimators as natural extensions of the classical MLE:

$$\widehat{\theta} \in \arg\min_\theta \mathcal{L}(\alpha_\theta, \beta_n).$$

For instance, in Wasserstein GANs (Arjovsky et al., 2017), they suggest using using the $\mathcal{W}_1$ divergence, and the class of distributions $\alpha_\theta = h_{\theta\#}\gamma$, where γ is some base distribution such as a standard Gaussian distribution, and h_θ is some flexible parametrization such as a deep neural network.

In their case, they then solve for:

$$\min_\theta \max_{f\in B} \left\{ \mathbb{E}_{X\sim\beta_n}[f(X)] - \mathbb{E}_{X\sim\alpha_\theta}[f(X)] \right\}.$$

Letting f_θ be the optimal Kantorovich potential for a given α_θ, by an application of an envelope theorem (Milgrom, Segal, 2002), and the reparameterization trick, it follows that the gradient of the objective with respect to θ can be written as

$$\nabla_\theta \mathcal{L}(\beta_n, \alpha_\theta) = -\mathbb{E}_{Z\sim\gamma} \nabla_\theta f_\theta(h_\theta(Z)),$$

which thus facilitates easily optimization.

In some cases we do not even have access to a parametric generative model for α_θ. Instead, we can simulate from it. This happens quite often, for example, in astronomy and climate science. In such settings, we can again replace the MLE with minimum Wasserstein distance, as was suggested in such contexts by Berntom et al (2017). That is, given data $\{x_i\}_{i=1}^n \sim \beta$, we can solve for:

$$\widehat{\theta} = \operatorname*{argmin}_\theta \mathcal{W}(\alpha_\theta, \beta_n),$$

where we approximate α_θ via the empirical measure $\alpha_{\theta,m} = \frac{1}{m}\sum_{j=1}^m \delta_{z_j}$, where $\{z_j\}_{j=1}^m \sim \alpha_\theta$.

14 Optimal Transport for Domain Adaptation

An interesting and natural application of optimal transport is to domain adaptation. 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.

15 Summary, References

Wasserstein distance has many nice properties and has become popular in statistics and machine learning. But the distance does have problems. First, it is hard to compute. Second, the distance is not a smooth functional which is not a good thing. We have also seen that the distance is not robust although the trimmed version may fix this.

Three good references for this topic are:

Peyre, Gabriel, and Marco Cuturi. Computational optimal transport. Foundations and Trends in Machine Learning 11.5-6 (2019): 355-607.

Kolouri, Soheil, Se Rim Park, Matthew Thorpe, Dejan Slepcev, and Gustavo K. Rohde. 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.