1 Basics

1.1 Optimal transport

Let X and Y be two Radon spaces and take $c : X \times Y \to [0, \infty]$ be a Borel-measurable function (with $c(x, y)$ indicating the cost of transportation from x to y). Given probability measures μ on X and ν on Y, the Kantorovich formulation of the optimal transportation problem seeks to find the measure ν on $X \times Y$ achieving the infimum

$$\inf \left\{ \int_{X \times Y} c(x, y) d\pi(x, y) \,\middle|\, \pi \in \Pi(\mu, \nu) \right\},$$

where $\Pi(\mu, \nu)$ denotes the set of all couplings of ν and μ. The existence of such a ν is guaranteed if c is lower semi-continuous. Often, we use the dual form of this problem given by

$$\sup \left(\int_X \varphi(x) \, \mathrm{d}\mu(x) + \int_Y \psi(y) \, \mathrm{d}\nu(y) \right),$$

where the supremum runs over all pairs of bounded and continuous functions $\varphi : X \to \mathbb{R}$ and $\psi : Y \to \mathbb{R}$ such that

$$\varphi(x) + \psi(y) \leq c(x, y).$$

See `https://en.wikipedia.org/wiki/Transportation_theory_(mathematics)` for more details.

1.2 Wasserstein metric

Let (M, d) be a Radon space. For $p \geq 1$, let $\mathcal{P}_p(M)$ denote the collection of all probability measures μ on M with finite pth moment, i.e. those μ for which there exists some $x_0 \in M$ such that

$$\int_M d(x, x_0)^p \, \mathrm{d}\mu(x) < \infty.$$

The pth Wasserstein distance between two probability measures μ and ν in $\mathcal{P}_p(M)$ is defined as

$$W_p(\mu, \nu) := \left(\inf_{\pi \in \Pi(\mu, \nu)} \int_{M \times M} d(x, y)^p \, \mathrm{d}\pi(x, y) \right)^{1/p}.$$

Equivalently, we have

$$W_p(\mu, \nu) = (\inf \mathbb{E}[d(X, Y)^p])^{1/p},$$

where the infimum is taken over all X and Y whose joint distribution is a coupling of μ and ν. Letting $\mathrm{Lip}_1(M)$ denote the space of all real functions on M with Lipchitz smoothness at most 1, we have a more specific duality result.

Theorem 1 (Kantorovich-Rubinstein duality)**.** *Let (M, d) be a Radon space and fix $\mu, \nu \in P_1(M)$. Then,*

$$W_1(\mu, \nu) = \sup_{f \in \mathrm{Lip}_1(M)} \left\{ \int_M f \, \mathrm{d}\mu - \int_M f \, \mathrm{d}\nu \right\}.$$

Proof. From the more general dual form, we find that

$$W_1(\mu, \nu) = \sup \left(\int_M f \, \mathrm{d}\mu + \int_M g \, \mathrm{d}\nu \right)$$

over bounded, continuous f and g with $f(x) + g(y) \leq d(x, y)$. Thus, for each $\varepsilon > 0$, there exist such f and g with

$$W_1(\mu, \nu) - \varepsilon \leq \int_M f \, \mathrm{d}\mu + \int_M g \, \mathrm{d}\nu.$$

Next, define $h : M \to \mathbb{R}$ by $h(x) = \inf_{y \in M}(d(x, y) - g(y))$, which is well defined by our boundedness assumption. Note that

$$\begin{aligned} |h(x) - h(x')| &= \left| \inf_{y \in M}(d(x, y) - g(y)) - \inf_{y \in M}(d(x', y) - g(y)) \right| \\ &\leq \sup_{y \in M} |d(x, y) - d(x', y)| \leq d(x, x'), \end{aligned}$$

so $h \in \mathrm{Lip}_1(M)$. Also, by design, $f \leq h \leq -g$ pointwise. Taking $\pi \in \Pi(\mu, \nu)$ to be a coupling of μ and ν, we have

$$\begin{aligned} W_1(\mu, \nu) - \varepsilon &\leq \int_M f \, \mathrm{d}\mu + \int_M g \, \mathrm{d}\nu \\ &\leq \int_M h \, \mathrm{d}\mu - \int_M h \, \mathrm{d}\nu \\ &\leq \sup_{f \in \mathrm{Lip}_1(M)} \left\{ \int_M f \, \mathrm{d}\mu - \int_M f \, \mathrm{d}\nu \right\} \\ &= \sup_{f \in \mathrm{Lip}_1(M)} \left\{ \int_{M \times M} (f(x) - f(y)) \, \mathrm{d}\pi(x, y) \right\} \\ &\leq \int_{M \times M} d(x, y) \, \mathrm{d}\pi(x, y), \end{aligned}$$

from which the theorem follows. □

Proposition 2. *$(\mathcal{P}_p, W_p)$ is a metric space.*

Proposition 3. *Convergence with respect to W_p is equivalent to weak convergence, plus convergence of pth moment.*

See `https://en.wikipedia.org/wiki/Wasserstein_metric` and `http://n.ethz.ch/~gbasso/download/A%20Hitchhikers%20guide%20to%20Wasserstein/A%20Hitchhikers%20guide%20to%20Wasserstein.pdf` for more details.

1.3 Gaussian-smoothed Wasserstein metric

In what follows, we will restrict ourselves to Borel probability distributions over $\mathbb{R}^d$, and we will denote the set of such measures with finite pth moments as $\mathcal{P}_p(\mathbb{R}^d)$. We will let $\mathcal{N}_\sigma$ denote the standard normal distribution with mean 0 and standard deviation σ, with corresponding probability density function φ_σ. We define the smoothed Wasserstein distance W_p^σ by

$$W_p^\sigma(\nu, \mu) := W_p(\nu * \mathcal{N}_\sigma, \mu * \mathcal{N}_\sigma).$$

Proposition 4. *W_p^σ is a metric on $\mathcal{P}_p(\mathbb{R}^d)$.*

Proof. The fact that $W_p^\sigma(\nu, \mu)$ is symmetric, non-negative, and equals zero for $\nu = \mu$ follows from the definition. Now, fix $\mu_1, \mu_2, \mu_3 \in \mathcal{P}_p(\mathbb{R}^d)$. Let $\pi_{12} \in \Pi(\mu_1 * \mathcal{N}_\sigma, \mu_2 * \mathcal{N}_\sigma)$ be the smoothed optimal coupling of μ_1 and μ_2, and let $\pi_{23} \in \Pi(\mu_2 * \mathcal{N}_\sigma, \mu_3 * \mathcal{N}_\sigma)$ be the optimal coupling of μ_2 and μ_3 (existence is guaranteed because metrics are continuous). Then, we can use the gluing lemma to construct a measure $\pi \in \mathcal{P}_p(\mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^d)$ with π_{12} and π_{23} as marginals in the natural way. Then, defining $\pi_{13} \in \Pi(\mu_1, \mu_3)$ by $\pi_{13}(A \times B) = \pi(A \times \mathbb{R}^d \times B)$, we have

$$\begin{aligned}
W_p^\sigma(\mu_1, \mu_3) &\le (\mathbb{E}_{\pi_{13}} \|X_1 - X_3\|^p)^{1/p} = (\mathbb{E}_\pi \|X_1 - X_3\|^p)^{1/p} \\
&\le (\mathbb{E}_\pi \|X_1 - X_2\|^p)^{1/p} + (\mathbb{E}_\pi \|X_2 - X_3\|^p)^{1/p} \\
&= (\mathbb{E}_{\pi_{12}} \|X_1 - X_2\|^p)^{1/p} + (\mathbb{E}_{\pi_{23}} \|X_2 - X_3\|^p)^{1/p} \\
&= W_p^\sigma(\mu_1, \mu_2) + W_p^\sigma(\mu_2, \mu_3).
\end{aligned}$$

Finally, suppose that $W_p^\sigma(\mu, \nu) = 0$. Then $\mu * \mathcal{N}_\sigma = \nu * \mathcal{N}_\sigma$ (since W_p is a metric), and so $\phi_\mu \phi_{\mathcal{N}_\sigma} = \phi_\nu * \phi_{\mathcal{N}_\sigma}$. Since $\phi_{\mathcal{N}_\sigma} \neq 0$ everywhere, we get $\phi_\nu = \phi_\mu$ pointwise, so $\nu = \mu$. □

In fact, this proof generalizes to any noise model $\mathcal{M}_\sigma$ for which $\phi_{\mathcal{M}_\sigma}$ is zero. A sufficient condition for this is infinite divisibility, i.e. that the noise can be expressed as a sum of an arbitrary number of i.i.d variables. This includes stable distributions but excludes distributions with bounded support.

See `http://people.ece.cornell.edu/zivg/GOT_AISTATS2020.pdf` for more details.

1.3.1 Smoothed W_1 metric

We have

$$\begin{aligned}
W_1^\sigma(\mu, \nu) &= W_1(\mu * \mathcal{N}_\sigma, \nu * \mathcal{N}_\sigma) \\
&= \sup_{f \in \mathrm{Lip}_1(\mathbb{R}^d)} \mathbb{E}_{\mu * \mathcal{N}_\sigma} f - \mathbb{E}_{\nu * \mathcal{N}_\sigma} f \\
&= \sup_{f \in \mathrm{Lip}_1(\mathbb{R}^d)} \mathbb{E}_\mu f * \varphi_\sigma - \mathbb{E}_\nu f * \varphi_\sigma \\
&\approx \sup_{\substack{\theta \in \Theta \\ f_\theta \in \mathrm{Lip}_1(\mathbb{R}^d)}} \mathbb{E}_\mu f_\theta * \varphi_\sigma - \mathbb{E}_\nu f_\theta * \varphi_\sigma,
\end{aligned}$$

for some parameterization of Lipschitz-1 functions $\{f_\theta\}_{\theta\in\Theta}$. (note: does equality 2 need any conditions on measures, or can I take a limit?) We have a closed form for neural networks with a single hidden layer using group sort activation

Another perspective is that

$$W_1^\sigma(P,Q) = \sup_{g\in\mathcal{F}_\sigma} \mathbb{E}_\mu g - \mathbb{E}_\nu g,$$

where $\mathcal{F}_\sigma = \{f * \varphi_\sigma \mid f \in \mathrm{Lip}_1(\mathbb{R}^d)\}$. This supremum domain is more well-behaved in some sense (Hölder ball?) than $\mathrm{Lip}_1(\mathbb{R}^d)$.

1.3.2 Emperical approximation with smoothed W_1 metric

In the non-smooth case, we have

$$\mathbb{E}[W_1(\hat{P}_n, P)] \lesssim \begin{cases} n^{-1/2}, & d = 1 \\ \frac{\log n}{\sqrt{n}}, & d = 2 \\ n^{-1/d}, & d \geq 3. \end{cases}$$

These are asymptotically tight, except for the second, which has some wiggle room (how much?). Thus, for $d = 1$, we have

$$\sqrt{n}\mathbb{E}W_1(\hat{P}_n, P) \to \text{const.}$$

A natural question is to find the limiting distribution of $\sqrt{n}W_1(\hat{P}_n, P)$ (for each n, this is a random variable using n samples for the emperical estimate). I believe we have a description of this case (find this, also why couldn't we then get the expectation). For $d \geq 3$, I think this is unknown (verify). The reason that the one-dimensional case is more straightforward is that

$$W_1(P,Q) = \|F - G\|_{L^1(\mathbb{R})},$$

where F and G are the corresponding CDFs. This follows (in one direction) because

$$\|F - G\|_{L^1(\mathbb{R})} = \|F^{-1} - G^{-1}\|_{L^1([0,1])}.$$

(This describes a greedy algorithm, sort of "pouring" one distribution into the other.)

In the smooth case, however, we have

$$\mathbb{E}[W_1^\sigma(\hat{P}_n, P)] \approx n^{-1/2}$$

for all $d \geq 1$. More generally, we have

$$\sqrt{n}W_1^\sigma(P_n, P) \xrightarrow{d} \sup_{f\in\mathrm{Lip}_1(\mathbb{R}^d)} G_P(f)$$

where $\{G_P(f)\}_{f\in\mathrm{Lip}}$ is a tight Gaussian process over $\ell^\infty(\mathrm{Lip}_2(\mathbb{R}^d))$ with $\mathbb{E}G_P(f) = 0$ and

$$\mathrm{Cov}(G_P(f), G_P(g)) = \int (f * \varphi_\sigma)(g * \varphi_\sigma) dP.$$

[todo: understand this]

1.4 Smooth Wasserstein Barycenter

As we have seen

$$W_2^\sigma(P,Q) = W_2(P * \mathcal{N}_\sigma, Q * \mathcal{N}_\sigma)$$

defines a metric which metrizes weak convergence.

[Check whether this is true or conjecture:]

$$\mathbb{E}[W_2^\sigma(\hat{P}_n, P)^2] \approx \frac{1}{n}$$

for P sub-Gaussian with constant $K \leq f(\sigma)$

The barycenter problem is defined as follows. Given a set of probability distributions $P_1, \ldots, P_k \in \mathcal{P}_2(\mathbb{R}^d)$, find Q minimizing

$$\sum_{i=1}^{k} W_2^2(P_i, Q).$$

Perhaps smoothed W_2 distance would have an advantage (from a statistical perspective) if we are given samples from the distributions. Then, we can look at the sequence of barycenters from the emperical distributions and see how quickly they converge to the true barycenter. Hopefully, this will have a statistical edge over the entropic optimal transport framework, which has strong convexity in its favor.

1.5 f-Divergences

Let P and Q be two probability distributions over Ω such that P is absolutely continuous with respect to Q. Then, for a convex function f such that $f(1) = 0$, the f-divergence of P from Q is defined as

$$D_f(P \parallel Q) := \int_\Omega f\left(\frac{\mathrm{d}P}{\mathrm{d}Q}\right) \mathrm{d}Q.$$

If P and Q are both absolutely continuous with respect to a reference distribution μ on Ω with probability density functions p and q, then

$$D_f(P \parallel Q) = \int_\Omega f\left(\frac{p(x)}{q(x)}\right) q(x)\, \mathrm{d}\mu(x).$$

In general, by Jensen's inequality,

$$D_f(P \parallel Q) = \int_\Omega f\left(\frac{\mathrm{d}P}{\mathrm{d}Q}\right) \mathrm{d}Q \geq f\left(\int_\Omega \frac{\mathrm{d}P}{\mathrm{d}Q}\, \mathrm{d}Q\right) = f(1) = 0,$$

with equality if and only if $P = Q$.

A standard f-divergence is the KL-divergence, given by $f = t \log t$.

$$D_{KL}(P \parallel Q) = \int_\Omega \frac{\mathrm{d}P}{\mathrm{d}Q} \log\left(\frac{\mathrm{d}P}{\mathrm{d}Q}\right) \mathrm{d}Q = \int_\Omega \log\left(\frac{\mathrm{d}P}{\mathrm{d}Q}\right) \mathrm{d}P$$

Also common is total variation distance, given by $f = |t-1|/2$.

$$D_{TV}(P \parallel Q) = \frac{1}{2}\int_\Omega d|P-Q| = \frac{1}{2}\|f-g\|_1,$$

where the last case holds if P and Q are both dominated by the measure with respect to which the norm is taken and have corresponding pdfs are f and g.

Some downsides:

$$\begin{aligned} D_{KL}(\mu \parallel \nu) &= \infty \quad \text{if } \nu \not\ll \nu \\ D_{TV}(\mu \parallel \nu) &= \text{const} \quad \text{if } \operatorname{supp}(\mu) \cap \operatorname{supp}(\nu) = \emptyset \end{aligned}$$

TV distance doesn't metrize weak convergence.

See `https://en.wikipedia.org/wiki/F-divergence` for more details.

1.6 Background proofs and definitions

Proposition 5. *The characteristic function of the normal distribution $\mathcal{N}(\mu, \sigma)$ is given by*

$$\phi(t) = e^{it\mu - \frac{1}{2}\sigma^2 t^2}.$$

Proof. For the standard normal $\mathcal{N}(0,1)$, we have

$$\begin{aligned} \phi_0(t) = \mathbb{E}_{X\sim\mathcal{N}(0,1)}[e^{itX}] &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{itx}e^{-\frac{1}{2}x^2}\,\mathrm{d}x \\ &= \frac{1}{\sqrt{2\pi}}\left[\int_0^{\infty} e^{itx}e^{-\frac{1}{2}x^2}\,\mathrm{d}x - \int_0^{\infty} e^{-itx}e^{-\frac{1}{2}x^2}\,\mathrm{d}x\right] \\ &= \sqrt{\frac{2}{\pi}}\int_0^{\infty} \cos(tx)e^{-\frac{1}{2}x^2}\,\mathrm{d}x. \end{aligned}$$

Hence, we can use integration by parts to obtain

$$\begin{aligned} \phi_0'(t) &= -\sqrt{\frac{2}{\pi}}\int_0^{\infty} \sin(tx)xe^{-\frac{1}{2}x^2}\,\mathrm{d}x \\ &= \sqrt{\frac{2}{\pi}}\int_0^{\infty} \sin(tx)\,\mathrm{d}[e^{-\frac{1}{2}x^2}] \\ &= \sqrt{\frac{2}{\pi}}\left[\sin(tx)e^{-\frac{1}{2}x^2}\Big|_0^{\infty} - x\int_0^{\infty}\cos(tx)e^{-\frac{1}{2}x^2}\,\mathrm{d}x\right] \\ &= -x\phi'(t) \end{aligned}$$

With initial condition $\phi_0(0) = 1$, this gives that $\phi_0(t) = e^{-\frac{1}{2}x^2}$. Thus,

$$\phi(t) = \mathbb{E}_{X\sim\mathcal{N}(\mu,\sigma)}[e^{itX}] = \mathbb{E}_{X\sim\mathcal{N}(0,1)}[e^{it(\sigma X+\mu)}] = e^{it\mu}\phi_0(\sigma t) = e^{it\mu-\frac{1}{2}\sigma^2t^2}.$$

□

Definition 6 (Sub-Gaussian distribution)**.** *We call $P \in \mathcal{P}(\mathbb{R}^d)$ β-sub-Gaussian, for $\beta > 0$, if $X \sim P$ satisfies*

$$\mathbb{E}[\exp(\alpha \cdot (X - \mathbb{E}[X]))] \leq e^{\beta^2 \|\alpha\|^2/2}$$

for all $\alpha \in \mathbb{R}^d$.