Notes on Wasserstein distances

Preliminaries

A special case of the Disintegration Theorem we need is the Following. Throughout the talk we assume that all spaces are Polish.

Theorem: Disintegration for Product spaces

Let X, Y be Polish, $p : X \times Y \to X$ the projection and $\pi \in \mathcal{P}(X \times Y)$.
Setting $\mu := p_{\#}(\pi)$ we get the existence of a paramatrized family of probability measures $\{\pi_x\}_{x \in X} \subset \mathcal{P}(X \times Y)$ such that

1. For all $A \in \mathcal{B}(X \times Y)$ the function $x \mapsto \pi_x(A)$ is measurable.
2. $\pi(A) = \int_X \pi_x(A) d\mu(x)$ for all $A \in \mathcal{B}(X \times Y)$
3. π_x lives on $p^{-1}(x) = \{x\} \times Y$ for $\mu-$almost all $x \in X$. This means that $\pi_x((X \times Y) \setminus p^{-1}(x)) = 0$.

For the proof of the triangle inequality for the distance, we need the Gluing Lemma.

Lemma: Dudley/Gluing

Let $(X_1, \mu_1), (X_1, \mu_2), (X_1, \mu_3)$ be Polish and $\pi^{1,2} \in \Gamma(\mu_1, \mu_2)$ and $\pi^{2,3} \in \Gamma(\mu_2, \mu_3)$. Then there exists some $\pi \in \mathcal{P}(X_1 \times X_2 \times X_3)$ such that

$$p^{1,2}_{\#}(\pi) = \pi^{1,2} \quad \text{and} \quad p^{2,3}_{\#}(\pi) = \pi^{2,3}$$

where $p^{1,2}(x_1, x_2, x_3) = (x_1, x_2)$ and $p^{2,3}(x_1, x_2, x_3) = (x_2, x_3)$.

Proof:

By the Disintegration Theorem, we have

$$\pi^{1,2}(\mathrm{d}x_1, \mathrm{d}x_2) = \pi^{1,2}_{x_2}(\mathrm{d}x_1)\mu_2(\mathrm{d}x_2), \quad \pi^{2,3}(\mathrm{d}x_2, \mathrm{d}x_3) = \pi^{2,3}_{x_2}(\mathrm{d}x_3)\mu_2(\mathrm{d}x_2).$$

Which in integral notation means for any bounded f

$$\int_{X_1 \times X_2} f(x_1, x_2) d\pi^{1,2}(x_1, x_2) = \int_{X_2} \int_{X_1} f(x_1, x_2) d\pi^{1,2}_{x_1}(x_1) d\mu_2(x_2).$$

and similar for $\pi^{2,3}$.
Now, we "glue" both of those disintegrations together with μ_2. Thus we define

$$\pi := \pi^{12}_{x_2} \times \pi^{23}_{x_2}(\mathrm{d}x_1, \mathrm{d}x_3)\mu_2(\mathrm{d}x_2) \in \mathcal{P}(X_1 \times X_2 \times X_3),$$

which in integral notation now is

$$\int_{X_1 \times X_2 \times X_3} f d\pi = \int_{X_2} \int_{X_1 \times X_3} f(x_1, x_2, x_3) d(\pi^{1,2}_{x_2} \times \pi^{2,3}_{x_2})(x_1, x_3) d\mu_2(x_2)$$

Introduction to Wasserstein distance

As a reminder, we have the following space

$$\mathcal{P}_p(X) \coloneqq \left\{ \mu \in \mathcal{P}(X) \mid \int_X d^p(x, x_0)\, \mathrm{d}\mu(x) < \infty \right\}$$

and the following will be a metric on this space.

Definition: Wasserstein distance

For any $p \in [1, \infty[$ and $\nu, \mu \in \mathcal{P}_p(X)$

$$W_p^p(\mu, \nu) \coloneqq \min \left\{ \int_{X \times X} d^p(x, y) \mathrm{d}\pi(x, y) \mid \pi \in \Gamma(\mu, \nu) \right\}.$$

Theorem: Wasserstein distance is a metric

For any $p \in [1, \infty[$ we have that

$$(\mathcal{P}_p(X), W_p) \text{ is a metric space!}$$

Proof:

We only proof $p = 2$, other p are similar.
Firstly, we show that W_2 maps to $\mathbb{R}$ and hence is always finite. For that, let $\mu, \nu \in \mathcal{P}_2(X)$, then $\pi = \mu \times \nu$ is in $\Gamma(\mu, \nu)$ and we have for any metric d that

$$d^2(x, y) \leq (d(x, x_0) + d(x_0, y))^2 = d^2(x, x_0) + 2d(x, x_0)d(x_0, y) + d^2(x_0, y) \leq 2(d^2(x, x_0) + d^2(x_0, y))$$

because $2ab \leq a^2 + b^2$ always. Now, applying what we derived for d in ! for some $x_0 \in X$, after inserting π in $*$ we get

$$\begin{aligned} W_2^2(\mu, \nu) &\overset{*}{\leq} \int_{X \times X} d^2(x, y) d\pi(x, y) \overset{!}{\leq} 2 \int_{X \times X} d^2(x, x_0) d\pi(x, y) + 2 \int_{X \times X} d^2(x_0, y) d\pi(x, y) \\ &= \int_X \int_X d^2(x, x_0) d\mu(x) d\nu(y) + \int_X \int_X d^2(x_0, y) d\nu(y) d\mu(x) \\ &= \int_X d^2(x, x_0) d\mu(x) + \int_X d^2(x_0, y) d\nu(y) < \infty \end{aligned}$$

where we used Fubini on the first line break and get the finiteness because $\mu, \nu \in \mathcal{P}_2(X)$.

Now, assume that $\int_{X \times X} d^2 d\pi = 0$, which implies that $d^2(x, y) = 0 \Leftrightarrow x = y$ for π-almost all $(x, y) \in X \times X$. This implies that the optimal π only lives on the diagonal $D = \{(x, x) \mid x \in X\}$. Now, for any $\varphi \in C_b(X)$ we have

$$\int_X \varphi(x) \mathrm{d}\mu(x) = \int_{X \times X} \varphi(x) \mathrm{d}\pi(x, y) = \int_{X \times X} \varphi(y) \mathrm{d}\pi(x, y) = \int_X \varphi(y) \mathrm{d}\nu(y)$$

where the first/last equality holds with the definition of marginal and the change of variables formula. The middle one because π lives on D. This shows $\mu = \nu$.

Conversely, assume $\mu = \nu$, then, choose $\pi = (id \times id)_{\#}(\mu)$, which implies

$$\int_{X\times X} d^2(x,y)d\pi(x,y) = \int_X d^2(x,x)d\mu(x) = 0$$

by the change of variables forumla

$$\int_Y f(y)d(T_{\#}(\mu)) = \int_X f(T(x))d\mu(x)$$

where in this case, $d^2 = f$, $T(x) = (id \times id)(x) = (x,x)$ and $Y = X \times X$.

Now, for symmetry we have if $\pi \in \Gamma(\mu,\nu)$ is optimal, that

$$W_2^2(\mu,\nu) = \int_{X\times X} d^2(x,y)d\pi(x,y) = \int_{X\times X} d^2(y,x)d\pi(x,y) = \int_{X\times X} d^2(x,y)d\pi_{\#}(S)(x,y) \overset{!}{=} W_2^2(\nu,\mu)$$

where S(x,y) = (y,x). That the quality in (i) holds, is (to me at least) not trivial to show that, we will show that $\pi_{\#}(S)$ is optimal in $\Gamma(\nu,\mu)$.

For that, assume that $\rho \in \Gamma(\nu,\mu)$ is some other Transport plan, similar to the above we have

$$\int_{X\times X} d^2(x,y)d\pi(x,y) \leq \int_{X\times X} d^2(x,y)d\rho_{\#}(S)(x,y)$$

Now we just look at the push forward of S on both sides to get what we want, where we use that pushing twice by S does nothing since $S \circ S = id$

$$\int_{X\times X} d^2(x,y)d\pi_{\#}(S)(x,y) \leq \int_{X\times X} d^2(x,y)d\rho(x,y)$$

Now, for the triangle inequality, take $\mu_1,\mu_2,\mu_3 \in \mathcal{P}_2(X)$. Let $\pi^{1,2}$ and $\pi^{2,3}$ be the optimal plans of the respective measures. Now, by the Gluing Lemma, we get some $\pi \in \mathcal{P}(X\times X\times X)$, let $\pi^{1,3}$ be the marginal of π in the first and third coordinate.

$$\begin{aligned}
W_2^2(\mu_1,\mu_3) &\leq \left(\int_{X\times X} d^2(x_1,x_3)\mathrm{d}\pi^{1,3}(x_1,x_3)\right) \\
&\overset{!}{=} \left(\int_{X\times X\times X} d^2(x_1,x_3)\mathrm{d}\pi(x_1,x_2,x_3)\right) \\
&\leq \left(\int_{X\times X\times X} \left[d(x_1,x_2)+d(x_2,x_3)\right]^2 \mathrm{d}\pi(x_1,x_2,x_3)\right) \\
&\leq \left(\int_{X\times X\times X} d^2(x_1,x_2)\mathrm{d}\pi(x_1,x_2,x_3)\right) + \left(\int_{X\times X\times X} d^2(x_2,x_3)\mathrm{d}\pi(x_1,x_2,x_3)\right) \\
&\overset{!}{=} \left(\int_{X\times X} d^2(x_1,x_2)\mathrm{d}\pi^{1,2}(x_1,x_2)\right) + \left(\int_{X\times X} d^2(x_2,x_3)\mathrm{d}\pi^{2,3}(x_2,x_3)\right) \\
&= W_2^2(\mu_1,\mu_2) + W_2^2(\mu_2,\mu_3).
\end{aligned}$$

where in !, we used the change of variables formula and the fact that we are working with marginals of π, which means that for example $\pi^{1,3} = p^{1,3}_{\#}(\pi)$.

Using what we have seen in earilier meetings on OT Theory, we have the following two basic properties. Firstly, we get that

$$x \mapsto \delta_x \text{ is an isometry.}$$

This, we directly get, because

$$d(x,y) = \int_{X\times X} d(a,y)d(\delta_x \otimes \delta_y)(a,b)$$

and the product measure $(\delta_x \otimes \delta_y)$ is the only measure that has δ_x, δ_y as its marginals.

Furthermore, by the Kantorovich-Rubinstein Duality we have

$$\begin{aligned} W_1(\mu,\nu) &= \sup_{(\phi,\psi)\in I_c} \left\{ \int_X \phi d\mu + \int_X \psi d\nu \right\} \\ &= \sup_{\|\phi\|_{Lip}\leq 1} \left\{ \int_X \phi d\mu + \int_X \phi^c d\nu \right\} = \sup_{\|\phi\|_{Lip}\leq 1} \left\{ \int_X \phi d\mu - \int_X \phi d\nu \right\} \end{aligned}$$

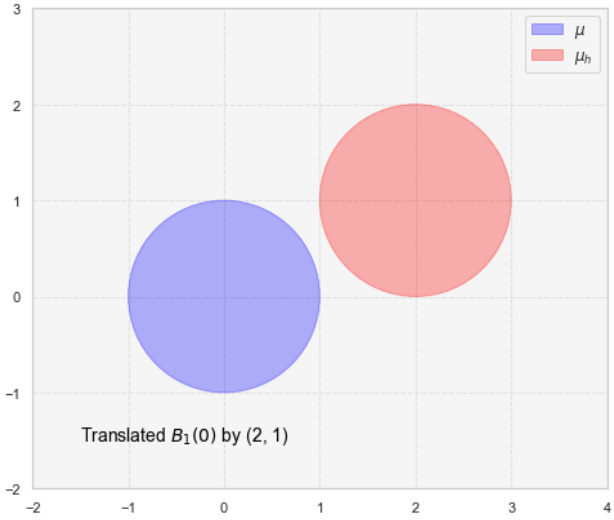
where the first equality actually holds for any p distance, since d^p is a continuous cost function. However, for the second and third equality we need that the cost function is exactly the metric, which is only the case for $d = c$. In that case, we get the the second equality because any c-concave function is 1-Lipschitz, which then implies the third one because in that case $\phi^c = -\phi$. This Duality might be very useful in applications where computing this might be much easier than the minimization over Transport Plans.

Examples for Wasserstein distance

Let $\mu \ll \lambda^n$ with Radon Nikodym derivative f such that $\text{supp}(f) \subseteq \overline{B_1(0)}$.
Additionally, let $\mu_h \ll \lambda^n$ and $f_h(x) = f(x+h)$ be its Radon Nikodym derivative.
As a reminder, this means by the Theorem of Radon Nikodym that

$$\mu(A) = \int_A f(x) d\lambda^n(x)$$

for some $A \in \mathcal{B}(\mathbb{R}^n)$ and simimarly for f_h. Then we have for $\|h\| > 2$ that $supp(\mu_h) \cap supp(\mu) = \emptyset$ since $\overline{B_1(0)} \cap \overline{B_1(h)} = \emptyset$.



Now, this gives us

$$\|f_h - f\|^2_{L^2(\mathbb{R}^n)} = \int_{\mathbb{R}^n} |f(x+h) - f(x)| d\lambda^n(x) = 2\|f\|^2_{L^2(\mathbb{R}^n)}$$

but on the other hand

$$W_2(\mu_h,\mu) = \int_{X\times X} d^2(x,y)\mathrm{d}\pi(x,y) = \int_{\mathbb{R}^n} d^2(x,T(x))d\mu(x) = \int_{\mathbb{R}^n} |x-h+x| d\mu(x) = \|h\|_2$$

by the Brenier, Knott-Smith Theorem we saw in Meeting 5. This Theorem, in this case, gives us that the solution to the Kantorovich Problem is unique and induced by a Transport Map T which in this case is just the translation $T(x) = x + h$ since this is the gradient of the convex and differentiable function $\phi(x) = \frac{1}{2}\|x + h\|^2$ and we have $\mu = \mu_h \circ T^{-1}$. This for large $\|h\|$ implies

$$W_2(\mu_h, \mu) \gg \|f_h - f\|_{L^2(\mathbb{R}^n)}.$$

In the case of small $\|h\|$ we can also find special f such that $W_2(\mu_h, \mu) \ll \|f_h - f\|_{L^2(\mathbb{R}^n)}$.

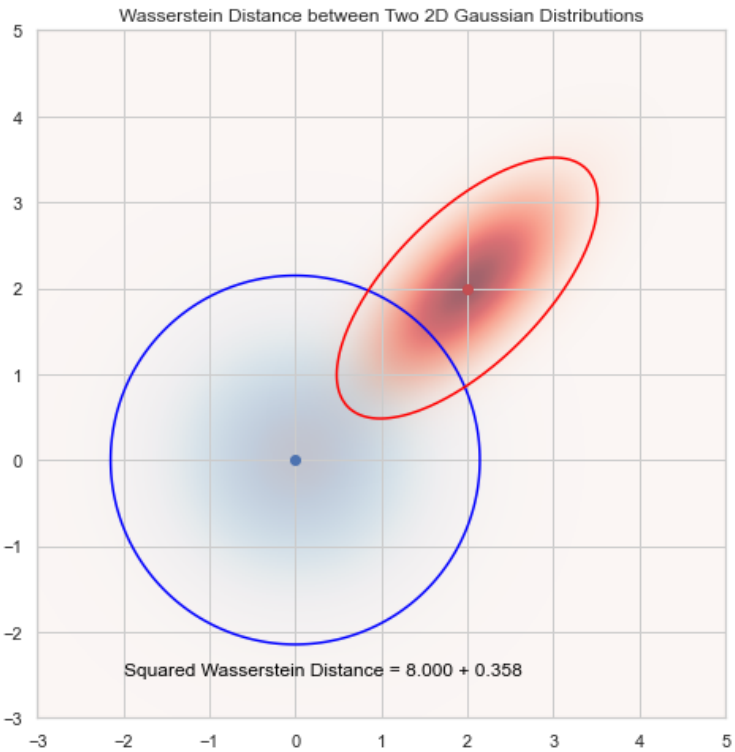
For any $\mu_X, \mu_Y \in \mathcal{P}_2(\mathbb{R}^d)$ with $\mathcal{L}(X) = \mu_X$ and $\mathcal{L}(Y) = \mu_Y$ where X and Y are normally distributed, the Wasserstein distance can be calculated as follows

$$W_2(\mu_x, \mu_Y) = \|m_X - m_Y\|_2^2 + \operatorname{tr}\left(\Sigma_X - 2 \cdot \left(\Sigma_Y^{1/2}\Sigma_X\Sigma_Y^{1/2}\right)^{1/2} + \Sigma_Y\right)$$

If Σ_X and Σ_Y commute, this simplifies to

$$W_2(\mu_x, \mu_Y) = \|m_X - m_Y\|_2^2 + \sum_{i=1}^{d}\left(\sqrt{\lambda_i^X} - \sqrt{\lambda_i^Y}\right)^2$$

where λ_i^X and λ_i^Y are the Eigenvalues of Σ_X and Σ_Y respectively.
As an example, let us take a look at $\Sigma_X = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ and $\Sigma_Y = \begin{pmatrix} \frac{1}{2} & \frac{1}{3} \\ \frac{1}{3} & \frac{1}{2} \end{pmatrix}$ with $m_X = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$ and $m_Y = \begin{pmatrix} 2 \\ 2 \end{pmatrix}$.
This looks like the following, where a darker colored area indicates more mass to be concentrated there.



We can also calculate this ourselves, we directly get since $\lambda^2 - \lambda + \frac{5}{36}$ is the char polynomial.

$$W_2^2(\mu_Y, \mu_X) = \left\|\begin{pmatrix} 2 \\ 2 \end{pmatrix}\right\|_2^2 + \left(1 - \sqrt{\frac{5}{6}}\right)^2 + \left(1 - \sqrt{\frac{1}{6}}\right)^2 = 8 + \left(1 - \sqrt{\frac{5}{6}}\right)^2 + \left(1 - \sqrt{\frac{1}{6}}\right)^2 \approx 8.358$$

If we restrict ourselves to $\mathcal{P}_\infty(X)$, the space of probability measures with bounded support. We can also define W_∞ as the limit of W_p.

$$\begin{aligned}\lim_{p\to\infty} W_p(\mu,\nu) &= W_\infty(\mu,\nu) = \inf\{\|d(x,y)\|_{L^\infty}(\pi) \mid \pi \in \Gamma(\mu,\nu)\} \\ &= \inf_{\pi\in\Gamma(\mu,\nu)} \inf\{C \geq 0 \mid |d(x,y)| \leq C \text{ for } \pi \text{ almost all } (x,y)\}\end{aligned}$$

and we have

$$(\mathcal{P}_\infty(X), W_\infty) \text{ is a metric space}$$

Lifting completeness

To lift the completeness from a complete metric space (X,d) to $(\mathcal{P}_p(X), W_p)$, we need a iterated version of the gluing Lemma.

Lemma: Iterated Gluing/Dudley

Let $N \geq 3$ and for any $n \leq N$ (X_n, d_n) Polish, $\mu_n \in \mathcal{P}(X_n)$ and $\theta_n \in \Gamma(\mu_{n-1}, \mu_n)$.
Then there exists $\pi_n \in \mathcal{P}(X_1 \times ... \times X_n)$ for any $n \leq N$ such that.

1. $p^{1,\dots,n-1}_{\#}\pi_n = \pi_{n-1}$ for $2 \leq n \leq N$
2. $p^{i}_{\#}\pi_n = \mu_i$ for $1 \leq i \leq n \leq N$
3. $p^{i-1,i}_{\#}\pi_n = \theta_i$ for $2 \leq i \leq n \leq N$

Proof:

For $N = 3$ this is just the Dudley Lemma. In other cases, we iteratively ;) apply the Dudley Lemma in the following way.

$$X_1 \times X_2 \times ... \times X_n = Z_1 \times Z_2 \times Z_3, \quad Z_1 = (X_1 \times ... \times X_{n-2}), \quad Z_2 = X_{n-1}, \quad Z_3 = X_n.$$

since we already have $\pi_{n-1} \in \mathcal{P}(Z_1 \times Z_2)$ and $\theta_n \in \Gamma(\mu_{n-1}, \mu_n) \subseteq \mathcal{P}(Z_2 \times Z_3)$ to get π_n. Note that this also works for $N = \infty$, in this case, the inequalities become strict for the three cases.
TODO DRAWING

We want to apply this Lemma in the setting that $X_i = X$ and (μ_n) is some sequence of measures on X.
In this case we, by the Lemma, get a sequence of meaures (π_n) from the Lemma. Now, we can (by for example Ionescu-Tulcea) get a measure on $\mathbb{X} = \prod_{i=0}^{\infty} X$, so we have the following

$$\pi_\infty \in \mathcal{P}(\mathbb{X}) \quad \text{ such that } \quad p^{1,\dots,n}_{\#}(\pi) = \pi_n.$$

We will also need the metric version of the L^p spaces.

$$L^p(\Omega, \mathcal{F}, P, X) \coloneqq \left\{ f : \Omega \to X \mid f \text{ measurable }, \int_\Omega d^p(f, z_0) dP < \infty \right\}$$

with

$$d^p_{L^p}(f,g) \coloneqq \int_\Omega d^p(f,g) dP$$

These are not necessarily vector spaces, but one can show that this more general notion of L^p space is still complete. (Which we will need in the following)

Theorem: Lifting completeness from X to $\mathcal{P}_p(X)$

Let (X, d) be a complete metric space, then $(\mathcal{P}_p(X), W_p)$ is complete.

Proof:

We again only prove it for $p = 2$, the other cases are similar.

Let (μ_n) be a cauchy sequence in $\mathcal{P}_2(X)$ with respect to W_2. Now, applying our idea from above using the Iterated Dudley Lemma for μ_n with $X_i = X$ and $\theta_n \in \Gamma_o(\mu_n, \mu_{n+1})$. We thus get the above mentioned π_∞ and by construction we have that the μ_n are the marginals of π_∞. So we have

$$p^n_\#(\pi) = \mu_n \quad (p_n, p_{n+1})_\#(\pi) = \theta_n.$$

Now, the key observation is the following, we can, without loss of generality assume that $\sum_{n=0}^{\infty} W_2(\mu_n, \mu_{n+1})$ exists, because we can always extract a subsequence such that this holds and because our sequence is cauchy, the limit of the subsequence has to be the limit of the enitre sequence. The subsequence we might construct could be the following, given n_k choose n_{k+1} such that

$$W_2(\mu_{n_k}, \mu_{n_{k+1}}) < 2^{-k},$$

which we can do because our sequence is Cauchy, then the series obviously converges. Using this assumption, we can then deduce that

$$\begin{aligned}
\|p_n - p_{n+1}\|_{L^2(\pi_\infty)} &= \int_{\mathbb{X}} d^2(p, p_{n+1}) d\pi_\infty \\
&= \int_{X\times X} d^2(x_n, x_{n+1}) d(p_n, p_{n+1})_\#(\pi_\infty) \\
&= \int_{X\times X} d^2(x_n, x_{n+1}) d\theta(x_n, x_{n+1}) = W_2^2(\mu_n, \mu_{n+1})
\end{aligned}$$

and thus the sequence of projections (p_n) is Cauchy in $L^2(\mathbb{X}, \mathcal{B}_\infty, \pi_\infty, X)$. Now, we know that L^2 is complete, let p_∞ be the limit of (p_n) and define

$$\mu_\infty = (p_\infty)_\#(\pi_\infty).$$

If we use what we had earlier in the Book, namely that

$$S_\# P = \mu, \quad T_\# P = \nu \Rightarrow (S, T)_\# P \in \Gamma(\mu, \nu).$$

we get

$$(p_n)_\#(\pi_\infty) = \mu_n, \quad (p_\infty)_\#(\pi_\infty) = \mu_\infty \Rightarrow (p_n, p_\infty)_\#(\pi_\infty) \in \Gamma(\mu_n, \mu_\infty).$$

and thus have

$$\begin{aligned}
W_2^2(\mu_n, \mu_\infty) &\leq \int_{X\times X} d^2(x_n, x_{n+1}) d(p_n, p_\infty)_\#(\pi_\infty)(x_n, x_{n+1}) \\
&= \int_{\mathbb{X}} d^2(p_n, p_\infty) d\pi_\infty = \|p_n - p_\infty\|_{L^2(\pi_\infty)} \xrightarrow{n\to\infty} 0
\end{aligned}$$

which gives us $\mu_n \xrightarrow{W_2} \mu_\infty$ and the completeness of $(\mathcal{P}_2(X), W_2)$.