Lecture #14: Optimal Transport

- Last time, we defined Wasserstein p-metrics on $\mathcal{P}_p(X)$ = probability measures with finite p^{th} moment for Polish space (X,d). Specifically,

$$W_p(\mu,\nu) := \left[\min_{\pi \in \Pi(\mu,\nu)} \int_{X\times X} d^p(x,y)\, d\pi(x,y) \right]^{1/p}.$$

- This is not the only way to ~~control~~ quantify probability measure difference. Indeed, we have the following, which bounds Wasserstein metrics in terms of total variation:

Proposition (Weighted TV upper bounds W_p): Let μ,ν be probability measures on (X,d) a Polish space. For any $p \geq 0$, $x_0 \in X$,

$$\min_{\pi \in \Pi(\mu,\nu)} \int_{X\times X} d^p(x,y)\, d\pi(x,y) \leq \max\{1, 2^{p-1}\} \cdot \int d^p(x_0,x)\, d|\mu-\nu|(x)$$

Proof: Let $\pi_0 := (Id \times Id)_{\#} (\mu \wedge \nu) + \frac{1}{a}[\mu-\nu]_+ \otimes [\mu-\nu]_-$,

where $\mu \wedge \nu := \mu - [\mu-\nu]_+$

$$a := [\mu-\nu]_+(X) = [\nu-\mu]_-[X]$$

Then $\min_{\pi \in \Pi(\mu,\nu)} \int_{X\times X} d^p(x,y)\, d\pi(x,y)$

N.B.: $[\mu-\nu]_+[A]$

$-[\mu-\nu]_-[A]$

$= [\mu-\nu][A]$ for all measurable A, $[\mu-\nu]_\pm$ are positive and singular w.r.t. each other

$$\leq \int d^p(x,y)\, d\pi_0(x,y)$$

$$= \frac{1}{a}\int d^p(x,y)\, d[\mu-\nu]_+(x)\, d[\mu-\nu]_-(y) \qquad \left(\text{since the } (Id\times Id)_\# (\mu\wedge\nu) \text{ maps to } d(x,x)=0\right)$$

$$\leq \frac{\max\{1,2^{p-1}\}}{a}\int \left[d(x,x_0)^p + d(x_0,y)^p\right] d[\mu-\nu]_+(x)\, d[\mu-\nu]_-(y).$$

Here, we have used the inequality that $[A+B]^p \leq \max\{1,2^{p-1}\}[A^p+B^p]$ for all $A,B\geq 0$ and $p\geq 0$.

Then, since $\frac{1}{a}\cdot\int d[\mu-\nu]_+(x) = 1 = \frac{1}{a}\cdot\int d[\nu-\mu]_-(y)$,

we have that the above is

$$= \max\{1,2^{p-1}\}\left[\int d(x,x_0)^p\, d[\mu-\nu]_+(x) + \int d(x_0,y)^p\, d[\mu-\nu]_-(y)\right]$$

$$= \max\{1,2^{p-1}\}\int d(x,x_0)^p\, d\left[(\mu-\nu)_+ + (\mu-\nu)_-\right](x)$$

$$= \max\{1,2^{p-1}\}\int d(x,x_0)^p\, d|\mu-\nu|(x). \blacksquare$$

<u>Corollary</u>: For $p\geq 1$, $W_p(\mu,\nu) \leq 2^{\frac{p-1}{p}}\cdot \|d(x_0,\cdot)^p(\mu-\nu)\|_{TV}^{\frac{1}{p}}$

//

- An important practical application of W_p is that it parametrizes a relatively lax notion of distance. Precisely, it parametrizes weak convergence of measures, as follows.

Theorem (Wp parametrizes weak convergence): Let $p \in (0,\infty)$ and let $\{\mu_k\}_{k=1}^{\infty}$ be a sequence of measures in $\mathcal{P}_p(X)$. Let $\mu \in \mathcal{P}(X)$. TFAE:

(i) $\lim_{k\to\infty} W_p(\mu_k,\mu) = 0$

(ii) For all $\ell \in \ell_b(X)$, $\lim_{k\to\infty} \int \ell \, d\mu_k = \int \ell \, d\mu$ and

for some $x_0 \in X$, $\lim_{R\to\infty} \limsup_{k\to\infty} \int_{d(x_0,x)\geq R} d(x_0,x)^p \, d\mu_k(x) = 0$

(iii) For all $\ell \in \ell_b(X)$, $\lim_{k\to\infty} \int \ell \, d\mu_k = \int \ell \, d\mu$ and

for some $x_0 \in X$, $\lim_{k\to\infty} \int d(x_0,x)^p \, d\mu_k(x) = \int d(x_0,x)^p \, d\mu(x) < \infty$

(iv) for any $\ell \in \ell(X)$ s.t. $|\ell(x)| \leq C[1 + d(x_0,x)^p]$ for some $x_0 \in X$, $C>0$, then
- $\int |\ell| \, d\mu_k < \infty$
- $\lim_{k\to\infty} \int \ell \, d\mu_k = \int \ell \, d\mu$.

Proof: See Villani. ∎

//

- As usual, things are especially nice when $d=1$. This is because we can work with cdfs. Recall that weak convergence of $\{\mu_k\}_{k=1}^{\infty}$ ($\int \ell \, d\mu_k \to \int \ell \, d\mu$ for all $\ell \in \ell_b(X)$) is equivalent to point wise convergence of the cdfs inverse at points of continuity:

Proposition (Weak convergence for $d=1$): Let $\{\mu_k\}_{k=1}^{\infty}$ and μ be measures on $\mathbb{R}$ with

respective cdfs $\{F_k\}_{k=1}^{\infty}$ and F. Then $\lim_{k\to\infty} \mu_k = \mu$ weakly

$\Updownarrow$

$\lim_{k\to\infty} F_k(x) = F(x)$ $\forall x$ s.t. F is continuous at x.

- But since we are in $d=1$, $W_1(\mu_k, \mu)$ (which parametrizes $\lim_{k\to\infty} \mu_k = \mu$ weakly) has a particularly nice form.
- Indeed, from Lecture 12, we know $W_1(\mu_k, \mu) = \|F_k - F\|_{L^1(\mathbb{R})}$. So, we have

$\lim_{k\to\infty} F_k(x) = F(x)$ $\forall x$ s.t. F is continuous at x

$\Updownarrow$

$\lim_{k\to\infty} \|F_k - F\|_{L^1(\mathbb{R})} = 0$

Remark: What is nice about "weak" convergence? It is somehow invariant enough to be useful:

- For any x, y, $W_p(\delta_x, \delta_y) = d(x,y)^p$ for $p \geq 1$
- Let $k \geq 1$ integer and set $d\mu_k(x) = [1 + \sin(2\pi k x)] \in \mathcal{P}(\mathbb{R})$. Then

$W_p(\mu_k, \mathcal{L}) \leq C_p \cdot \frac{1}{k}$ for $\mathcal{L}$ Lebesgue measure on $[0,1]$,

C_p a uniform universal constant.

These suggest that W_p can really see close versus far things in probability space. Note in particular that $\|\delta_x - \delta_y\|_{TV} = 2$ $\forall x \neq y$. This point is made for other ways of comparing measures in "[redacted] Information Geometric

Adversarial Networks" a landmark ML paper.

- How to "add" measures in a way compatible with the Wasserstein metrics?
- The most obvious way is through linear mixture models, as follows. Let $\{\mu_i\}_{i=1}^n \subset \mathcal{P}(\mathbb{R}^d)$ and let $(\lambda_1, \dots, \lambda_n)$ be s.t. $\lambda_i \geq 0$ for all i and $\sum_{i=1}^n \lambda_i = 1$. We can define a measure $\mu := \sum_{i=1}^n \lambda_i \mu_i$, which practically means that sampling from μ means sampling from the multinomial with parameters $(\lambda_1, \dots, \lambda_n)$, then sampling from the corresponding μ_i.
- Alternatively, one could construct a measure that is "on average" close to the $\{\mu_i\}_{i=1}^n$:

$$\mu_\lambda := \operatorname*{argmin}_{\mu \in \mathcal{P}(\mathbb{R}^d)} \sum_{i=1}^n \lambda_i W_p^p(\mu_i, \mu).$$

- Why does this look reasonable? Consider the analogous construction in $\mathbb{R}$ with distance function $d(x,y) = |x-y|^p$:

$$X_\lambda := \operatorname*{argmin}_{x \in \mathbb{R}} \frac{1}{n} \sum_{i=1}^n |x_i - x|^2$$

Differentiating w.r.t. x yields: $F(x) := \sum_{i=1}^n |x_i - x|^2$

$$\Rightarrow F'(x) := \sum_{i=1}^n \left(2(x_i - x)\right)$$

$$\Rightarrow F'(x) = 0 \iff x = \frac{1}{n} \sum_{i=1}^n x_i;$$ no difference if we replace $\frac{1}{n}$ with non-uniform weights

- Similar analysis works in $\mathbb{R}^d$ / not $\mathcal{P}(\mathbb{R}^d)$...