Wasserstein Distance

Tong Zhou

November 30, 2020

Contents

1 Definitions

The p-Wasserstein distance between probability measures μ and ν on $\mathbb{R}^d$ is defined as

$$W_p(\mu,\nu) = \inf_{\substack{X\sim\mu \\ Y\sim\nu}} \left(\mathrm{E}\left[\|X-Y\|^p\right]\right)^{1/p}, \quad p \geqslant 1.$$

Usually, $W_p(\mu,\nu)$ is also denoted by $W_p(X,Y)$.

Remark.

- If a general metric space $(\mathcal{X},\rho)$ is complete and separable, the above norm in the definition is replaced by $\rho(X,Y)$.
- W_p is a proper metric: it is nonnegative, symmetric and satisfies triangle inequality.

2 Properties of Wasserstein distance

Let X and Y be random variables taking values in $\mathcal{X} = \mathbb{R}^d$; the notation $(\mathcal{X}, \|\cdot\|)$ is maintained.

- For any real number a, $W_p(aX, aY) = |a|\, W_p(X,Y)$.
- For any fixed vector $x \in \mathcal{X}$, $W_p(X+x, Y+x) = W_p(X,Y)$.
- For any fixed $x \in \mathcal{X}$, we have $W_2^2(X+x, Y) = \|x + \mathrm{E}[X] - \mathrm{E}[Y]\| + W_2^2(X,Y)$.
- For product measures and when $p=2$, we have $W_2^2\left(\otimes_{i=1}^n \mu_i, \otimes_{i=1}^n \nu_i\right) = \sum_{i=1}^n W_2^2(\mu_i,\nu_i)$ in the analytic notation.

3 Relation to Other Metrics

For random variables X and Y on $\mathcal{X}$, let Ω be the union of their ranges and set

$$D = \sup_{x,y\in\Omega} \|x-y\|, \quad d_{\min} = \inf_{x\neq y\in\Omega} \|x-y\|$$

It possesses the following properties:

- If $p \leqslant q$, then $W_p \leqslant W_q$, by Jensen's inequality.
- On the other hand, $W_q^q \leqslant W_p^p D^{q-p}$.
- Duality arguments yield the particularly useful *Kantorovich-Rubinstein* representation for W_1 as

$$W_1(X,Y) = \sup_{\|f\|_{\mathsf{Lip}}\leqslant 1} \left|\mathrm{E}[f(X)] - \mathrm{E}[f(Y)]\right|, \quad \|f\|_{\mathsf{Lip}} = \sup_{x\neq y} \frac{|f(x)-f(y)|}{\|x-y\|},$$

- This shows that W_1 is larger than the *Bounded Lipschitz* (BL) metric

$$W_1(X,Y) \geqslant \mathsf{BL}(X,Y) = \sup_{\|f\|_\infty + \|f\|_{\mathsf{Lip}} \leqslant 1} |\mathrm{E}[f(X)] - \mathrm{E}[f(Y)]|$$

  that metrises convergence in distribution.

- Let P denote the Prokhorov distance. Then $P^2(X,Y) \leqslant W_1(X,Y) \leqslant (D+1)P(X,Y)$.

- For the class of random variables supported on a fixed bounded subset $K \subset \mathcal{X}$, BL and W_1 are equivalent up to constant, and all metrics W_p are topologically equivalent.

- The Wasserstein distances W_p can be bounded by a version of total variation TV. A weaker but more explicit bound for $p = 1$ is $W_1(X,Y) \leqslant D \times \mathsf{TV}(X,Y)$.

- For discrete random variables, there is an opposite bound $\mathsf{TV} \leqslant \frac{W_1}{\mathsf{d}_{\mathsf{min}}}$.

- The total variation between convolutions with a sufficiently smooth measure is abounded above by W_1.

- The *Toscani* (or *Toscani-Fourier*) distance is also bounded above by W_1.

$$\boldsymbol{\beta}_\lambda(\mathcal{T}, \boldsymbol{W}) = \left(\boldsymbol{W}\boldsymbol{X}\boldsymbol{X}^\intercal\boldsymbol{W}^\intercal + \lambda\boldsymbol{I}\right)^{-1}\boldsymbol{W}\boldsymbol{X}\boldsymbol{y},$$

We may now calculate the bias and variance of the model described above via the following formulations:

$$\begin{aligned}\mathsf{Bias}_\lambda(\boldsymbol{\theta})^2 &= \mathbb{E}_{\boldsymbol{x}}\left[\mathbb{E}_{\mathcal{T},\boldsymbol{W}}\left[f_\lambda(\boldsymbol{x};\mathcal{T},\boldsymbol{W})\right] - f_0(\boldsymbol{x})\right]^2, \\ \mathsf{Variance}_\lambda(\boldsymbol{\theta}) &= \mathbb{E}_{\boldsymbol{x}}\left[\mathsf{Var}_{\mathcal{T},\boldsymbol{W}}\left(f_\lambda(\boldsymbol{x};\mathcal{T},\boldsymbol{W})\right)\right].\end{aligned}$$