Smooth Wasserstein Distance: Metric Structure and Statistical Efficiency

Smooth Wasserstein Distance: Metric Structure and Statistical Efficiency

Ziv Goldfeld
Cornell University
goldfeld@cornell.edu

***Abstract*—The Wasserstein distance has seen a surge of interest and applications in machine learning. Its popularity is driven by many advantageous properties it possesses, such as metric structure (metrization of weak convergence), robustness to support mismatch, compatibility to gradient-based optimization, and rich geometric properties. However, empirical approximation under the Wasserstein distance suffers from a severe curse of dimensionality, rendering it impractical in high dimensions. We propose a novel Gaussian-smoothed Wasserstein distance, that achieves the best of both worlds: preserving the Wasserstein metric structure while alleviating the empirical approximation curse of dimensionality. Furthermore, as the smoothing parameter shrinks to zero, smooth Wasserstein converges towards the classic metric (with convergence of optimizers), thus serving as a natural extension. These theoretic properties establish the smooth Wasserstein distance as favorable alternative to its classic counterpart for high-dimensional analysis and applications.**

I. Extended Abstract

The 1-Wasserstein distance (W_1) between two probability measures P and Q, with finite first moments, is

$$\mathsf{W}_1(P,Q) := \inf_{\pi\in\Pi(P,Q)} \int \|x-y\| \,\mathsf{d}\pi(x,y),$$

where $\Pi(P,Q)$ is the set of couplings of P and Q. This distance has many appealing properties, such as: (i) robustness to mismatched supports of P and Q (crucial for generative modeling applications); (ii) metrization of weak convergence of probability measures; (iii) defining a constant speed geodesic in the space of probability measures (giving rise to a natural interpolation between measures). These advantages, however, come at a price of slow empirical convergence rates, known as the 'curse of dimensionality'.

Suppose $(X_i)_{i=1}^n$ are i.i.d. samples from a Borel probability measure P on $\mathbb{R}^d$. Consider the rate at which the empirical measure $P_n \triangleq \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ approaches P in the 1-Wasserstein distance, i.e., the $\mathbb{E}\mathsf{W}_1(P_n,P)$ rate of decay. Since W_1 metrizes narrow convergence, the Glivenko-Cantelli theorem implies $\mathsf{W}_1(P_n,P)\to 0$ as $n\to\infty$. Unfortunately, the convergence *rate* drastically deteriorates with dimension, scaling as $n^{-\frac{1}{d}}$ for any P absolutely continuous w.r.t. the Lebesgue measure [1]. This rate is sharp for all $d>2$. Thus, empirical approximation under W_1 is effectively infeasible in high dimensions – a disappointing shortcoming given the dimensionality of data in modern ML tasks.

To alleviate this impasse, we propose a novel framework, termed Gaussian-smooth Wasserstein distance that inherits the metric structure of W_1 while attaining much stronger statistical guarantees. The smooth Wasserstein distance of parameter $\sigma \geq 0$ between two d-dimensional probability measures P and Q is

$$\mathsf{W}_1^{(\sigma)}(P,Q) \triangleq \mathsf{W}_1(P*\mathcal{N}_\sigma, Q*\mathcal{N}_\sigma),$$

where $*$ stands for convolution and $\mathcal{N}_\sigma \triangleq \mathcal{N}(0,\sigma^2\mathrm{I}_d)$ is the isotropic Gaussian measure of parameter σ. In other words, $\mathsf{W}_1^{(\sigma)}(P,Q)$ is simply the W_1 distance between P and Q after each is smoothed by an isotropic Gaussian kernel.

Theorem 1 of [2] shows that just like W_1, for any $\sigma\in[0,+\infty)$, $\mathsf{W}_1^{(\sigma)}$ is a metric on the space of probability measures that metrizes weak topology. Namely, a sequence of probability measures $(P_k)_{k\in\mathbb{N}}$ converges weakly to P if and only if $\mathsf{W}_1^{(\sigma)}(P_k,P)\to 0$. This further implies that convergence to zero of W_1 and $\mathsf{W}_1^{(\sigma)}$ are equivalent (see [2, Theorem 2]). We next explore properties of $\mathsf{W}_1^{(\sigma)}(P,Q)$ as a function of σ for fixed P and Q. Theorem 3 in [2] establishes continuity and non-increasing monotonicity of $\mathsf{W}_1^{(\sigma)}(P,Q)$ in $\sigma\in[0,+\infty)$. These, in particular, imply that $\lim_{\sigma\to 0}\mathsf{W}_1^{(\sigma)}(P,Q)=\mathsf{W}_1(P,Q)$. Additionally, using the notion of Γ-convergence, Theorem 4 of the aforementioned work establishes convergence of optimal couplings. Namely, if $(\pi_k)_{k\in\mathbb{N}}$ is sequence of optimal couplings for $\mathsf{W}_1^{(\sigma_k)}(P,Q)$, where $\sigma_k\to 0$, then $(\pi_k)_{k\in\mathbb{N}}$ converges weakly to an optimal coupling for $\mathsf{W}_1(P,Q)$.

Lastly, consider empirical approximation under smooth Wasserstein, i.e., the convergence rate of $\mathbb{E}\mathsf{W}_1^{(\sigma)}(P_n,P)$. It was shown in [3, Proposition 1] that Gaussian smoothing alleviates the curse of dimensionality, with $\mathbb{E}\mathsf{W}_1^{(\sigma)}(P_n,P)$ converging as $n^{-\frac{1}{2}}$ in all dimensions. Although $\mathsf{W}_1^{(\sigma)}$ is specialized to Gaussian noise, Theorem 5 of [2] generalizes the empirical approximation result to account for subgaussian noise densities. The expected value analysis is followed by a concentration inequality for $\mathsf{W}_1^{(\sigma)}(P_n,P)$ derived through McDiarmid's inequality [2, Theorem 6].

References

[1] R. M. Dudley. The speed of mean Glivenko-Cantelli convergence. *Ann. Math. Stats.*, 40(1):40–50, Feb. 1969.

[2] Z. Goldfeld and K. Greenewald. Gaussian-smoothed optimal transport: Metric structure and statistical efficiency. In *International Conference on Artificial Intelligence and Statistics (AISTATS-2020)*, Palermo, Sicily, Italy, Jun. 2020.

[3] Z. Goldfeld, K. Greenewald, Y. Polyanskiy, and J. Weed. Convergence of smoothed empirical measures with applications to entropy estimation. *arXiv preprint arXiv:1905.13576*, May 2019.