Gaussian-Smoothed Optimal Transport: Metric Structure and Statistical Efficiency

Ziv Goldfeld
Cornell University

Kristjan Greenewald
MIT-IBM Watson AI Lab

Abstract

Optimal transport (OT), in particular the Wasserstein distance, has seen a surge of interest and applications in machine learning. However, empirical approximation under Wasserstein distances suffers from a severe curse of dimensionality, rendering them impractical in high dimensions. As a result, entropically regularized OT has become a popular workaround. While it enjoys fast algorithms and better statistical properties, it however loses the metric structure that Wasserstein distances enjoy. This work proposes a novel Gaussian-smoothed OT (GOT) framework, that achieves the best of both worlds: preserving the 1-Wasserstein metric structure while alleviating the empirical approximation curse of dimensionality. Furthermore, as the Gaussian-smoothing parameter shrinks to zero, GOT Γ-converges towards classic OT (with convergence of optimizers), thus serving as a natural extension. An empirical study that supports the theoretical results is provided, promoting Gaussian-smoothed OT as a powerful alternative to entropic OT.

1 Introduction

In recent years optimal transport (OT) has been applied to a host of machine learning (ML) tasks as a powerful means of comparing probability measures. The Kantorovich OT [Kantorovich, 1942] problem between two probability measures μ and ν with cost $c(x,y)$ is given by

$$\inf_{\pi\in\Pi(\mu,\nu)} \int c(x,y)\,\mathsf{d}\pi(x,y), \tag{1}$$

where $\Pi(\mu,\nu)$ is the set of transport plans (or couplings) between μ and ν. Applications of the Kantorovich formulation include data clustering [Ho et al., 2017], density ratio estimation [Iyer et al., 2014], domain adaptation [Courty et al., 2016, Courty et al., 2014], generative models [Arjovsky et al., 2017, Gulrajani et al., 2017], image recognition [Rubner et al., 2000, Sandler and Lindenbaum, 2011, Li et al., 2013], word and document embedding [Alvarez-Melis and Jaakkola, 2018, Yurochkin et al., 2019, Grave et al., 2019], and many others.

This surge in popularity has been driven by some highly advantageous properties of OT. Beyond its robustness to mismatched supports of μ and ν (crucial for learning generative models), when $c(x,y)=\|x-y\|$, (1) becomes the 1-Wasserstein distance[1], which (i) has the operational interpretation of minimizing work (or expected cost); (ii) metrizes weak (also known as, weak*) convergence of probability measures; and (iii) defines a constant speed geodesic in the space of probability measures (giving rise to a natural interpolation between measures). These advantages, however, come with a price as OT is generally hard to compute and suffers from the so-called curse of dimensionality.

Specifically, suppose we have n independent samples $(X_i)_{i=1}^n$ from a Borel probability measure μ on $\mathbb{R}^d$. Consider the fundamental question of how quickly the empirical measure $\hat{\mu}_n \triangleq \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ approaches μ in the 1-Wasserstein distance, i.e., the $\mathbb{E}\mathsf{W}_1(\hat{\mu}_n,\mu)$ rate of decay. This quantity is at the heart of empirical approximation under W_1 since it controls the error in various additional approximation setups, such as $\mathbb{E}\big|\mathsf{W}_1(\hat{\mu}_n,\nu)-\mathsf{W}_1(\mu,\nu)\big|$ (one-sample goodness of fit test), $\mathbb{E}\big|\mathsf{W}_1(\hat{\mu}_n,\hat{\nu}_n)-\mathsf{W}_1(\mu,\nu)\big|$ (two-sample tests)[2], and others; see [Panaretos and Zemel, 2019] for a

Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS) 2020, Palermo, Italy. PMLR: Volume 108.

[1] Any p-Wasserstein distance has these properties.

[2] Note that while Wasserstein-type GANs in practice typically use the two-sample setup since the generator

review on statistical applications of the Wasserstein distance. Since W_1 metrizes weak convergence [Villani, 2008, Cor. 6.18], the Glivenko-Cantelli theorem [Varadarajan, 1958] implies $\mathsf{W}_1(\hat{\mu}_n, \mu) \to 0$ as $n \to \infty$. Unfortunately, the convergence *rate* in n drastically deteriorates with dimension, scaling at best as $n^{-\frac{1}{d}}$ for any measure μ that is absolutely continuous with respect to (w.r.t.) the Lebesgue measure [Dudley, 1969]. Note that the $n^{-\frac{1}{d}}$ rate is sharp for all $d > 2$ (see [Dobrić and Yukich, 1995] for sharper results). This renders empirical approximation under the Wasserstein distance infeasible in high dimensions – a disappointing shortcoming given the dimensionality of data in modern ML tasks.

In light of the above, entropic OT emerged as an appealing alternative to Kantorovich OT. Its popularity has been driven both by algorithmic advances [Cuturi, 2013, Altschuler et al., 2017] and some better statistical properties it possesses [M. et al., 2017, Montavon et al., 2016, Rigollet and Weed, 2018]. Entropic OT regularizes the expected cost by a Kullback-Leibler (KL) divergence, forming:

$$\mathsf{S}_c^{(\epsilon)}(\mu,\nu) \triangleq \inf_{\pi\in\Pi(\mu,\nu)} \int c(x,y)\,\mathsf{d}\pi(x,y) + \epsilon \mathsf{D}(\pi\|\mu\times\nu),$$

where $c(x,y)$ is the cost and $\mathsf{D}(\alpha\|\beta) \triangleq \int \log\left(\frac{\mathsf{d}\alpha}{\mathsf{d}\beta}\right)\mathsf{d}\alpha$ if $\alpha \ll \beta$ and $+\infty$ otherwise. While the Wasserstein distance suffers from the curse of dimensionality, [Genevay et al., 2019] showed that if c is Lipschitz and infinitely differentiable, then $\mathbb{E}\big|\mathsf{S}_c^{(\epsilon)}(\hat{\mu}_n,\hat{\nu}_n) - \mathsf{S}_c^{(\epsilon)}(\mu,\nu)\big| \in O\left(n^{-\frac{1}{2}}\right)$ in all dimensions (see [Mena and Niles-Weed, 2019] for sharper results specialized to quadratic cost). Despite this fast convergence in the two-sample test, sample complexity bounds in the (stronger) one-sample regime are not available. More importantly, the assumptions from [Genevay et al., 2019] exclude the distance cost $c(x,y) = \|x-y\|$, which is our main interest. Another drawback is that $\mathsf{S}_c^{(\epsilon)}(\mu,\nu)$ is not a metric, even when $c(x,y)$ is [Feydy et al., 2019, Bigot et al., 2019] (e.g, $\mathsf{S}_c^{(\epsilon)}(\mu,\mu) \neq 0$).[3] Hence entropic OT retains several gaps in statistical convergence guarantees, and more importantly, it surrenders desirable properties of the Wasserstein distance. We thus seek an alternative OT framework that enjoys the best of both worlds.

Contributions. This paper proposes a novel OT framework, termed Gaussian-smoothed OT (GOT) that inherits the metric structure of W_1 while attaining stronger statistical guarantees than available for entropic OT. GOT of parameter $\sigma \geq 0$ between two d-dimensional probability measures μ and ν is defined as

$$\mathsf{W}_1^{(\sigma)}(\mu,\nu) \triangleq \mathsf{W}_1(\mu * \mathcal{N}_\sigma, \nu * \mathcal{N}_\sigma), \tag{2}$$

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, $\mathsf{W}_1^{(\sigma)}(\mu,\nu)$ is simply the W_1 distance between μ and ν after each is smoothed by an isotropic Gaussian kernel.

We first show that just as W_1, for any fixed $\sigma \in [0,+\infty)$, $\mathsf{W}_1^{(\sigma)}$ is a metric on the space of probability measures that metrizes the weak topology. Namely, a sequence of probability measures $(\mu_k)_{k\in\mathbb{N}}$ converges weakly to μ if and only if $\mathsf{W}_1^{(\sigma)}(\mu_k,\mu) \to 0$. We then turn to study properties of $\mathsf{W}_1^{(\sigma)}(\mu,\nu)$ as a function of σ for fixed μ and ν. We establish continuity and non-increasing monotonicity. These, in particular, imply convergence of the optimal transportation costs, i.e., $\lim_{\sigma\to 0}\mathsf{W}_1^{(\sigma)}(\mu,\nu) = \mathsf{W}_1(\mu,\nu)$. Additionally, using the notion of Γ-convergence [Maso, 2012], we establish convergence of optimizing transport plans. Thus, if $(\pi_k)_{k\in\mathbb{N}}$ is sequence of optimal transport plans for $\mathsf{W}_1^{(\sigma_k)}(\mu,\nu)$, where $\sigma_k \to 0$, then $(\pi_k)_{k\in\mathbb{N}}$ converges weakly to an optimal plan for $\mathsf{W}_1(\mu,\nu)$.

Lastly, we explore the one-sample empirical approximation under GOT, i.e., the convergence rate of $\mathbb{E}\mathsf{W}_1^{(\sigma)}(\hat{\mu}_n,\mu)$. It was shown in [Goldfeld et al., 2020] that Gaussian smoothing alleviates the curse of dimensionality, with $\mathbb{E}\mathsf{W}_1^{(\sigma)}(\hat{\mu}_n,\mu)$ converging as $n^{-\frac{1}{2}}$ in all dimensions. Although GOT is specialized to Gaussian noise, we present a generalized empirical approximation result that accounts for any subgaussian noise density. This, in turn, implies fast convergence of $\mathbb{E}\big|\mathsf{W}_1^{(\sigma)}(\hat{\mu}_n,\nu) - \mathsf{W}_1^{(\sigma)}(\mu,\nu)\big|$ and $\mathbb{E}\big|\mathsf{W}_1^{(\sigma)}(\hat{\mu}_n,\hat{\nu}_n) - \mathsf{W}_1^{(\sigma)}(\mu,\nu)\big|$ via the triangle inequality. The expected value analysis is followed by a high probability claim derived through McDiarmid's inequality. Numerical results that validate these theoretical findings are provided. We conclude that GOT is an appealing alternative to entropic optimal transport, both in terms of its analytic and its statistical properties.

2 Notation and Preliminaries

Let $\mathcal{P}(\mathbb{R}^d)$ be the set of Borel probability measures on $\mathbb{R}^d$, while $\mathcal{P}_1(\mathbb{R}^d) \subset \mathcal{P}(\mathbb{R}^d)$ are those with finite first moments, i.e., $\int_{\mathbb{R}^d}\|x\|\,\mathsf{d}\mu(x) < \infty$, where $\|\cdot\|$ is the Euclidean norm. We denote by $\Pi(\mu,\nu) \subset \mathcal{P}(\mathbb{R}^d)$ the set of transport plans (or couplings) between measures $\mu,\nu \in \mathcal{P}(\mathbb{R}^d)$. Namely, any $\pi \in \Pi(\mu,\nu)$ is a probability

distribution is intractable to compute, fundamentally the GAN actually corresponds to a one-sample setup since infinite samples can be obtained from the generator network.

[3] $\mathsf{S}_c(\epsilon)$ can be transformed into a Sinkhorn divergence for which $\mathsf{S}_c^{(\epsilon)}(\mu,\mu) = 0$), but it still is not a metric [Bigot et al., 2019] since it lacks the triangle inequality.

measure on $\mathbb{R}^d \times \mathbb{R}^d$ whose first and second marginals are μ and ν, respectively.

The n-fold product extension of $\mu \in \mathcal{P}(\mathbb{R}^d)$ is $\mu^{\otimes n}$. The probability density function (PDF) of the isotropic Gaussian measure $\mathcal{N}_\sigma$ is φ_σ. Given $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$, their convolution $\mu * \nu \in \mathcal{P}(\mathbb{R}^d)$ is $(\mu * \nu)(\mathcal{A}) = \int\int \mathbb{1}_{\mathcal{A}}(x+y)\, \mathsf{d}\mu(x)\, \mathsf{d}\nu(y)$, where $\mathbb{1}_{\mathcal{A}}$ is the indicator of $\mathcal{A}$. For two independent random variables $X \sim \mu$ and $Y \sim \nu$, we have $X + Y \sim \mu * \nu$.

We use $\mathbb{E}_\mu f$ for the expectation of a measurable f w.r.t. μ, sometimes writing $\mathbb{E}_\mu f(X)$ to emphasize its dependence on $X \sim \mu$. When the underlying probability measure is clear from the context, the subscript is omitted. Accordingly, the characteristic function of $\mu \in \mathcal{P}(\mathbb{R}^d)$ is $\phi_\mu(t) \triangleq \mathbb{E}_\mu\big[e^{it^\top X}\big]$. For any $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$, we have $\phi_{\mu*\nu}(t) = \phi_\mu(t)\phi_\nu(t)$; if $\mu \times \nu \in \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d)$ is the product measure of μ and ν, then $\phi_{\mu\times\nu}(t,s) = \phi_\mu(t)\phi_\nu(s)$.

Definition 1 (Weak Topology) *The* weak topology *on $\mathcal{P}(\mathbb{R}^d)$ is induced by integration against the set $C_b^0(\mathbb{R}^d)$ of bounded and continuous functions. Accordingly, we say that $(\mu_k)_{k\in\mathbb{N}} \subset \mathcal{P}(\mathbb{R}^d)$ converges weakly to $\mu \in \mathcal{P}(\mathbb{R}^d)$, denoted by $\mu_k \rightharpoonup \mu$, if $\int_{\mathbb{R}^d} f(x)\, \mathsf{d}\mu_k(x) \to \int_{\mathbb{R}^d} f(x)\, \mathsf{d}\mu(x)$, for all $f \in C_b^0(\mathbb{R}^d)$.*

It is a well-known fact that $\big(\mathcal{P}_1(\mathbb{R}^d), \mathsf{W}_1\big)$ is a metric space, and that the 1-Wasserstein distance metrizes the weak topology (cf. [Villani, 2008, Thm. 6.9]). As shown in the sequel, this statement remains true if the 1-Wasserstein distance is replaced with its Gaussian-smoothed version, as defined next.

Definition 2 (Gaussian-Smoothed W_1) *The Gaussian-smoothed 1-Wasserstein distance between $\mu, \nu \in \mathcal{P}_1(\mathbb{R}^d)$ is $\mathsf{W}_1^{(\sigma)}(\mu,\nu) \triangleq \mathsf{W}_1(\mu * \mathcal{N}_\sigma, \nu * \mathcal{N}_\sigma)$.*

Letting $X \sim \mu$, $Y \sim \nu$ and $Z, Z' \sim \mathcal{N}_\sigma$ be independent random variables, $\mathsf{W}_1^{(\sigma)}(\mu,\nu)$ is the 1-Wasserstein distance between the probability laws of $X + Z \sim \mu * \mathcal{N}_\sigma$ and $Y + Z' \sim \nu * \mathcal{N}_\sigma$. Thus, $\mathsf{W}_1^{(\sigma)}(\mu,\nu)$ can be understood as a 'smoothed' version of W_1, where 'smoothing' is applied to the probability measures via convolution with a Gaussian kernel (or, equivalently, via additive white Gaussian noise).

The theoretical results in this paper are organized as follows. Section 3 studies the metric properties of $\mathsf{W}_1^{(\sigma)}$. Section 4 establishes properties of $\mathsf{W}_1^{(\sigma)}$ as a function of σ. One-sample empirical approximation rates under $\mathsf{W}_1^{(\sigma)}$ are explored in Section 5.

3 Metrizing the Weak Topology

Clearly, $\mathsf{W}_1^{(\sigma)}(\mu,\nu) < +\infty$, for any $\mu, \nu \in \mathcal{P}_1(\mathbb{R}^d)$. Furthermore, similar to the regular 1-Wasserstein distance, $\mathsf{W}_1^{(\sigma)}$ is a metric on $\mathcal{P}_1(\mathbb{R}^d)$, whose convergence is equivalent to convergence in the weak topology.

Theorem 1 (GOT Metric) *For any $\sigma \geq 0$, $\mathsf{W}_1^{(\sigma)} : \mathcal{P}_1(\mathbb{R}^d) \times \mathcal{P}_1(\mathbb{R}^d) \to [0, +\infty)$ is a metric on $\mathcal{P}_1(\mathbb{R}^d)$.*

This result mostly follows from W_1 being a metric. Some work is needed to establish the 'identity of indiscernibles' properties. See Section 7.1 for the proof.

Theorem 2 (Weak Topology Metrization) *Let $\sigma \geq 0$, $(\mu_k) \subset \mathcal{P}(\mathbb{R}^d)$ and $\mu \in \mathcal{P}(\mathbb{R}^d)$. Then $\mathsf{W}_1^{(\sigma)}(\mu_k,\mu) \to 0$ if and only if (iff) $\mathbb{E}_{\mu_k}\|X\| \to \mathbb{E}_\mu\|X\|$ and $\mu_k \rightharpoonup \mu$. Consequently, $\mathsf{W}_1^{(\sigma)}(\mu_k,\mu) \to 0$ iff $\mathsf{W}_1(\mu_k,\mu) \to 0$.*

Theorem 2 with W_1 in place of $\mathsf{W}_1^{(\sigma)}$ is a well-known result [Villani, 2008, Thm. 6.9]). The above can be therefore understood as the statement that 'the 1-Wasserstein topology is invariant to convolutions with Gaussian kernels'. See Section 7.2 for the proof.

4 Dependence on Noise Parameter

We study properties on $\mathsf{W}_1^{(\sigma)}(\mu,\nu)$, for fixed $\mu, \nu \in \mathcal{P}_1(\mathbb{R}^d)$, as a function of $\sigma \in [0, +\infty)$.

Theorem 3 (GOT Dependence on $\boldsymbol{\sigma}$) *Fix $\mu, \nu \in \mathcal{P}_1(\mathbb{R}^d)$. The following hold:*

i) $\mathsf{W}_1^{(\sigma)}(\mu,\nu)$ is continuous and monotonically non-increasing in $\sigma \in [0, +\infty)$;

ii) $\lim_{\sigma\to 0} \mathsf{W}_1^{(\sigma)}(\mu,\nu) = \mathsf{W}_1(\mu,\nu)$;

iii) $\lim_{\sigma\to\infty} \mathsf{W}_1^{(\sigma)}(\mu,\nu) \neq 0$, for some $\mu, \nu \in \mathcal{P}_1(\mathbb{R}^d)$.

While $\mathsf{W}_1^{(\sigma)}(\mu,\nu)$ is a monotonically non-increasing function of σ, as $\sigma \to \infty$ it is interestingly not true in general that $\mathsf{W}_1(\mu * \mathcal{N}_\sigma, \nu * \mathcal{N}_\sigma)$ decays to zero. The proof of Theorem 3 (Section 7.3) shows this via a simple Dirac measure example.

A key technical tool (that may be of independent interest) for establishing item (i) above is the following lemma, which ties GOT at different noise levels to one another. Its proof (Section 7.4) uses the Kantorovich-Rubinstein duality.

Lemma 1 (Stability Across σ) *Fix $\mu,\nu \in \mathcal{P}_1(\mathbb{R}^d)$, and $0 \le \sigma_1 < \sigma_2 < +\infty$. We have*

$$\mathsf{W}_1^{(\sigma_2)}(\mu,\nu) \le \mathsf{W}_1^{(\sigma_1)}(\mu,\nu) \le \mathsf{W}_1^{(\sigma_2)}(\mu,\nu) + 2\sqrt{d\,(\sigma_2^2-\sigma_1^2)}.$$

Theorem 3 established convergence of transport costs, i.e., that $\mathsf{W}_1^{(\sigma_k)}(\mu,\nu) \to \mathsf{W}_1^{(\sigma)}(\mu,\nu)$ as $\sigma_k \to \sigma$. The next result shows we also have convergence of optimal plans. Namely, a sequence of optimal couplings $(\pi_k)_{k\in\mathbb{N}}$ for $\mathsf{W}_1^{(\sigma_k)}(\mu,\nu)$ (weakly) approaches an optimal coupling for $\mathsf{W}_1^{(\sigma)}(\mu,\nu)$ as k goes to infinity.

Theorem 4 (Convergence of Optimal Plans) *Fix $\mu,\nu \in \mathcal{P}_1(\mathbb{R}^d)$ and let $(\sigma_k)_{k\in\mathbb{N}}$ be a sequence with $\sigma_k \searrow \sigma \ge 0$. Let $\pi_k \in \Pi(\mu * \mathcal{N}_{\sigma_k}, \nu * \mathcal{N}_{\sigma_k})$, $k \in \mathbb{N}$, be an optimal coupling for $\mathsf{W}_1^{(\sigma_k)}(\mu,\nu)$. Then there exists $\pi \in \Pi(\mu * \mathcal{N}_\sigma, \nu * \mathcal{N}_\sigma)$ such that $\pi_k \rightharpoonup \pi$ (weakly) as $k \to \infty$ and π is optimal for $\mathsf{W}_1^{(\sigma)}(\mu,\nu)$.*

The proof of Theorem 4 (Section 7.5) relies on the notion of Γ-convergence. Convergence of optimal transport plans then follows by standard tightness arguments. In particular, this theorem implies that a sequence of optimal transport plans for $\mathsf{W}_1^{(\sigma)}(\mu,\nu)$ converges to an optimal plan for the regular 1-Wasserstein distance $\mathsf{W}_1(\mu,\nu)$ as $\sigma \to 0$.

5 Empirical Approximation

We now explore statistical properties of $\mathsf{W}_1^{(\sigma)}$. In fact, our derivation accounts for any isotropic noise distribution $\mathcal{G}_\sigma$ that along each coordinate is σ-subgaussian with a bounded and monotone (in a proper sense) density.[4] Gaussian noise is captured as a special case.

Consider the fundamental one-sample empirical approximation, where $\mu \in \mathcal{P}_1(\mathbb{R}^d)$ is approximated by $\hat{\mu}_n \triangleq \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$, with $(X_1,\dots,X_n) \sim \mu^{\otimes n}$ and δ_x as the Dirac measure centered at x. We study how fast $\mathsf{W}_1^{(\mathcal{G}_\sigma)}(\hat{\mu}_n,\mu) \triangleq \mathsf{W}_1(\hat{\mu}_n * \mathcal{G}_\sigma, \mu * \mathcal{G}_\sigma) \to 0$ with n.[5] In a remarkable contrast to the 1-Wasserstein curse of dimensionality, we show $\mathbb{E}_{\mu^{\otimes n}}\mathsf{W}_1^{(\sigma)}(\hat{\mu}_n,\mu) \in O\big(n^{-\frac{1}{2}}\big)$ in all dimensions [Goldfeld et al., 2020], thus attaining the parametric rate.

To state the results, we first define subgaussianity.

Definition 3 (Subgaussian Measure) *A probability measure $\mu \in \mathcal{P}_1(\mathbb{R}^d)$ is K-subgaussian, for $K>0$, if for any $\alpha \in \mathbb{R}^d$, $X \sim \mu$ satisfies*

$$\mathbb{E}_\mu\left[e^{\alpha^T(X-\mathbb{E}X)}\right] \le e^{\frac{1}{2}K^2\|\alpha\|^2}. \tag{3}$$

[4] A further extension to nonisotropic noise is possible via similar techniques, but we do not delve into it here.

[5] Of course, $\mathsf{W}_1^{(\mathcal{N}_\sigma)}(\mu,\nu) = \mathsf{W}_1^{(\sigma)}(\mu,\nu)$.

We first bound the expected value and then give a high probability bound. The next theorem generalizes [Goldfeld et al., 2020, Prop. 1] to non-Gaussian noise.

Theorem 5 (GOT Empirical Approximation) *Fix $d \ge 1$, $\sigma > 0$ and $K > 0$. Let $\mathcal{G}_\sigma \in \mathcal{P}_1(\mathbb{R}^d)$ have a density g_σ that decomposes as $g_\sigma(x) = \prod_{j=1}^d \tilde{g}_\sigma(x_j)$. Assume that $\tilde{g}_\sigma$ is σ-subgaussian, bounded and monotonically decreases as its argument goes away from zero in either direction. For any K-subgaussian $\mu \in \mathcal{P}_1(\mathbb{R}^d)$, we have*

$$\mathbb{E}_{\mu^{\otimes n}}\mathsf{W}_1^{(\mathcal{G}_\sigma)}(\hat{\mu}_n,\mu) \le c_{\sigma,d,K}\, n^{-\frac{1}{2}}, \tag{4}$$

where $c_{\sigma,d,K} = e^{O(d)}$ is given in (20). In particular $\mathsf{W}_1^{(\sigma)}(\hat{\mu}_n,\mu) \in O\left(n^{-\frac{1}{2}}\right)$.

The proof of Theorem 5 is given in Section 7.6.

Corollary 1 (Concentration Inequality) *Under the paradigm of Theorem 5, denote $\mathcal{X} \triangleq \mathsf{supp}(\mu)$ and suppose $\mathsf{diam}(\mathcal{X}) < \infty$, where $\mathsf{diam}(\mathcal{X}) = \sup_{x\neq y\in\mathcal{X}}\|x-y\|$. For any $t>0$ we have*

$$\mathbb{P}_{\mu^{\otimes n}}\left(\left|\mathsf{W}_1^{(\mathcal{G}_\sigma)}(\hat{\mu}_n,\mu) - \mathbb{E}\mathsf{W}_1^{(\mathcal{G}_\sigma)}(\hat{\mu}_n,\mu)\right| \ge t\right) \le 2e^{-\frac{2t^2 n}{\mathsf{diam}(\mathcal{X})^2}} \tag{5}$$

and consequently,

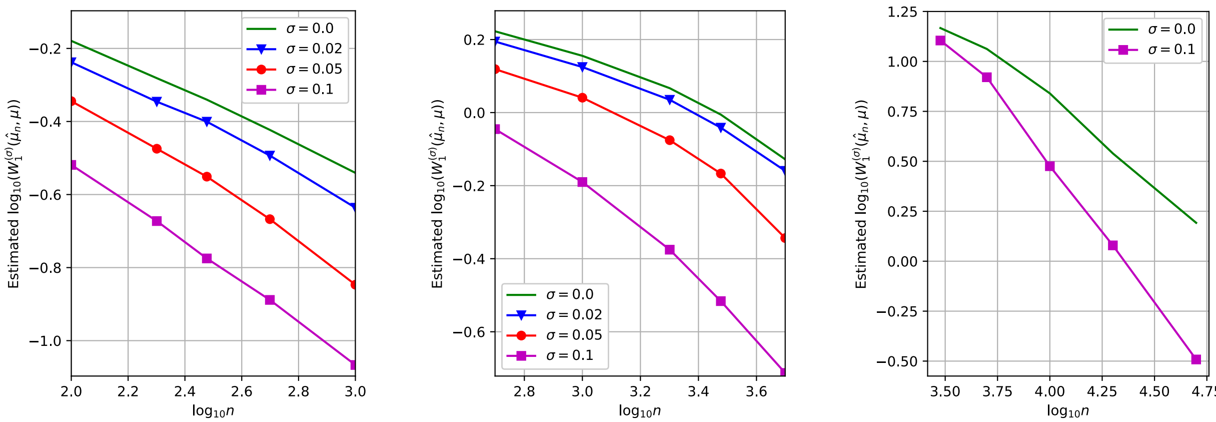
$$\mathbb{P}_{\mu^{\otimes n}}\left(\mathsf{W}_1^{(\mathcal{G}_\sigma)}(\hat{\mu}_n,\mu) \in \omega\left(\frac{\log n}{\sqrt{n}}\right)\right) \le \frac{1}{\mathsf{poly}(n)}. \tag{6}$$

The proof Theorem 1 is given in Section 7.7. It uses the W_1 duality and McDiarmid's inequality.

6 Empirical Results

We turn to some numerical experiments demonstrating the difference in empirical approximation convergence rates between the regular 1-Wasserstein distance and GOT. We compute $\mathsf{W}_1(\hat{\mu}_n,\mu)$ and $\mathsf{W}_1^{(\sigma)}(\hat{\mu}_n,\mu)$, for $\mu = \mathsf{Unif}\big([0,1]^d\big)$ the uniform measure on $[0,1]^d$, and $\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ the empirical measure based on i.i.d. samples $X_1,\dots,X_n \sim \mu$. This simple setup also hints at the breadth of the class of distributions for which $\mathsf{W}_1(\hat{\mu}_n,\mu)$ attains the poor convergence rate.

The GOT framework corresponds to the 1-Wasserstein distance between two continuous (smooth) distributions. To evaluate this 1-Wasserstein distance we chose to employ the neural network (NN) based dual optimization approach of [Gulrajani et al., 2017]. This approach seems to be better suited for continuous probability measures than, e.g., the Sinkhorm algorithm



(a) $d = 5$ (b) $d = 10$ (c) $d = 100$

Figure 1: Convergence of $\mathsf{W}_1^{(\sigma)}(\hat{\mu}_n, \mu)$ as a function of the number of samples n for various values of σ, shown in log-log space. The measure μ is the uniform distribution over $[0,1]^d$. Note that $\sigma = 0$ corresponds to the vanilla Wasserstein distance, which converges slower than GOT (note the difference in slopes), especially with larger d.

[Cuturi, 2013]. Starting from Kantorovich-Rubinstein duality

$$\mathsf{W}_1(\mu,\nu) = \sup_{\|f\|_{\mathsf{Lip}}\leq 1} \mathbb{E}_\mu f - \mathbb{E}_\nu f, \tag{7}$$

the function f is first parametrized by a NN f_θ, with parameter set $\theta \in \Theta$,[6] and then the $\|f_\theta\|_{\mathsf{Lip}\leq 1}$ constraint is relaxed to a regularization penalty on the expected gradient of $f_\theta(x)$ (w.r.t. to x). In sum, as in [Gulrajani et al., 2017], we use the ADAM stochastic gradient ascent method to optimize

$$\sup_{\theta\in\Theta} \mathbb{E}_\mu f_\theta - \mathbb{E}_\nu f_\theta + \lambda \mathbb{E}_\eta\Big[\big(\|\nabla_x f_\theta\| - 1\big)^2\Big], \tag{8}$$

where η interpolates between μ and ν in a manner compatible with the gradient penalty (GP) theoretical justification [Gulrajani et al., 2017, Prop. 1]. Specializing to $\mathsf{W}_1^{(\sigma)}$, μ and ν above are replaced with $\mu * \mathcal{N}_\sigma$ and $\nu * \mathcal{N}_\sigma$, respectively. To approximate expectations with empirical sums, we sample from these Gaussian-smoothed measures by adding (sampled) Gaussian noise to the original samples. This makes use of the fact that convolution of probability measures corresponds to sums of independent random variables.

Figure 1 shows the results for $d = 5$, $d = 10$, and $d = 100$, with each curve averaged over 10 random trials.[7] Note the slower decay of the $\sigma = 0$ case, which corresponds to vanilla W_1, compared to the approximately $O(n^{-1/2})$ rate of $W_1^{(\sigma)}$ for $\sigma > 0$. In particular, this divergence of rates increases as the dimension increases, as expected. In the $d = 10$ plot, the curves slightly accelerate as n increases instead of staying linear. This seems to originate from a two-fold imperfection in the NN-based approximation of the Lipschitz function f. First, the GP regularization does not perfectly enforce the Lipschitz constraint especially in high dimensions. Second, to accurately evaluate $\mathsf{W}_1^{(\sigma)}(\hat{\mu}_n, \mu)$ the network effectively needs to overfit $\hat{\mu}_n$. As NNs tend to avoid overfitting (especially once the number of modes n in $\hat{\mu}_n$ becomes large), additional slackness might be introduced.

As expected, the $\mathsf{W}_1(\hat{\mu}_n, \mu)$ estimate converges significantly slower than its Gaussian-smoothed counterpart, as evident by comparing the slopes of the curves in log-log space. In particular, the convergence of the $\mathsf{W}_1(\hat{\mu}_n, \mu)$ estimate is much slower for $d = 10$ than for $d = 5$ as predicted. The $\mathsf{W}_1^{(\sigma)}$ estimate, on the other hand, still converges approximately as $O(n^{-1/2})$ in both cases. The fact that $\mathsf{W}_1^{(\sigma)}$ is monotonically decreasing in σ can also be seen from the plots. These results are comparable with the ones from [Genevay et al., 2019] for two-sample empirical approximation of entropic OT.

7 Proofs

7.1 Proof of Theorem 1

The fact that $\mathsf{W}_1^{(\sigma)}(\mu, \nu)$ is symmetric, non-negative and equals zero when $\mu = \nu$ follows from its definition.

To prove the triangle inequality, i.e., $\mathsf{W}_1^{(\sigma)}(\mu_1, \mu_3) \leq$

[6] We used a fully connected DNN with 3 hidden ReLU layers, each of 1024 nodes. The network was trained until convergence of the estimated Wasserstein distance.

[7] Error bars were omitted since they were too small, and for $d = 100$ we restricted the values of σ to alleviate the computational burden.

$\mathsf{W}_1^{(\sigma)}(\mu_1,\mu_2) + \mathsf{W}_1^{(\sigma)}(\mu_2,\mu_3)$, for any $\mu_1,\mu_2,\mu_3 \in \mathcal{P}_1(\mathbb{R}^d)$, let $\pi_{12} \in \Pi(\mu_1 * \mathcal{N}_\sigma, \mu_2 * \mathcal{N}_\sigma)$ and $\pi_{23} \in \Pi(\mu_2 * \mathcal{N}_\sigma, \mu_3 * \mathcal{N}_\sigma)$ be optimal couplings for $\mathsf{W}_1^{(\sigma)}(\mu_1,\mu_2)$ and $\mathsf{W}_1^{(\sigma)}(\mu_2,\mu_3)$, respectively. Applying the Gluing Lemma [Villani, 2008], let $\pi \in \mathcal{P}_1(\mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^d)$ be a probability measure with π_{12} and π_{23} as its marginals on the corresponding coordinates. Letting $\pi_{13}(A \times B) \triangleq \pi(A \times \mathbb{R}^d \times B)$, we have $\pi_{13} \in \Pi(\mu_1,\mu_3)$ and

$$\begin{aligned}\mathsf{W}_1^{(\sigma)}(\mu_1,\mu_3) &\leq \mathbb{E}_{\pi_{13}}\|X_1 - X_3\| \\ &\leq \mathbb{E}_{\pi_{12}}\|X_1 - X_2\| + \mathbb{E}_{\pi_{23}}\|X_2 - X_3\| \\ &= \mathsf{W}_1^{(\sigma)}(\mu_1,\mu_2) + \mathsf{W}_1^{(\sigma)}(\mu_2,\mu_3). \qquad (9)\end{aligned}$$

It remains to show that $\mathsf{W}_1^{(\sigma)}(\mu,\nu) = 0$ implies that $\mu = \nu$. Since W_1 is a metric, we know that if $\mathsf{W}_1^{(\sigma)}(\mu,\nu) = 0$ then $\mu * \mathcal{N}_\sigma = \nu * \mathcal{N}_\sigma$. This implies pointwise equality between characteristic functions: $\phi_\mu \phi_{\mathcal{N}_\sigma} = \phi_\nu \phi_{\mathcal{N}_\sigma}$. Since $\phi_{\mathcal{N}_\sigma} \neq 0$ everywhere, we get $\phi_\mu = \phi_\nu$ pointwise, implying $\mu = \nu$.

7.2 Proof of Theorem 2

The claim relies on the equivalence between weak convergence and pointwise convergence of characteristic functions. Since W_1 metrizes weak convergence:

$$\begin{aligned}&\mathsf{W}_1^{(\sigma)}(\mu_k,\mu) \to 0 \\ \iff &\quad \mu_k * \mathcal{N}_\sigma \to \mu * \mathcal{N}_\sigma \\ \iff &\quad \phi_{\mu_k}(t)\phi_{\mathcal{N}_\sigma}(t) = \phi_\mu(t)\phi_{\mathcal{N}_\sigma}(t), \quad \forall t \in \mathbb{R}^d \\ \iff &\quad \phi_{\mu_k}(t) = \phi_\mu(t), \quad \forall t \in \mathbb{R}^d.\end{aligned}$$

7.3 Proof of Theorem 3

For Claim (ii), the fact that $\lim_{\sigma\to 0}\mathsf{W}_1^{(\sigma)}(\mu,\nu) = \mathsf{W}_1(\mu,\nu)$ follows from Lemma 1 by taking $\sigma_1 = 0$ and $\sigma_2 = \sigma \to 0$.

For Claim (i), $\mathsf{W}_1^{(\sigma)}(\mu,\nu)$ being monotonically non-increasing in σ also follows directly from Lemma 1. To prove continuity at $\sigma \in (0,+\infty)$, we consider left- and right- continuity separately. Let $\sigma_k \nearrow \sigma$ as $k \to \infty$. Lemma 1 gives

$$\mathsf{W}_1^{(\sigma)}(\mu,\nu) \leq \mathsf{W}_1^{(\sigma_k)}(\mu,\nu) \leq \mathsf{W}_1^{(\sigma)}(\mu,\nu) + 2d\sqrt{\sigma^2 - \sigma_k^2},$$

and left-continuity follows.

To see that $\mathsf{W}_1^{(\sigma)}(\mu,\nu)$ is right-continuous in σ, let $\sigma_k \searrow \sigma$ and denote $\epsilon_k \triangleq \sqrt{\sigma_k^2 - \sigma^2}$. We have

$$\mathsf{W}_1^{(\sigma_k)}(\mu,\nu) = \mathsf{W}_1^{(\epsilon_k)}(\mu * \mathcal{N}_\sigma, \nu * \mathcal{N}_\sigma) \xrightarrow[k\to\infty]{} \mathsf{W}_1^{(\sigma)}(\mu,\nu),$$

where the last step uses $\mathsf{W}_1^{(\sigma)}$ continuity at $\sigma = 0$.

Moving to Claim (iii), let $\mu = \delta_x$ and $\nu = \delta_y$ be two Dirac measures at $x \neq y \in \mathbb{R}^d$. For any $\sigma \in [0,+\infty)$,

$$\begin{aligned}\mathsf{W}_1^{(\sigma)}(\mu,\nu) &= \mathsf{W}_1\big(\mathcal{N}(x,\sigma^2\mathrm{I}_d), \mathcal{N}(y,\sigma^2\mathrm{I}_d)\big) \\ &\geq \Big\|\mathbb{E}_{\mathcal{N}(x,\sigma^2\mathrm{I}_d)}X - \mathbb{E}_{\mathcal{N}(y,\sigma^2\mathrm{I}_d)}Y\Big\| \\ &= \|x - y\|,\end{aligned}$$

where the equality uses Jensen's inequality and convexity of norms.

7.4 Proof of Lemma 1

The first inequality immediately follows because W_1 is non-increasing under convolutions and since $\mathcal{N}_{\sigma_2} = \mathcal{N}_{\sigma_1} * \mathcal{N}_{\sqrt{\sigma_2^2 - \sigma_1^2}}$.

For the second inequality, we use Kantorovich-Rubinstein duality to write

$$\begin{aligned}\mathsf{W}_1^{(\sigma_1)}(\mu,\nu) &= \sup_{\|f_1\|_{\mathsf{Lip}}\leq 1} \mathbb{E}_{\mu*\mathcal{N}_{\sigma_1}} f_1 - \mathbb{E}_{\nu*\mathcal{N}_{\sigma_1}} f_1; \\ \mathsf{W}_1^{(\sigma_2)}(\mu,\nu) &= \sup_{\|f_2\|_{\mathsf{Lip}}\leq 1} \mathbb{E}_{\mu*\mathcal{N}_{\sigma_2}} f_2 - \mathbb{E}_{\nu*\mathcal{N}_{\sigma_2}} f_2.\end{aligned}$$

Letting $f_1^\star$ be optimal for $\mathsf{W}_1^{(\sigma_1)}(\mu,\nu)$, we have

$$\mathsf{W}_1^{(\sigma_2)}(\mu,\nu) \geq \mathbb{E}_{\mu*\mathcal{N}_{\sigma_2}} f_1^\star - \mathbb{E}_{\nu*\mathcal{N}_{\sigma_2}} f_1^\star. \qquad (10)$$

Set $X \sim \mu$, $Z_1 \sim \mathcal{N}_{\sigma_1}$ and $Z_{21} \sim \mathcal{N}_{\sqrt{\sigma_2^2 - \sigma_1^2}}$ as independent random variables; clearly, $Z_2 \triangleq Z_1 + Z_{21} \sim \mathcal{N}_{\sigma_2}$. Consider:

$$\begin{aligned}\big|\mathbb{E}_{\mu*\mathcal{N}_{\sigma_1}} f_1^\star - \mathbb{E}_{\mu*\mathcal{N}_{\sigma_2}} f_1^\star\big| &= \mathbb{E}f_1^\star(X + Z_1) - \mathbb{E}f_1^\star(X + Z_2) \\ &\leq \mathbb{E}\|Z_{21}\| \\ &= \sqrt{d\,(\sigma_2^2 - \sigma_1^2)}, \qquad (11a)\end{aligned}$$

where the last in equality uses $\|f_1^\star\|_{\mathsf{Lip}} \leq 1$. Similarly, one has

$$\big|\mathbb{E}_{\nu*\mathcal{N}_{\sigma_1}} f_1^\star - \mathbb{E}_{\nu*\mathcal{N}_{\sigma_2}} f_1^\star\big| \leq \sqrt{d\,(\sigma_2^2 - \sigma_1^2)}. \qquad (11b)$$

Inserting (11) into (10) concludes the proof.

7.5 Proof of Theorem 4

We first include the definitions of tightness of measures and Γ-convergence of functionals.

Definition 4 (Tightness of Measures) *A subset $\mathcal{S} \subset \mathcal{P}(\mathbb{R}^d)$ is tight if for any $\epsilon > 0$ there is a compact set $\mathcal{K}_\epsilon \subset \mathbb{R}^d$ such that $\mu(\mathcal{K}_\epsilon) \geq 1 - \epsilon$, for all $\mu \in \mathcal{P}(\mathbb{R})^d$.*

Definition 5 (Γ-Convergence) *Let $\mathcal{X}$ be a metric space and $\mathsf{F}_k : \mathcal{X} \to \mathbb{R}$, $k \in \mathbb{N}$ be a sequence of functionals. We say $(\mathsf{F}_k)_{k\in\mathbb{N}}$ Γ-converges to $\mathsf{F} : \mathcal{X} \to \mathbb{R}$, and we write $\mathsf{F}_k \xrightarrow{\Gamma} \mathsf{F}$, if:*

i) For every $x_k, x \in \mathcal{X}$, $k \in \mathbb{N}$, with $x_k \to x$, we have $\mathsf{F}(x) \le \liminf_{k\to\infty} \mathsf{F}_k(x_k)$*;*

ii) For any $x \in \mathcal{X}$, there exists $x_k \in \mathcal{X}$, $k \in \mathbb{N}$, with $x_k \to x$, and $\mathsf{F}(x) \ge \limsup_{k\to\infty} \mathsf{F}_k(x_k)$

By pointwise convergence of characteristic functions, $P_k \triangleq \mu * \mathcal{N}_{\sigma_k}$ and $Q_k \triangleq \nu * \mathcal{N}_{\sigma_k}$ are weakly convergent measures on $\mathbb{R}^d$. Prokhorov's Theorem then implies they are tight. By [Villani, 2008, Lemma 4.4] we have that $\Pi\Big((P_k)_{k\in\mathbb{N}}, (Q_k)_{k\in\mathbb{N}}\Big)$, the set of all couplings with marginals in $(P_k)_{k\in\mathbb{N}}$ and $(Q_k)_{k\in\mathbb{N}}$ is also tight. Hence, the sequence of optimal couplings $(\pi_k)_{k\in\mathbb{N}}$ is tight and weakly converges to some $\pi \in \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d)$. Taking the limit of the relation $\pi_k \in \Pi(P_k, Q_k)$ we obtain $\pi \in \Pi(P, Q)$, where $P \triangleq \mu * \mathcal{N}_\sigma$ and $Q \triangleq \nu * \mathcal{N}_\sigma$.

With that in mind, recall that if $(\mathsf{F}_k)_{k\in\mathbb{N}}$ Γ-converges to F, then $\lim_{k\to\infty} \inf \mathsf{F}_k = \inf \mathsf{F}$ [Maso, 2012, Thm. 7.8]. Furthermore, if $(x_k)_{k\in\mathbb{N}}$ is a sequence of minimizers of F_k, for each $k \in \mathbb{N}$, then any cluster (limit) point of $(x_k)_{k\in\mathbb{N}}$ is a minimizer of $\mathcal{F}$ [Maso, 2012, Cor. 7.20]. Thus, to conclude the proof of Theorem 4 it suffices to establish Γ-convergence of $\mathsf{F}_k : \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d) \to \mathbb{R} \cup \{\infty\}$ to $\mathsf{F} : \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d) \to \mathbb{R} \cup \{\infty\}$ defined as

$$\mathsf{F}_k(\pi) = \begin{cases} \mathbb{E}_\pi \|X - Y\|, & \pi \in \Pi(\mu * \mathcal{N}_{\sigma_k}, \nu * \mathcal{N}_{\sigma_k}) \\ \infty, & \text{otherwise} \end{cases}$$
$$\mathsf{F}(\pi) = \begin{cases} \mathbb{E}_\pi \|X - Y\|, & \pi \in \Pi(\mu * \mathcal{N}_{\sigma}, \nu * \mathcal{N}_{\sigma}) \\ \infty, & \text{otherwise} \end{cases}.$$

We start with the $\liminf$ Γ-convergence inequality. First observe that if $(\pi_k)_{k\in\mathbb{N}}$ does not contain a subsequence (without relabeling) such that $\pi_k \in \Pi(\mu * \mathcal{N}_{\sigma_k}, \nu * \mathcal{N}_{\sigma_k})$, then the claim is trivial. Accordingly, assume (again, up to extraction of subsequences) that $\pi_k \in \Pi(\mu * \mathcal{N}_{\sigma_k}, \nu * \mathcal{N}_{\sigma_k})$, for all $k \in \mathbb{N}$. Since $x \mapsto \|x\|$ is a non-negative and continuous, the $\liminf$ condition directly follows from the Portmanteau Theorem:

$$\mathsf{F}(\pi) \le \liminf_{k\to\infty} \int \|x - y\| d\pi_k = \liminf_{k\to\infty} \mathsf{F}_k(\pi_k). \quad (12)$$

For the $\limsup$ let $\pi \in \Pi(\mu * \mathcal{N}_\sigma, \nu * \mathcal{N}_\sigma)$. For convenience, we use random variable notation. There exists a tuple (X, Y, Z', Z'') with marginal distributions $X \sim \mu$, $Y \sim \nu$ and $Z', Z'' \sim \mathcal{N}_\sigma$, such that (X, Z') are independent, (Y, Z'') are independent, and $(X + Z', Y + Z'') \sim \pi$.

To construct the sequence $(\pi_k)_{k\in\mathbb{N}}$, let $Z_k \sim \mathcal{N}_{\sqrt{\sigma_k^2 - \sigma^2}}$ be independent of (X, Y, Z', Z''). Setting π_k as the joint probability law of $(X + Z' + Z_k, Y + Z'' + Z_k)$, we have $\pi_k \in \Pi(\mu * \mathcal{N}_{\sigma_k}, \nu * \mathcal{N}_{\sigma_k})$, $k \in \mathbb{N}$. Evaluating F_k we obtain

$$\mathsf{F}_k(\pi_k) = \mathbb{E}\|X + Z' - Y - Z''\| = \mathsf{F}(\pi), \quad (13)$$

which in particular implies the $\limsup$ condition.

7.6 Proof of Theorem 5

The 1-Wasserstein distance is upper bounded by weighted total variation (TV) as follows [Villani, 2008, Theorem 6.15]:

$$\mathsf{W}_1(\hat{\mu}_n * \mathcal{G}_\sigma, \mu * \mathcal{G}_\sigma) \le \int_{\mathbb{R}^d} \|t\| \big|r_n(t) - q(t)\big| \, \mathsf{d}t, \quad (14)$$

where r_n and q are the densities of $\hat{\mu}_n * \mathcal{G}_\sigma$ and $\mu * \mathcal{G}_\sigma$, respectively. The inequality is proved using the maximal TV coupling of $\hat{\mu}_n * \mathcal{G}_\sigma$ with $\mu * \mathcal{G}_\sigma$.

Let $a > 0$ (to be specified later) and set $f_a : \mathbb{R}^d \to \mathbb{R}$ as the density of $\mathcal{N}\left(0, \frac{1}{2a}\mathrm{I}_d\right)$. By Cauchy-Schwarz, we have

$$\begin{aligned} &\mathbb{E}_{\mu^{\otimes n}} \int_{\mathbb{R}^d} \|t\| \big|r_n(t) - q(t)\big| \, \mathsf{d}t \\ &\le \left(\int_{\mathbb{R}^d} \|t\|^2 f_a(t) \, \mathsf{d}t \right)^{\frac{1}{2}} \left(\mathbb{E}_{\mu^{\otimes n}} \int_{\mathbb{R}^d} \frac{\big(q(t) - r_n(t)\big)^2}{f_a(t)} \, \mathsf{d}t \right)^{\frac{1}{2}}. \end{aligned} \quad (15)$$

The first term equals $\frac{d}{2a}$. Turning to the second integral, note that $r_n(t) = \frac{1}{n}\sum_{i=1}^n g_\sigma(t - X_i)$, where $\{X_i\}_{i=1}^n$ are i.i.d. and $\mathbb{E}_\mu g_\sigma(t - X_i) = q(t)$. Using the definition of subgaussianity (Definition 3), we have the following lemma (proven in Appendix **??**) that bounds g_σ everywhere in terms of the Gaussian density φ_σ.

Lemma 2 *Let $\delta \triangleq \min\left\{1, \frac{1}{4\sigma^2}\right\}$. There exists a constant $c_1 > 0$ such that*

$$g_\sigma(t) \le c_1^d e^{\delta\|t\|^2} \varphi_\sigma(t), \quad \forall t \in \mathbb{R}^d. \quad (16)$$

We now can bound the second integrand of (15):

$$\begin{aligned} \mathbb{E}_{\mu^{\otimes n}} \big(q(z) - r_n(z)\big)^2 &= \mathsf{var}_{\mu^{\otimes n}} \big(r_n(z)\big) \\ &= \mathsf{var}_{\mu^{\otimes n}} \left(\frac{1}{n} \sum_{i=1}^n g_\sigma(z - X_i) \right) \\ &= \frac{1}{n} \mathsf{var}_\mu \big(g_\sigma(z - X)\big) \\ &\le \mathbb{E}_\mu g_\sigma^2(z - X) \\ &\le c_1^d \delta^{2d} \mathbb{E}_\mu e^{2\delta\|z - X\|^2} \varphi_\sigma^2(z - X) \\ &\le \frac{c_2^2}{n} \mathbb{E}_\mu e^{-\frac{1}{2\sigma^2}\|z - X\|^2}, \end{aligned} \quad (17)$$

with $c_2 = c_1^d (2\pi\sigma^2)^{-d/2}$. This further implies

$$\int_{\mathbb{R}^d} \mathbb{E}_{\mu^{\otimes n}} \frac{\big(q(t) - r_n(t)\big)^2}{f_a(t)} \, \mathsf{d}z \le \frac{c_2}{n 2^{d/2}} \mathbb{E} \frac{1}{f_a(X + Z)}, \quad (18)$$

where $X \sim \mu$ and $Z \sim \mathcal{N}_\sigma$ are independent.

Starting from (18), we finish the proof via steps similar to [Goldfeld et al., 2020]. Specifically, for $c_3 \triangleq \left(\frac{\pi}{a}\right)^{\frac{d}{2}}$, it holds that $\left(f_a(t)\right)^{-1} = c_3 e^{a\|t\|^2}$. Since X is K-subgaussian and Z is σ-subgaussian, $X+Z$ is $(K+\sigma)$-subgaussian. Following (18), for any $0 < a < \frac{1}{2(K+\sigma)^2}$, we have [Hsu et al., 2012, Rmk. 2.3]

$$\begin{aligned}
&\frac{c_2}{n2^{d/2}}\mathbb{E}\frac{1}{f(X+Z)} \\
&= \frac{c_2 c_3}{n2^{d/2}}\mathbb{E}\exp\left(a\left\|X+Z\right\|^2\right) \qquad (19)\\
&\leq \frac{c_2 c_3}{n2^{d/2}}\exp\left(\left(K+\sigma\right)^2 ad + \frac{(K+\sigma)^4 a^2 d}{1-2(K+\sigma)^2 a}\right),
\end{aligned}$$

Setting $a = \frac{1}{4(K+\sigma)^2}$ and combining (15)-(19) yields

$$\mathbb{E}_{\mu^{\otimes n}}\mathsf{W}_1^{(\mathcal{G}_\sigma)}(\hat{\mu}_n,\mu) \leq c_1^d \sigma\sqrt{2d}\left(1+\frac{K}{\sigma}\right)^{\frac{d}{2}+1} e^{\frac{3d}{16}}\frac{1}{\sqrt{n}}, \qquad (20)$$

where c_1 is the constant from Lemma 2. We note that a better constant can be achieved by assuming $\mathcal{G}_\sigma = \mathcal{N}_\sigma$ [Goldfeld et al., 2020], but we chose to sacrifice that in favor of generality.

7.7 Proof of Corollary 1

The main tool we use is McDiarmid's inequality:

Lemma 3 (McDiarmid's Inequality) *Let $X^n \triangleq (X_1,\ldots,X_n)$ be an n-tuple of $\mathcal{X}$-valued independent random variables. Suppose $g : \mathcal{X}^n \to \mathbb{R}$ is a map that for any $i = 1,\ldots,n$ and $x_1,\ldots,x_n,x_i' \in \mathcal{X}$ satisfies*

$$\left|g(x^n) - g(x_1,\ldots,x_{i-1},x_i',x_{i+1},\ldots,x_n)\right| \leq c_i, \quad (21)$$

for some non-negative $\{c_i\}_{i=1}^n$. Then for any $t > 0$:

$$\mathbb{P}\Big(g(X^n) - \mathbb{E}g(X^n) \geq t\Big) \leq e^{-\frac{2t^2}{\sum_{i=1}^n c_i^2}} \qquad (22a)$$

$$\mathbb{P}\Big(\left|g(X^n) - \mathbb{E}g(X^n)\right| \geq t\Big) \leq 2e^{-\frac{2t^2}{\sum_{i=1}^n c_i^2}} \qquad (22b)$$

Let $g(X^n) \triangleq \mathsf{W}_1^{(\mathcal{G}_\sigma)}(\hat{\mu}_n,\mu)$ and use Kantorovich-Rubinstein duality:

$$\begin{aligned}
g(X^n) &= \sup_{\|f\|_{\mathsf{Lip}}\leq 1} \mathbb{E}_{\hat{\mu}_n * \mathcal{G}_\sigma} f - \mathbb{E}_{\mu*\mathcal{G}_\sigma} f \\
&= \sup_{\|f\|_{\mathsf{Lip}}\leq 1} \frac{1}{n}\sum_{i=1}^n (f*g_\sigma)(X_i) - \mathbb{E}_\mu\left[f*g_\sigma\right].
\end{aligned}$$

Fix $i \in \{1,\ldots,n\}$ and $x_1,\ldots,x_n,x_i' \in \mathcal{X}$. Property (21) follows by first observing that:

$$\begin{aligned}
&n\Big(g(x^n) - g(x_1,\ldots,x_{i-1},x_i',x_{i+1},\ldots,x_n)\Big) \\
&= \sup_{\|f\|_{\mathsf{Lip}}\leq 1}\left\{\sum_{j\neq i}(f*g_\sigma)(x_j) - \mathbb{E}_\mu\left[f*g_\sigma\right] + (f*g_\sigma)(x_i)\right\} \\
&- \sup_{\|h\|_{\mathsf{Lip}}\leq 1}\left\{\sum_{j\neq i}(h*g_\sigma)(x_j) - \mathbb{E}_\mu\left[h*g_\sigma\right] + (h*g_\sigma)(x_i')\right\} \\
&\leq \sup_{\|f\|_{\mathsf{Lip}}\leq 1} (f*g_\sigma)(x_i) - (f*g_\sigma)(x_i'). \qquad (23)
\end{aligned}$$

Then we note that Lipschitzness of f implies that $f*g_\sigma$ is also Lipschitz.

Lemma 4 (Lipschitz after Convolution) *If $f : \mathbb{R}^d \to \mathbb{R}$ has $\|f\|_{\mathsf{Lip}} \leq L$, then $\|f*g\|_{\mathsf{Lip}} \leq L$ for any PDF $g : \mathbb{R}^d \to \mathbb{R}_{\geq 0}$.*

The proof is immediate and thus omitted. Combining Lemma 4 and (23), we obtain

$$\Big|g(x^n) - g(x_1,\ldots,x_{i-1},x_i',x_{i+1},\ldots,x_n)\Big| \leq \frac{\mathsf{diam}(\mathcal{X})^2}{n},$$

for all $i = 1,\ldots,n$ and $x_1,\ldots,x_n,x_i' \in \mathcal{X}$.

Applying McDiarmid's inequality (22b) for $g(X^n) = \mathsf{W}_1^{(\mathcal{G}_\sigma)}(\hat{\mu}_n,\mu)$ produces (5). Taking $t = \Theta\left(\frac{\log n}{\sqrt{n}}\right)$ and inserting into (22a) gives (6).

8 Summary and Concluding Remarks

We proposed a novel Gaussian-smoothed framework for OT defined as $\mathsf{W}_1^{(\sigma)}(\mu,\nu) \triangleq \mathsf{W}_1(\mu*\mathcal{N}_\sigma,\nu*\mathcal{N}_\sigma)$. This GOT distance was shown to inherit the metric structure (and the metrization of weak convergence) from the regular 1-Wasserstein distance. As a function of σ, $\mathsf{W}_1^{(\sigma)}(\mu,\nu)$ is a continuous and monotonically decreasing function maximized at $\mathsf{W}_1^{(0)}(\mu,\nu) = \mathsf{W}_1(\mu,\nu)$. Furthermore, as $\mathsf{W}_1^{(\sigma)}(\mu,\nu) \xrightarrow[\sigma\to 0]{} \mathsf{W}_1(\mu,\nu)$, optimal transport plans for $\mathsf{W}_1^{(\sigma)}(\mu,\nu)$ weakly converge to an optimal plan for $\mathsf{W}_1(\mu,\nu)$. Finally, we explored statistical properties of $\mathsf{W}_1^{(\sigma)}$, studying the convergence rate of $\mathbb{E}\mathsf{W}_1^{(\sigma)}(\hat{\mu}_n,\mu)$ to 0, where $\hat{\mu}_n$ is the empirical measure induced by n i.i.d. samples from μ. Building on [Goldfeld et al., 2020], we showed that $\mathsf{W}_1(\hat{\mu}_n*\mathcal{G}_\sigma,\mu*\mathcal{G}_\sigma) \in O\left(n^{-\frac{1}{2}}\right)$ in all dimensions, for any subgaussian noise distribution $\mathcal{G}_\sigma$ with a monotone and bounded density. In particular, $\mathsf{W}_1^{(\sigma)}$ alleviates the curse of dimensionality in the one-sample (and hence also in the weaker two-sample) regime. This stands in striking contrast to the classic 1-Wasserstein distance, which converges at most as $n^{-\frac{1}{d}}$, while no results are available for entropic OT with distance cost. These theoretical findings were verified through an empirical study, posing GOT as an appealing alternative to the popular entropically regularized OT methods.

References

[Altschuler et al., 2017] Altschuler, J., Weed, J., and Rigollet, P. (2017). Near-linear time approximation algorithms for optimal transport via Sinkhorn iteration. In *Advances in Neural Information Processing Systems (NeurIPS-2017)*, pages 1964–1974, Long Beach, CA, US.

[Alvarez-Melis and Jaakkola, 2018] Alvarez-Melis, D. and Jaakkola, T. (2018). Gromov-Wasserstein alignment of word embedding spaces. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1881–1890.

[Arjovsky et al., 2017] Arjovsky, M., Chintala, S., and Bottou, L. (2017). Wasserstein generative adversarial networks. In *International Conference on Machine Learning (ICML-2017)*, pages 214–223, Sydney, Australia.

[Bigot et al., 2019] Bigot, J., Cazelles, E., and Papadakis, N. (2019). Central limit theorems for entropy-regularized optimal transport on finite spaces and statistical applications. *Electronic Journal of Statistics*, 13(2):5120–5150.

[Courty et al., 2016] Courty, N., Flamary, R., D., Tuia, and Rakotomamonjy, A. (2016). Optimal transport for domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(9):1853–1865.

[Courty et al., 2014] Courty, N., Flamary, R., and Tuia, D. (2014). Domain adaptation with regularized optimal transport. In *European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD 2014)*, pages 274–289, Nancy, France.

[Cuturi, 2013] Cuturi, M. (2013). Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems (NeurIPS-2013)*, pages 2292–2300, Stateline, NV, US.

[Dobrić and Yukich, 1995] Dobrić, V. and Yukich, J. E. (1995). Asymptotics for transportation cost in high dimensions. *J. Theoretical Prob.*, 8(1):97–118.

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

[Feydy et al., 2019] Feydy, J., Séjourné, T., Vialard, F.-X., Amari, S.-I., Trouve, A., and Peyré, G. (2019). Interpolating between optimal transport and MMD using Sinkhorn divergences. In *International Conference on Artificial Intelligence and Statistics (AISTATS-2019)*, pages 2681–2690.

[Genevay et al., 2019] Genevay, A., Chizat, L., Bach, F., Cuturi, M., and Peyré, G. (2019). Sample complexity of Sinkhorn divergences. In *International Conference on Artificial Intelligence and Statistics (AISTATS-2019)*, pages 1574–1583, Okinawa, Japan.

[Goldfeld et al., 2020] Goldfeld, Z., Greenewald, K., Niles-Weed, J., and Polyanskiy, Y. (2020). Convergence of smoothed empirical measures with applications to entropy estimation. *IEEE Transactions on Information Theory*.

[Grave et al., 2019] Grave, E., Joulin, A., and Berthet, Q. (2019). Unsupervised alignment of embeddings with Wasserstein Procrustes. In *International Conference on Artificial Intelligence and Statistics (AISTATS-2019)*, pages 1880–1890, Okinawa, Japan.

[Gulrajani et al., 2017] Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., and Courville, A. C. (2017). Improved training of Wasserstein GANs. In *Advances in Neural Information Processing Systems (NeurIPS-2017)*, pages 5767–5777, Long Beach, CA, US.

[Ho et al., 2017] Ho, N., Nguyen, X. L., Yurochkin, M., Bui, H. H., Huynh, V., and Phung, D. (2017). Multilevel clustering via Wasserstein means. In *International Conference on Machine Learning (ICML-2017)*, pages 1501–1509, Sydney, Australia.

[Hsu et al., 2012] Hsu, D., Kakade, S., and Zhang, T. (2012). A tail inequality for quadratic forms of subgaussian random vectors. *Electronic Communications in Probability*, 17.

[Iyer et al., 2014] Iyer, A., Nath, S., and Sarawagi, S. (2014). Maximum mean discrepancy for class ratio estimation: Convergence bounds and kernel selection. In *International Conference on Machine Learning*, pages 530–538.

[Kantorovich, 1942] Kantorovich, L. V. (1942). On the translocation of masses. In *USSR Academy of Science (Doklady Akademii Nauk USSR*, volume 37, pages 199–201.

[Li et al., 2013] Li, P., Wang, Q., and Zhang, L. (2013). A novel earth mover's distance methodology for image matching with Gaussian mixture models. In *IEEE International Conference on Computer Vision (ICCV-2013)*, pages 1689–1696, Sydney, Australia.

[M. et al., 2017] M., A. G., Cuturi, Peyré, G., and Bach, F. (2017). Stochastic optimization for large-scale optimal transport. In *Advances in Neural Information Processing Systems (NeurIPS-2016)*, pages 3440–3448, Barcelona, Spain.

[Maso, 2012] Maso, G. D. (2012). *An introduction to* Γ*-convergence*, volume 8. Springer Science & Business Media.

[Mena and Niles-Weed, 2019] Mena, G. and Niles-Weed, J. (2019). Statistical bounds for entropic optimal transport: sample complexity and the central limit theorem. In *Advances in Neural Information Processing Systems (NeurIPS-2019)*, pages 4543–4553.

[Montavon et al., 2016] Montavon, G., Müller, K.-R., and Cuturi, M. (2016). Wasserstein training of restricted Boltzmann machines. In *Advances in Neural Information Processing Systems (NeurIPS-2016)*, pages 3718–3726, Barcelona, Spain.

[Panaretos and Zemel, 2019] Panaretos, V. M. and Zemel, Y. (2019). Statistical aspects of Wasserstein distances. *Annual Review of Statistics and its Application*, 6:405–431.

[Rigollet and Weed, 2018] Rigollet, P. and Weed, J. (2018). Entropic optimal transport is maximum-likelihood deconvolution. *Comptes Rendus Mathematique*, 356(11-12):1228–1235.

[Rubner et al., 2000] Rubner, Y., Tomasi, C., and Guibas, L. J. (2000). The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision*, 40(2):99–121.

[Sandler and Lindenbaum, 2011] Sandler, R. and Lindenbaum, M. (2011). Nonnegative matrix factorization with earth mover's distance metric for image analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(8):1590–1602.

[Varadarajan, 1958] Varadarajan, V. S. (1958). On the convergence of sample probability distributions. *Sankhyā: The Indian Journal of Statistics (1933-1960)*, 19(1/2):23–26.

[Villani, 2008] Villani, C. (2008). *Optimal transport: old and new*, volume 338. Springer Science & Business Media.

[Yurochkin et al., 2019] Yurochkin, M., Claici, S., Chien, E., Mirzazadeh, F., and Solomon, J. (2019). Hierarchical optimal transport for document representation. In *Advances in Neural Information Processing Systems (NeurIPS-2019)*, pages 1599–1609.