Smooth p-Wasserstein Distance: Structure, Empirical Approximation, and Statistical Applications

Sloan Nietert [1] **Ziv Goldfeld** [2] **Kengo Kato** [3]

Abstract

Discrepancy measures between probability distributions, often termed statistical distances, are ubiquitous in probability theory, statistics and machine learning. To combat the curse of dimensionality when estimating these distances from data, recent work has proposed smoothing out local irregularities in the measured distributions via convolution with a Gaussian kernel. Motivated by the scalability of this framework to high dimensions, we investigate the structural and statistical behavior of the Gaussian-smoothed p-Wasserstein distance $\mathsf{W}_p^{(\sigma)}$, for arbitrary $p \geq 1$. After establishing basic metric and topological properties of $\mathsf{W}_p^{(\sigma)}$, we explore the asymptotic statistical behavior of $\mathsf{W}_p^{(\sigma)}(\hat{\mu}_n, \mu)$, where $\hat{\mu}_n$ is the empirical distribution of n independent observations from μ. We prove that $\mathsf{W}_p^{(\sigma)}$ enjoys a parametric empirical convergence rate of $n^{-1/2}$, which contrasts the $n^{-1/d}$ rate for unsmoothed W_p when $d \geq 3$. Our proof relies on controlling $\mathsf{W}_p^{(\sigma)}$ by a pth-order smooth Sobolev distance $\mathsf{d}_p^{(\sigma)}$ and deriving the limit distribution of $\sqrt{n}\,\mathsf{d}_p^{(\sigma)}(\hat{\mu}_n, \mu)$, for all dimensions d. As applications, we provide asymptotic guarantees for two-sample testing and minimum distance estimation using $\mathsf{W}_p^{(\sigma)}$, with experiments for $p = 2$ using a maximum mean discrepancy formulation of $\mathsf{d}_2^{(\sigma)}$.

1. Introduction

The Wasserstein distance W_p is a discrepancy measure between probability distributions rooted in the theory of optimal transport (Villani, 2003; 2008). It has seen a surge of applications in statistics and ML, ranging from generative modeling (Arjovsky et al., 2017; Gulrajani et al., 2017; Tolstikhin et al., 2018) and image recognition (Rubner et al., 2000; Sandler & Lindenbaum, 2011; Li et al., 2013) to domain adaptation (Courty et al., 2014; 2016) and robust optimization (Mohajerin Esfahani & Kuhn, 2018; Blanchet et al., 2018; Gao & Kleywegt, 2016). The widespread use of this statistical distance is driven by an array of desirable properties, including its metric structure, a convenient dual form, and its robustness to support mismatch.

In applications, the Wasserstein distance is often estimated from samples. However, the error of these empirical estimates suffers from an exponential dependence on dimension that presents an obstacle to sample-efficient bounds for inference and learning. More specifically, the rate at which $\mathsf{W}_p(\hat{\mu}_n, \mu)$ converges to 0, where $\hat{\mu}_n$ is an empirical measure based on n independent samples from μ, scales as $n^{-1/d}$ under mild moment conditions, for $d \geq 3$ (Dereich et al., 2013; Boissard & Le Gouic, 2014; Fournier & Guillin, 2015; Panaretos & Zemel, 2019; Weed & Bach, 2019; Lei, 2020). This rate deteriorates poorly with dimension and seems at odds with the scalability of empirical W_p observed in modern machine learning practice.

1.1. Smooth Wasserstein Distances

Gaussian smoothing was recently introduced as a means to alleviate the curse of dimensionality of empirical W_p, while preserving the virtuous structural properties of the classic framework (Goldfeld et al., 2020b; Goldfeld & Greenewald, 2020; Goldfeld et al., 2020a). Specifically the σ-smooth p-Wasserstein distance is $\mathsf{W}_p^{(\sigma)}(\mu, \nu) := \mathsf{W}_p(\mu * \mathcal{N}_\sigma, \nu * \mathcal{N}_\sigma)$, where $\mathcal{N}_\sigma = \mathcal{N}(0, \sigma^2 \mathrm{I})$ is the isotropic Gaussian measure of parameter σ. Goldfeld & Greenewald (2020) showed that $\mathsf{W}_1^{(\sigma)}$ inherits the metric and topological structure of W_1 and approximates it within an $O(\sigma\sqrt{d})$ gap. At the same time, empirical convergence rates for $\mathsf{W}_p^{(\sigma)}$ are much faster. As shown in (Goldfeld et al., 2020b), $\mathbb{E}\left[\mathsf{W}_1^{(\sigma)}(\hat{\mu}_n, \mu)\right] = O(n^{-1/2})$ when μ is sub-Gaussian in any dimension, i.e., it exhibits a parametric convergence rate. This fast rate was also established for $\mathsf{W}_2^{(\sigma)}$ but only when the sub-Gaussian constant is smaller than $\sigma/2$. These results significantly

[1]Department of Computer Science, Cornell University, Ithaca, NY [2]School of Electrical and Computer Engineering, Cornell University, Ithaca, NY [3]Department of Statistics and Data Science, Cornell University, Ithaca, NY. Correspondence to: Sloan Nietert <sbn45@cornell.edu>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

depart from the $n^{-1/d}$ rate in the unsmoothed case.

Follow-up work (Goldfeld et al., 2020a) developed a limit distribution theory for $\mathsf{W}_1^{(\sigma)}$, showing that $\sqrt{n}\,\mathsf{W}_1^{(\sigma)}(\hat{\mu}_n, \mu)$ converges in distribution to the supremum of a tight Gaussian process, for all dimensions d and under a milder moment condition. Such limit distribution results are known for unsmoothed W_p only when $p \in \{1, 2\}$ and $d = 1$ (del Barrio et al., 1999; 2005) or μ is supported on a finite or countable set (Sommerfeld & Munk, 2018; Tameling et al., 2019), but are a wide open question otherwise. Other works investigated the behavior of $\mathsf{W}_p^{(\sigma)}$ as $\sigma \to \infty$ (Chen & Niles-Weed, 2020), and established related results for other statistical distances, including total variation (TV), Kullback-Leibler (KL) divergence, and χ^2-divergence (Goldfeld et al., 2020b; Chen & Niles-Weed, 2020).

1.2. Contributions

We focus on the smooth p-Wasserstein distance $\mathsf{W}_p^{(\sigma)}$ for $p > 1$ and arbitrary dimension d. We first explore basic structural properties of $\mathsf{W}_p^{(\sigma)}$, proving that many of the beneficial attributes of W_p carry over to the smooth setting. We show that $\mathsf{W}_p^{(\sigma)}$ is a metric and induces the same topology as W_p. Then, we prove that $\mathsf{W}_p^{(\sigma)}$ is stable under small perturbations of the smoothing parameter σ, implying, in particular, that $\mathsf{W}_p^{(\sigma)} \to \mathsf{W}_p$ as $\sigma \to 0$. We then extend the stability of optimal transport distances to that of transport plans, establishing weak convergence of the optimal couplings for $\mathsf{W}_p^{(\sigma)}$ to those of W_p as σ shrinks.

Moving on to a statistical analysis, we explore empirical convergence for $\mathsf{W}_p^{(\sigma)}$. Elementary techniques imply that $\mathbb{E}\left[\mathsf{W}_p^{(\sigma)}(\hat{\mu}_n, \mu)\right] = O(n^{-1/(2p)})$ under a mild moment condition. While this rate is independent of d, it is suboptimal in p, with the expected answer being $n^{-1/2}$ as previously established for $p = 1, 2$ (Goldfeld et al., 2020a;b). To get the correct rate, we establish a comparison between $\mathsf{W}_p^{(\sigma)}$ and a smooth pth-order Sobolev integral probability metric (IPM), $\mathsf{d}_p^{(\sigma)}$; the latter lends itself well to tools from empirical process theory. Under a sub-Gaussian assumption, we prove that the function class defining $\mathsf{d}_p^{(\sigma)}$ is μ-Donsker, giving a limit distribution for $\sqrt{n}\,\mathsf{d}_p^{(\sigma)}(\hat{\mu}_n, \mu)$ that implies the $n^{-1/2}$ rate for $\mathsf{W}_p^{(\sigma)}$. We conclude with a concentration inequality for $\mathsf{W}_p^{(\sigma)}(\hat{\mu}_n, \mu)$.

We next turn to computational aspects, first showing that $\mathsf{d}_2^{(\sigma)}$ is efficiently computable as a maximum mean discrepancy (MMD) and characterizing its reproducing kernel. Next, we consider applications to two-sample testing and generative modeling using $\mathsf{W}_p^{(\sigma)}$. We construct two-sample tests based on the smooth p-Wasserstein test statistic that achieve asymptotic consistency and correct asymptotic level. For generative modeling, we examine minimum distance estimation with $\mathsf{W}_p^{(\sigma)}$ and establish measurability, consistency, and parametric convergence rates, along with finite-sample generalization guarantees in arbitrary dimension. Many of these directions (beyond measurability and consistency) are intractable with standard W_p unless $d = 1$. We conclude with numerical results that support our theory.

1.3. Related Discrepancy Measures

The sliced Wasserstein distance (Rabin et al., 2011) takes an average (or maximum (Deshpande et al., 2019)) of one-dimensional Wasserstein distances over random projections of the d-dimensional distributions. Like the smooth framework considered herein, the sliced distance also exhibits an $n^{-1/2}$ empirical convergence rate and has characterized limit distributions in some cases (Nadjahi et al., 2019; 2020). In (Nadjahi et al., 2019), sliced W_1 was shown to be a metric that induces a topology at least as fine as that of weak convergence, akin to $\mathsf{W}_1^{(\sigma)}$. They further examined generative modeling via minimum sliced Wasserstein estimation, establishing measurability, consistency, and some limit theorems. However, while $\mathsf{W}_p^{(\sigma)}$ converges to W_p as $\sigma \to 0$, there is no approximation parameter for these sliced distances, and comparisons to the standard distances typically require compact support and feature dimension-dependent multiplicative constants (see, e.g., (Bonnotte, 2013)).

Another relevant framework is entropic optimal transport (EOT), which admits efficient algorithms (Cuturi, 2013; Altschuler et al., 2017) and some desirable statistical properties (Genevay et al., 2016; Rigollet & Weed, 2018). In particular, two-sample EOT has empirical convergence rate $n^{-1/2}$ for smooth costs and compactly supported distributions (Genevay et al., 2019). For the quadratic cost, (Mena & Niles-Weed, 2019) extended this rate to sub-Gaussian distributions and derived a central limit theorem (CLT) for empirical EOT, mirroring a result for W_2 established in (del Barrio & Loubes, 2019). The Mena & Niles-Weed (2019) CLT is notably different from ours: it uses as a centering constant the expected empirical distance between $\hat{\mu}_m$ and $\hat{\nu}_n$ as opposed to the population distance between μ and ν (which corresponds to our centering about 0 in the one-sample case). Finally, while EOT can be computed efficiently, it is no longer a metric, even if the underlying cost is (Feydy et al., 2019; Bigot et al., 2019).

Sobolev IPMs have proven independently useful for generative modeling, often referred to as 'dual Sobolev norms'. For example, alternative Sobolev IPMs are the basis for multiple generative adversarial network (GAN) frameworks (Mroueh et al., 2018; Xu et al., 2020) and are featured in (Si et al., 2020), which examines Wasserstein projections of empirical measures onto a chosen hypothesis class.

2. Preliminaries

2.1. Notation

Let $|\cdot|$ and $\langle\cdot,\cdot\rangle$ denote the Euclidean norm and inner product. For a (signed) measure μ and a measurable function f on $\mathbb{R}^d$, we write $\mu(f) = \int f \,\mathrm{d}\mu$. For a non-empty set $\mathcal{T}$, let $\ell^\infty(\mathcal{T})$ be the space of all bounded functions $f : \mathcal{T} \to \mathbb{R}$, equipped with the sup-norm $\|f\|_{\infty,\mathcal{T}} = \sup_{t\in\mathcal{T}} |f(t)|$. The space of compactly supported, infinitely differentiable real functions on $\mathbb{R}^d$ is C_0^∞. For any $p \in [1,\infty)$ and any Borel measure γ on $\mathbb{R}^d$, we denote by $L^p(\gamma;\mathbb{R}^k)$ the space of measurable maps $f : \mathbb{R}^d \to \mathbb{R}^k$ such that $\|f\|_{L^p(\gamma;\mathbb{R}^k)} = (\int_{\mathbb{R}^d} |f|^p \mathrm{d}\gamma)^{1/p} < \infty$. The space $(L^p(\gamma;\mathbb{R}^k), \|\cdot\|_{L^p(\gamma;\mathbb{R}^k)})$ is a Banach space, and we also write $L^p(\gamma) = L^p(\gamma;\mathbb{R}^1)$.

The class of Borel probability measures on $\mathbb{R}^d$ is $\mathcal{P}$ and the subset of measures $\mu \in \mathcal{P}$ with finite p-th moment $\int |x|^p \,\mathrm{d}\mu(x)$ is $\mathcal{P}_p$. The convolution of measures $\mu,\nu \in \mathcal{P}$ is defined by $(\mu * \nu)(A) := \int\int \mathbb{1}_A(x+y)\,\mathrm{d}\mu(x)\,\mathrm{d}\nu(y)$, where $\mathbb{1}_A$ is the indicator of A. The convolution of measurable functions f, g on $\mathbb{R}^d$ is $(f * g)(x) := \int f(x-y)g(y)\,\mathrm{d}y$. Recall that $\mathcal{N}_\sigma = \mathcal{N}(0,\sigma^2\mathrm{I})$ and use $\varphi_\sigma(x) = (2\pi\sigma^2)^{-d/2}e^{-|x|^2/(2\sigma^2)}$, $x \in \mathbb{R}^d$, for the Gaussian density. Write $\mu \otimes \nu$ for the product measure of $\mu,\nu \in \mathcal{P}$. Let $\xrightarrow{w}$ and $\xrightarrow{d}$ denote weak convergence of probability measures and convergence in distribution of random variables.

2.2. Background

We next provide some background on the statistical distances used in this paper.

(Smooth) Wasserstein Distance. For $p \geq 1$, the p-Wasserstein distance between $\mu,\nu \in \mathcal{P}_p$ is defined by

$$\mathsf{W}_p(\mu,\nu) := \left(\inf_{\pi\in\Pi(\mu,\nu)} \int_{\mathbb{R}^d} |x-y|^p \,\mathrm{d}\pi(x,y)\right)^{1/p}$$

where $\Pi(\mu,\nu)$ is the set of couplings of μ and ν. See (Villani, 2003; 2008; Santambrogio, 2015) for additional background. The σ-smooth p-Wasserstein distance between probability measures $\mu,\nu \in \mathcal{P}_p$ is defined by

$$\mathsf{W}_p^{(\sigma)}(\mu,\nu) := \mathsf{W}_p(\mu * \mathcal{N}_\sigma, \nu * \mathcal{N}_\sigma).$$

Integral Probability Metrics. Let $\mathcal{F}$ be a class of measurable real functions on $\mathbb{R}^d$. The IPM with respect to (w.r.t.) $\mathcal{F}$ between probability measures $\mu,\nu \in \mathcal{P}$ is defined by

$$\|\mu-\nu\|_{\infty,\mathcal{F}} = \sup_{f\in\mathcal{F}} \mu(f) - \nu(f).$$

We subsequently control $\mathsf{W}_p^{(\sigma)}$ via an IPM whose functions have bounded Sobolev norm.

Smooth Sobolev IPM. Let γ be a Borel measure on $\mathbb{R}^d$ and fix $p \geq 1$. For a differentiable function $f : \mathbb{R}^d \to \mathbb{R}$, let

$$\|f\|_{\dot{H}^{1,p}(\gamma)} := \|\nabla f\|_{L^p(\gamma;\mathbb{R}^d)} = \left(\int_{\mathbb{R}^d} |\nabla f|^p \mathrm{d}\gamma\right)^{1/p}$$

be its Sobolev seminorm. We define the homogeneous Sobolev space $\dot{H}^{1,p}(\gamma)$ as the completion of $\dot{C}_0^\infty = \{f+a : a \in \mathbb{R}, f \in C_0^\infty\}$ w.r.t. $\|\cdot\|_{\dot{H}^{1,p}(\gamma)}$. The dual Sobolev norm of a signed measure ℓ on $\mathbb{R}^d$ with zero total mass is

$$\|\ell\|_{\dot{H}^{-1,p}(\gamma)} := \sup\{\ell(f) : f \in C_0^\infty, \|f\|_{\dot{H}^{1,q}(\gamma)} \leq 1\},$$

where q is the conjugate index of p, i.e., $1/p + 1/q = 1$. We define the pth-order smooth Sobolev IPM by

$$\mathsf{d}_p^{(\sigma)}(\mu,\nu) := \|(\mu-\nu) * \mathcal{N}_\sigma\|_{\dot{H}^{-1,p}(\mathcal{N}_\sigma)}$$

for measures $\mu,\nu \in \mathcal{P}$. Observe that $\mathsf{d}_p^{(\sigma)}$ is an IPM w.r.t. the class $\mathcal{F} * \varphi_\sigma = \{f * \varphi_\sigma : f \in \mathcal{F}\}$ with $\mathcal{F} = \{f \in C_0^\infty : \|f\|_{\dot{H}^{1,q}(\mathcal{N}_\sigma)} \leq 1\}$.

3. Structure of Smooth Wasserstein Distance and Comparison with Smooth Sobolev IPM

We now examine basic properties of smooth Wasserstein distances, including a useful connection to the smooth Sobolev IPM. The case of $\mathsf{W}_1^{(\sigma)}$ has been well-studied in (Goldfeld & Greenewald, 2020; Goldfeld et al., 2020a). Herein we present results that hold for arbitrary $p \geq 1$ and $\sigma \geq 0$ unless stated otherwise, with proofs left for the supplement. Extending beyond $p = 1$ requires new techniques, most prominently a comparison result between $\mathsf{W}_p^{(\sigma)}$ and $\mathsf{d}_p^{(\sigma)}$.

3.1. Structural Properties

We first consider the topology induced by $\mathsf{W}_p^{(\sigma)}$. Since convolution acts as a contraction, we have $\mathsf{W}_p^{(\sigma)} \leq \mathsf{W}_p$. In fact, the two distances induce the same topology on $\mathcal{P}_p$, which coincides with that of weak convergence in addition to convergence of pth moments.

Proposition 1 (Metric and topological structure of $\mathsf{W}_p^{(\sigma)}$)**.** *$\mathsf{W}_p^{(\sigma)}$ is a metric on $\mathcal{P}_p$ inducing the same topology as W_p.*

The proof uses existence of optimal couplings and uniform integrability arguments. Next, we examine the behavior of $\mathsf{W}_p^{(\sigma)}$ as a function of the smoothing parameter σ. We start from the following stability lemma, guaranteeing that small changes in σ result only in slight perturbations of $\mathsf{W}_p^{(\sigma)}$.

Lemma 1 (Stability of $\mathsf{W}_p^{(\sigma)}$)**.** *For $\mu,\nu \in \mathcal{P}_p$ and $0 \leq \sigma_1 \leq \sigma_2 < \infty$, we have*

$$\left|\mathsf{W}_p^{(\sigma_2)}(\mu,\nu) - \mathsf{W}_p^{(\sigma_1)}(\mu,\nu)\right| \leq 2\sqrt{(\sigma_2^2-\sigma_1^2)(d+2p+2)}.$$

This result generalizes Lemma 1 of (Goldfeld & Greenewald, 2020), which covers $p = 1$, and establishes uniform continuity of $\mathsf{W}_p^{(\sigma)}$ in σ. Its proof takes a different approach, using Minkowski's inequality instead of the Kantorovich-Rubinstein duality. An immediate consequence of Lemma 1 is given next, mirroring Theorem 3 of (Goldfeld & Greenewald, 2020).

Corollary 1 ($\mathsf{W}_p^{(\sigma)}$ dependence on σ)**.** *For $\mu, \nu \in \mathcal{P}_p$, the following hold:*

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

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

(iii) $\lim_{\sigma \to \infty} \mathsf{W}_p^{(\sigma)}(\mu, \nu) = \left| \mathbb{E}[X] - \mathbb{E}[Y] \right|$, for $X \sim \mu$ and $Y \sim \nu$ sub-Gaussian.

Remark 1 (Infinite smoothing)**.** A detailed study of $\mathsf{W}_p^{(\sigma)}$ in the infinite smoothing regime (i.e., when $\sigma \to \infty$) is conducted in (Chen & Niles-Weed, 2020). Therein, the authors prove Item 3 above and examine the convergence of $\mathsf{W}_p^{(\sigma)}(\mu, \nu)$ to 0 when $\mathbb{E}[X] = \mathbb{E}[Y]$. For that case, they show that if μ and ν have matching moment tensors up to order n (but not $n+1$), then $\mathsf{W}_2^{(\sigma)}(\mu, \nu) \asymp \sigma^{-n}$ as $\sigma \to \infty$.

Corollary 1 guarantees the convergence of transport costs as $\sigma \to 0$. It is natural to ask whether optimal transport plans (i.e., couplings) that achieve these costs converge as well. We answer this question to the affirmative.

Proposition 2 (Convergence of transport plans)**.** *Fix $\mu, \nu \in \mathcal{P}_p$ and let $(\sigma_k)_{k \in \mathbb{N}}$ be a sequence with $\sigma_k \searrow \sigma \geq 0$. For each $k \in \mathbb{N}$, let $\pi_k \in \Pi(\mu * \mathcal{N}_{\sigma_k}, \nu * \mathcal{N}_{\sigma_k})$ be an optimal coupling for $\mathsf{W}_p^{(\sigma_k)}(\mu, \nu)$. Then there exists $\pi \in \Pi(\mu * \mathcal{N}_\sigma, \nu * \mathcal{N}_\sigma)$ such that $\pi_k \xrightarrow{w} \pi$ as $k \to \infty$, and π is optimal for $\mathsf{W}_p^{(\sigma)}(\mu, \nu)$.*

The proof observes that the arguments for Theorem 4 of (Goldfeld & Greenewald, 2020) extend from $p = 1$ to the general case with minor changes.

So far we have studied metric, topological, and limiting properties of $\mathsf{W}_p^{(\sigma)}$. In Section 4 we explore its statistical behavior, when distributions are estimated from samples. To that end, we now establish a relation between $\mathsf{W}_p^{(\sigma)}$ and the smooth Sobolev IPM $\mathsf{d}_p^{(\sigma)}$. This result is later used to study the empirical convergence under $\mathsf{W}_p^{(\sigma)}$ using tools from empirical process theory (applied to the $\mathsf{d}_p^{(\sigma)}$ upper bound).

Theorem 1 (Comparison between $\mathsf{W}_p^{(\sigma)}$ and $\mathsf{d}_p^{(\sigma)}$)**.** *Fix $p > 1$ and let q be the conjugate index of p. Then, for $X \sim \mu \in \mathcal{P}_p$ with mean 0 and $\nu \in \mathcal{P}$, we have*

$$\mathsf{W}_p^{(\sigma)}(\mu, \nu) \leq p\, e^{\mathbb{E}[|X|^2]/(2q\sigma^2)}\, \mathsf{d}_p^{(\sigma)}(\mu, \nu). \tag{1}$$

The proof builds upon related inequalities established for standard W_p (Dolbeault et al., 2009; Peyre, 2018; Ledoux, 2019), exploiting the metric structure of the Wasserstein space and the Benamou-Brenier dynamic formulation of optimal transport (Benamou & Brenier, 2000). Namely, we note that $\mathsf{W}_p(\mu_0, \mu_1)$ is upper bounded by the length of any continuous path from μ_0 to μ_1 in $(\mathcal{P}_p, \mathsf{W}_p)$ and examine the path $t \mapsto t\mu_1 + (1-t)\mu_0$ which interpolates linearly between the two densities. The theorem follows upon applying the resulting bound to $\mu * \mathcal{N}_\sigma$ and $\nu * \mathcal{N}_\sigma$. We also give a lower bound for $\mathsf{W}_p^{(\sigma)}(\mu, \nu)$ using $\|(\mu - \nu) * \mathcal{N}_\sigma\|_{\dot{H}^{-1,p}(\mathcal{N}_{\sqrt{2}\sigma})}$, though the constant factor restricts its usefulness.

Remark 2. When $p = 1$, one can show that W_1 and the dual Sobolev norm $\|\cdot\|_{\dot{H}^{-1,1}(\gamma)}$ coincide (Dolbeault et al., 2009). In particular, this implies that $\mathsf{W}_1^{(\sigma)}(\mu, \nu) = \mathsf{d}_1^{(\sigma)}(\mu, \nu)$. For larger p, the gap between $\mathsf{W}_p^{(\sigma)}(\mu, \nu)$ and the upper bound given by Theorem 1 can grow quite large, so we view the comparison as a useful theoretical tool rather than a device for practical approximation guarantees.

Finally, we establish some basic properties of $\mathsf{d}_p^{(\sigma)}$.

Proposition 3 ($\mathsf{d}_p^{(\sigma)}$ dependence on σ)**.** *For $\mu, \nu \in \mathcal{P}$, the following hold:*

(i) $\lim_{\sigma \to 0} \mathsf{d}_p^{(\sigma)}(\mu, \nu) = \infty$ for $\mu \neq \nu$;

(ii) $\lim_{\sigma \to \infty} \mathsf{d}_2^{(\sigma)}(\mu, \nu) = |\mathbb{E}[X] - \mathbb{E}[Y]|$, for $X \sim \mu$ and $Y \sim \nu$ sub-Gaussian.

We focus on $p = 2$ for (ii) due to a convenient MMD formulation for $\mathsf{d}_2^{(\sigma)}$ established in Section 5.

4. Empirical Approximations

Fix $p > 1$, $\sigma > 0$, and let $\mu \in \mathcal{P}_p$ with $X \sim \mu$. Given independently and identically distributed (i.i.d.) samples $X_1, \ldots, X_n \sim \mu$ with empirical distribution $\hat{\mu}_n := n^{-1} \sum_{i=1}^n \delta_{X_i}$, we study the convergence rate of $\mathbb{E}\left[\mathsf{W}_p^{(\sigma)}(\hat{\mu}_n, \mu)\right]$ to zero. To start, we observe that elementary techniques imply $\mathbb{E}\left[\mathsf{W}_p^{(\sigma)}(\hat{\mu}_n, \mu)\right] = O(n^{-1/(2p)})$ under mild conditions on μ. Although the rate $n^{-1/(2p)}$ is dimension-free, its dependence on p is sub-optimal.

Theorem 2 (Slow rate)**.** *If $X \sim \mu$ satisfies*

$$\int_0^\infty r^{d+p-1} \sqrt{\mathbb{P}(|X| > r)}\, \mathrm{d}r < \infty, \tag{2}$$

then $\mathbb{E}\left[\mathsf{W}_p^{(\sigma)}(\hat{\mu}_n, \mu)\right] = O(n^{-1/(2p)})$. Condition (2) holds if μ has finite $(2d + 2p + \epsilon)$-th moment for some $\epsilon > 0$.

The proof follows by coupling μ and $\hat{\mu}_n$ via the maximal coupling. This bounds $\left(\mathsf{W}_p^{(\sigma)}(\hat{\mu}_n, \mu)\right)^p$ from above by a

weighted TV distance, which converges as $n^{-1/2}$, provided that the above moment condition holds. This proof technique was previously applied in (Goldfeld et al., 2020b) to achieve the same rate when $p = 1$.

We next turn to show that the $n^{-1/2}$ rate is attainable for $\mathsf{W}_p^{(\sigma)}(\hat{\mu}_n, \mu)$ itself (rather than for its pth power). To this end, we first establish a limit distribution result for the empirical smooth Sobolev IPM $\mathsf{d}_p^{(\sigma)}(\hat{\mu}_n, \mu)$. This, in turn, yields the desired rate for $\mathbb{E}\left[\mathsf{W}_p^{(\sigma)}(\hat{\mu}_n, \mu)\right]$ via the comparison from Theorem 1. Recall the function class $\mathcal{F} = \{f \in C_0^\infty : \|f\|_{\dot{H}^{1,q}(\mathcal{N}_\sigma)} \leq 1\}$.

Theorem 3 (Limit distribution for empirical $\mathsf{d}_p^{(\sigma)}$)**.** *Suppose there exists $\theta > p - 1$ for which $X \sim \mu$ satisfies*

$$\int_0^\infty e^{\frac{\theta r^2}{2\sigma^2}} \sqrt{\mathbb{P}(|X| > r)} \mathrm{d}r < \infty. \tag{3}$$

Then $\sqrt{n}\mathsf{d}_p^{(\sigma)}(\hat{\mu}_n, \mu) \xrightarrow{d} \|G\|_{\infty,\mathcal{F}}$ as $n \to \infty$, where $G = (G(f))_{f\in\mathcal{F}}$ is a tight Gaussian process in $\ell^\infty(\mathcal{F})$ with mean zero and covariance function $\mathrm{Cov}(G(f), G(g)) = \mathrm{Cov}(f * \varphi_\sigma(X), g * \varphi_\sigma(X))$.

Corollary 2 (Fast rate)**.** *Under the conditions of Theorem 3, we have $\lim_{n\to\infty} \sqrt{n}\,\mathbb{E}\left[\mathsf{d}_p^{(\sigma)}(\hat{\mu}_n, \mu)\right] = \mathbb{E}\left[\|G\|_{\infty,\mathcal{F}}\right] < \infty$. Consequently, $\mathbb{E}\left[\mathsf{W}_p^{(\sigma)}(\hat{\mu}_n, \mu)\right] = O(n^{-1/2})$.*

The proof of Theorem 3 shows that the smoothed function class $\mathcal{F} * \varphi_\sigma = \{f * \varphi_\sigma : f \in \mathcal{F}\}$ is μ-Donsker. Specifically, we prove that functions in $\mathcal{F} * \varphi_\sigma$ are smooth with derivatives uniformly bounded on domains within a fixed radius of the origin. Using these bounds, we apply techniques from empirical process theory to establish the Donsker property. Importantly, the preceding argument hinges on the convolution with the smooth Gaussian density and does *not* hold for the unsmoothed function class. No mean zero requirement appears because we can center $\hat{\mu}_n$ and μ by the mean of μ.

Condition (3) requires that $\mathbb{P}(|X| > r) \to 0$ faster than e^{-Cr^2} as $r \to \infty$ for some $C > (p-1)/\sigma^2$, which in turn requires $|X|$ to be sub-Gaussian. The requirement is trivially satisfied if μ is compactly supported. We can also relate Condition (3) to a more standard notion of sub-Gaussianity for random vectors.

Definition 1 (Sub-Gaussian distribution)**.** Let $Y \sim \nu \in \mathcal{P}$ with $\mathbb{E}[|Y|] < \infty$. We say that ν or Y is *β-sub-Gaussian* for $\beta \geq 0$ if $\mathbb{E}[\exp(\langle \alpha, Y - \mathbb{E}[Y]\rangle)] \leq \exp\left(\beta|\alpha|^2/2\right)$ for all $\alpha \in \mathbb{R}^d$.

Proposition 4 (Sub-Gaussianity implies (3))**.** *If μ is β-sub-Gaussian with $\beta < \sigma/\sqrt{2(p-1)}$, then (3) holds.*

Next, we consider the concentration of $\mathsf{W}_p^{(\sigma)}(\hat{\mu}_n, \mu)$.

Proposition 5 (Concentration inequality)**.** *If μ has compact support, then, for all $t > 0$ and $n \in \mathbb{N}$, we have*

$$\mathbb{P}\left(\mathsf{W}_p^{(\sigma)}(\hat{\mu}_n, \mu) \geq Cn^{-1/2} + t\right) \leq \exp\left(-cnt^2\right)$$

with constants C, c independent of n and t.

For unbounded domains, concentration results mirroring those of Corollary 3 in (Goldfeld et al., 2020a) can be established in the same way under a stronger sub-Gaussianity assumption; we omit the details for brevity.

Remark 3 (Constants)**.** While the rates provided in this section are dimension-free, the constants necessarily exhibit an exponential dependence on dimension. Indeed, minimax results for estimation of standard W_p due to Singh & Póczos (2018), combined with Lemma 1, imply that achieving dimension-free rates with constants scaling only polynomially in dimension is impossible in general. We provide further details in the proofs of Theorem 2 and Theorem 3.

5. Smooth Sobolev IPM Efficient Computation

We next consider computation of $\mathsf{d}_p^{(\sigma)}$, which takes a convenient form when $p = 2$. Specifically, the Hilbertian structure of $\dot{H}^{1,2}(\mathcal{N}_\sigma)$ enables to streamline calculations significantly for all dimensions d. In the following, fix $\sigma > 0$, $\mu, \nu \in \mathcal{P}$, $X, X' \sim \mu \otimes \mu$, and $Y, Y' \sim \nu \otimes \nu$.

Consider the function space $\dot{H}_0^{1,2}(\mathcal{N}_\sigma) := \{f \in \dot{H}^{1,2}(\mathcal{N}_\sigma) : \mathcal{N}_\sigma(f) = 0\}$ with norm $\|\cdot\|_{\dot{H}^{1,2}(\mathcal{N}_\sigma)}$ (this norm is proper because of the constraint $\mathcal{N}_\sigma(f) = 0$). This space becomes a Hilbert space when equipped with inner product $\langle f, g\rangle_{\dot{H}^{1,2}(\mathcal{N}_\sigma)} = \int_{\mathbb{R}^d} \langle \nabla f, \nabla g\rangle \,\mathrm{d}\mathcal{N}_\sigma$. Likewise, the space $\dot{H}_0^{1,2}(\mathcal{N}_\sigma) * \varphi_\sigma = \{f * \varphi_\sigma : f \in \dot{H}_0^{1,2}(\mathcal{N}_\sigma)\}$ is a Hilbert space with inner product $\langle f * \varphi_\sigma, g * \varphi_\sigma\rangle_{\dot{H}^{1,2}(\mathcal{N}_\sigma)*\varphi_\sigma} = \langle f, g\rangle_{\dot{H}^{1,2}(\mathcal{N}_\sigma)}$ (see Appendix A.3 for a proof that this is well-defined). In fact, we can say a bit more, first recalling some definitions.

Reproducing Kernel Hilbert Space (RKHS). Let $\mathcal{H}$ be a Hilbert space of real-valued functions on $\mathbb{R}^d$. We say that $\mathcal{H}$ is an RKHS if there is a positive semidefinite function $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, called a reproducing kernel, such that $k(\cdot, x) \in \mathcal{H}$ and $h(x) = \langle h, k(\cdot, x)\rangle$ for all $x \in \mathbb{R}^d$ and $h \in \mathcal{H}$. See (Steinwart & Christmann, 2008; Aronszajn, 1950) for comprehensive background on RKHSs.

Maximum Mean Discrepancy. Let $\mathcal{H}$ be an RKHS with kernel k. The IPM corresponding to its unit ball, termed MMD, is given by

$$\mathsf{MMD}_{\mathcal{H}}(\mu, \nu) := \sup_{f\in\mathcal{H}:\, \|f\|_{\mathcal{H}} \leq 1} \mu(f) - \nu(f).$$

Proposition 6 (Borgwardt et al. (2006))**.** *If $\mathbb{E}\left[\sqrt{k(X,X)}\right]$, $\mathbb{E}\left[\sqrt{k(Y,Y)}\right] < \infty$, then*

$$\mathsf{MMD}_{\mathcal{H}}(\mu,\nu)^2 = \mathbb{E}[k(X,X')] - 2\mathbb{E}[k(X,Y)] + \mathbb{E}[k(Y,Y')]. \tag{4}$$

We prove that $\dot{H}_0^{1,2}(\mathcal{N}_\sigma) * \varphi_\sigma$ is an RKHS whose kernel is expressed in terms of the entire exponential integral (Oldham et al., 2009) $\mathrm{Ein}(z) := \int_0^z (1-e^{-t})\frac{dt}{t} = \sum_{k=1}^\infty \frac{(-1)^{k+1}z^k}{k\cdot k!}$, giving an MMD form for $\mathsf{d}_2^{(\sigma)}$.

Theorem 4 ($\mathsf{d}_2^{(\sigma)}$ as an MMD)**.** *The space $\dot{H}_0^{1,2}(\mathcal{N}_\sigma) * \varphi_\sigma$ is an RKHS with reproducing kernel $\kappa^{(\sigma)}(x,y) := -\sigma^2 \mathrm{Ein}\left(-\langle x,y\rangle/\sigma^2\right)$. Thus, if $\mathbb{E}\left[\sqrt{\kappa^{(\sigma)}(X,X)}\right], \mathbb{E}\left[\sqrt{\kappa^{(\sigma)}(Y,Y)}\right] < \infty$, then*

$$\mathsf{d}_2^{(\sigma)}(\mu,\nu)^2 = \mathbb{E}\left[\kappa^{(\sigma)}(X,X')\right] + \mathbb{E}\left[\kappa^{(\sigma)}(Y,Y')\right] - 2\,\mathbb{E}\left[\kappa^{(\sigma)}(X,Y)\right]. \tag{5}$$

The proof begins with a reduction to $\sigma = 1$ and observes that properly normalized multivariate Hermite polynomials form an orthonormal basis for $\dot{H}_0^{1,2}(\mathcal{N}_1)$. Convolving these polynomials with φ_1, we obtain an orthonormal basis for $\dot{H}_0^{1,2}(\mathcal{N}_1) * \varphi_1$ comprising scaled monomials, which can then be used to calculate the kernel.

The MMD formulation (5) gives a convenient way to compute $\mathsf{d}_2^{(\sigma)}$ in practice. Suppose that we generate i.i.d. samples $X_1,\dots,X_m \sim \mu$ and $Y_1,\dots,Y_n \sim \nu$ with empirical distributions $\hat{\mu}_m = m^{-1}\sum_{i=1}^m \delta_{X_i}$ and $\hat{\nu}_n = n^{-1}\sum_{j=1}^n \delta_{Y_j}$. Then, we can compute

$$\mathsf{d}_2^{(\sigma)}(\hat{\mu}_m,\hat{\nu}_n)^2 = \frac{1}{m^2}\sum_{i=1}^m\sum_{j=1}^m \kappa^{(\sigma)}(x_i,x_j) + \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n \kappa^{(\sigma)}(y_i,y_j) - \frac{2}{mn}\sum_{i=1}^m\sum_{j=1}^n \kappa^{(\sigma)}(x_i,y_j).$$

Provided that μ and ν are compactly supported or β-sub-Gaussian with $\beta < \sigma/\sqrt{2}$, Corollary 2 and the triangle inequality imply

$$\mathbb{E}\left[\left|\mathsf{d}_2^{(\sigma)}(\hat{\mu}_m,\hat{\nu}_n) - \mathsf{d}_2^{(\sigma)}(\mu,\nu)\right|\right] = O\left(\min\{m,n\}^{-1/2}\right).$$

Hence, we can approximate $\mathsf{d}_2^{(\sigma)}$ up to expected error ε with $O(\varepsilon^{-2})$ samples from the measured distributions and $O(\varepsilon^{-4})$ evaluations of $\kappa^{(\sigma)}$ for any dimension d.

6. Statistical Applications

With the empirical approximation and computational results in hand, we now present applications to two-sample testing and minimum distance estimation. These results highlight the benefits of smoothing, as several of the subsequent claims are unavailable for standard W_p due to the lack of parametric rates and limit distributions.

6.1. Two-Sample Testing

We start from two-sample testing with $\mathsf{W}_p^{(\sigma)}$ and $\mathsf{d}_p^{(\sigma)}$, where $p > 1$ and $\sigma > 0$ are fixed throughout. Let $\mu,\nu \in \mathcal{P}_p$ and take $X_1,\dots,X_m \sim \mu$ and $Y_1,\dots,Y_n \sim \nu$ to be mutually independent samples. The goal of nonparametric two-sample testing is to detect, based on the samples, whether the null hypothesis $H_0 : \mu = \nu$ holds, without imposing parametric assumptions on the distributions.

A standard class of tests rejects H_0 if $D_{m,n} > c_{m,n}$, where $D_{m,n} = D_{m,n}(X_1,\dots,X_m,Y_1,\dots,Y_n)$ is a scalar test statistic and $c_{m,n}$ is a critical value chosen according to the desired level $\alpha \in (0,1)$. Precisely, we say that such a sequence of tests has asymptotic *level* α if $\limsup_{m,n\to\infty} \mathbb{P}(D_{m,n} > c_{m,n}) \le \alpha$ whenever $\mu = \nu$. We say that these tests are asymptotically *consistent* if $\lim_{m,n\to\infty} \mathbb{P}(D_{m,n} > c_{m,n}) = 1$ whenever $\mu \neq \nu$. In what follows, we assume that $m,n \to \infty$ and $m/N \to \tau \in (0,1)$ with $N = m+n$.

The previous theorems will help us construct tests that enjoy asymptotic consistency and correct asymptotic level based on the smooth p-Wasserstein distance, using $W_{m,n} := \sqrt{\frac{mn}{N}}\,\mathsf{W}_p^{(\sigma)}(\hat{\mu}_m,\hat{\nu}_n)$. Two-sample testing using the Wasserstein distance was previously explored in (Ramdas et al., 2017), but these results are fundamentally restricted to the one-dimensional setting. Specifically, while the authors designed tests with data-independent critical values for $d = 1$, they rely heavily on limit distributions of empirical Wasserstein distances that do not extend to higher dimensions. Our results use data-dependent critical values but scale to arbitrary dimension, demonstrating the compatibility of smooth distances for multivariate two-sample testing.

We use the bootstrap to calibrate critical values. Consider the pooled data $(Z_1,\dots,Z_N) = (X_1,\dots,X_m,Y_1,\dots,Y_n)$ with empirical distribution $\hat{\gamma}_N = N^{-1}\sum_{i=1}^N \delta_{Z_i}$. Let $X_1^B,\dots,X_m^B$ and $Y_1^B,\dots,Y_n^B$ be i.i.d. from $\hat{\gamma}_N$ given $Z_1,\dots,Z_N$, and take $\hat{\mu}_m^B = m^{-1}\sum_{i=1}^m \delta_{X_i^B}$ and $\hat{\nu}_n^B = n^{-1}\sum_{i=1}^n \delta_{Y_i^B}$ to be the corresponding bootstrap empirical measures.

Specifying critical values requires a bit of care, as the comparison inequality (1) requires centering of one of measures. So we center the bootstrap empirical measures $\hat{\mu}_m^B$ and $\hat{\nu}_n^B$ by the pooled sample mean $\bar{Z} = N^{-1}\sum_{i=1}^n Z_i$, namely, we apply the bootstrap as

$$W_{m,n}^B = p\, e^{\frac{\mathrm{tr}\,\hat{\Sigma}_Z}{2q\sigma^2}} \sqrt{\frac{mn}{N}}\,\mathsf{d}_p^{(\sigma)}\left(\hat{\mu}_m^B * \delta_{-\bar{Z}_N}, \hat{\nu}_n^B * \delta_{-\bar{Z}_N}\right),$$

where $\hat{\Sigma}_Z = N^{-1}\sum_{i=1}^N (Z_i - \bar{Z}_N)(Z_i - \bar{Z}_N)^\top$. Denote the conditional $(1-\alpha)$-quantile of $W_{m,n}^B$ by $w_{m,n}^B(1-\alpha)$, i.e.,

$$w_{m,n}^B(1-\alpha) = \inf\left\{t : \mathbb{P}^B(W_{m,n}^B \le t) \ge 1-\alpha\right\}.$$

Then, we have the following result.

Proposition 7 (Asymptotic validity)**.** *For $\mu,\nu \in \mathcal{P}_p$ satisfying the condition of Theorem 3, the sequence of tests that re-*

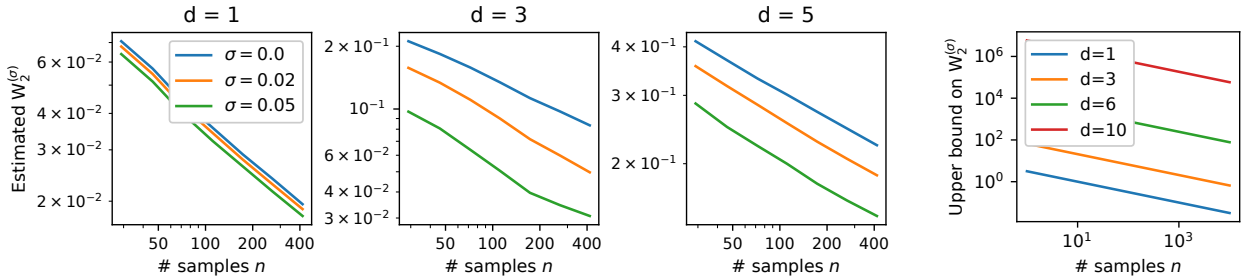

Figure 1. (Left) Empirical convergence of estimated $\mathbb{E}[\mathsf{W}_2^{(\sigma)}(\hat{\mu}_n, \mu)]$ to 0 for $\mu = \mathsf{Unif}([-1,1]^d)$ with $d \in \{1,3,5\}$. (Right) Loose upper bound on $\mathbb{E}[\mathsf{W}_2^{(0.5)}(\hat{\mu}_n, \mu)]$ provided by $\mathsf{d}_2^{(0.5)}$ for $\mu = \mathsf{Unif}([-1,1]^d)$ with $d \in \{1,3,6,10\}$.

ject the null hypothesis $H_0 : \mu = \nu$ if $W_{m,n} > w^B_{m,n}(1-\alpha)$ is asymptotically consistent with level α.

We remark that the same argument gives a simpler result when $p = 1$. Because $\mathsf{W}_1^{(\sigma)}$ is an IPM itself w.r.t. a function class that is μ- and ν-Donsker under moment conditions (see (Goldfeld et al., 2020a, Theorem 1)), no centering is necessary, and $W^B_{m,n}$ can be replaced by $\sqrt{\frac{mn}{N}}\mathsf{W}_1^{(\sigma)}(\hat{\mu}^B_m, \hat{\nu}^B_n)$.

6.2. Generative Modeling

In the unsupervised learning task of generative modeling, we obtain an i.i.d. sample $X_1, \ldots, X_n$ from a distribution $\mu \in \mathcal{P}$ and aim to learn a generative model from a parameterized family $\{\nu_\theta\}_{\theta\in\Theta} \subset \mathcal{P}$ which approximates μ under some statistical distance. We adopt the smooth Wasserstein distance as the figure of merit and use the empirical distribution $\hat{\mu}_n$ as an estimate for μ. Generative modeling is thus formulated as the following minimum smooth Wasserstein estimation (M-SWE) problem:

$$\inf_{\theta\in\Theta} \mathsf{W}_p^{(\sigma)}(\hat{\mu}_n, \nu_\theta).$$

In the unsmooth case, this objective with W_1 inspired the Wasserstein GAN (W-GAN) framework that continues to underlie state-of-the-art methods in generative modeling (Arjovsky et al., 2017; Gulrajani et al., 2017). M-SWE with $p = 1$ was studied in (Goldfeld et al., 2020a), and here we pursue similar measurability and consistency results for $p > 1$.

In what follows, we take both $\mu \in \mathcal{P}_p$ and $\{\nu_\theta\}_{\theta\in\Theta} \subset \mathcal{P}_p$. Further, we suppose that $\Theta \subset \mathbb{R}^{d_0}$ is compact with nonempty interior and that $\theta \mapsto \nu_\theta$ is continuous w.r.t. the weak topology, i.e., $\nu_\theta \xrightarrow{w} \nu_{\bar{\theta}}$ whenever $\theta \to \bar{\theta}$. We start by establishing measurability, consistency, and parametric convergence rates for M-SWE.

Proposition 8 (M-SWE measurability)**.** *For each $n \in \mathbb{N}$, there exists a measurable function $\omega \mapsto \hat{\theta}_n(\omega)$ such that $\hat{\theta}_n(\omega) \in \operatorname{argmin}_{\theta\in\Theta} \mathsf{W}_p^{(\sigma)}(\hat{\mu}_n(\omega), \nu_\theta)$.*

Proposition 9 (M-SWE consistency)**.** *The following hold:*

1. $\inf_{\theta\in\Theta} \mathsf{W}_p^{(\sigma)}(\hat{\mu}_n, \nu_\theta) \to \inf_{\theta\in\Theta} \mathsf{W}_p^{(\sigma)}(\mu, \nu_\theta)$ a.s.

2. There exists an event with probability one on which the following holds: for any sequence $\{\hat{\theta}_n\}_{n\in\mathbb{N}}$ of measurable estimators such that $\mathsf{W}_p^{(\sigma)}(\hat{\mu}_n, \nu_{\hat{\theta}_n}) \le \inf_{\theta\in\Theta} \mathsf{W}_p^{(\sigma)}(\hat{\mu}_n, \nu_\theta) + o_{\mathbb{P}}(1)$, the set of cluster points of $\{\hat{\theta}_n\}_{n\in\mathbb{N}}$ is included in $\operatorname{argmin}_{\theta\in\Theta} \mathsf{W}_p^{(\sigma)}(\mu, \nu_\theta)$.

3. If $\operatorname{argmin}_{\theta\in\Theta} \mathsf{W}_p^{(\sigma)}(\mu, \nu_\theta) = \{\theta^\star\}$, then $\hat{\theta}_n \to \theta^\star$ a.s.

Proposition 10 (M-SWE convergence rate)**.** *If μ satisfies the conditions of Theorem 3, then*

$$\left| \inf_{\theta\in\Theta} \mathsf{W}_p^{(\sigma)}(\hat{\mu}_n, \nu_\theta) - \inf_{\theta\in\Theta} \mathsf{W}_p^{(\sigma)}(\mu, \nu_\theta) \right| = O_{\mathbb{P}}(n^{-1/2}).$$

Likewise, under additional regularity conditions, the solutions $\hat{\theta}_n$ to M-SWE converge at the parametric rate, i.e., $|\hat{\theta}_n - \theta^\star| = O_{\mathbb{P}}(n^{-1/2})$, where $\theta^\star$ is the unique solution as above; see Supplement A.4 for details.

These propositions follow by similar arguments to those in (Goldfeld et al., 2020a), which build on (Pollard, 1980), with arbitrary $p \ge 1$ instead of $p = 1$ as considered therein (the needed results from (Villani, 2008) hold for all $p \ge 1$). We thus omit their proofs for brevity.

We next examine a high probability generalization bound for generative modeling via M-SWE, in accordance to the framework from (Arora et al., 2017; Zhang et al., 2018). Thus, we want to control the gap between the $\mathsf{W}_p^{(\sigma)}$ loss attained by approximate, possibly suboptimal, empirical minimizers and the population loss $\inf_{\theta\in\Theta} \mathsf{W}_p^{(\sigma)}(\mu, \nu_\theta)$. Upper bounding this gap by the rate of empirical convergence, the concentration result Proposition 5 implies the following.

Corollary 3 (M-SWE generalization error)**.** *Assume μ has compact support and let $\hat{\theta}_n$ be an estimator with $\mathsf{W}_p^{(\sigma)}(\hat{\mu}_n, \nu_{\hat{\theta}_n}) \le \inf_{\theta\in\Theta} \mathsf{W}_p^{(\sigma)}(\hat{\mu}_n, \nu_\theta) + \epsilon$, for some $\epsilon > 0$. We have*

$$\mathbb{P}\left(\mathsf{W}_p^{(\sigma)}(\mu, \nu_{\hat{\theta}_n}) - \inf_{\theta\in\Theta} \mathsf{W}_p^{(\sigma)}(\mu, \nu_\theta) > \epsilon + t \right) \le Ce^{-cnt^2},$$

for constants C, c independent of n and t.

7. Experiments

We present several numerical experiments supporting the theoretical results established in the previous sections. We focus on $p = 2$ so that we can use the MMD form for $\mathsf{d}_2^{(\sigma)}$. Code is provided at https://github.com/sbnietert/smooth-Wp.

First, we examine $\mathsf{W}_2^{(\sigma)}(\mu, \hat{\mu}_n)$ directly, with computations feasible for small sample sizes using the stochastic averaged gradient (SAG) method as proposed by (Genevay et al., 2016) and implemented by (Hallin et al., 2020). In Figure 1 (left), we take $\mu = \mathsf{Unif}([-1,1]^d)$ and estimate $\mathbb{E}[\mathsf{W}_2^{(\sigma)}(\hat{\mu}_n, \mu)]$ averaged over 10 trials, for varied d and σ. We observe the contractive property of $\mathsf{W}_2^{(\sigma)}$ and the speed up in convergence rate due to smoothing. However, SAG is not well-suited for computation in high dimensions and larger values of σ. Indeed, this method computes standard W_2 between the convolved measures, which needs an exponential in d number of samples from the Gaussian measure. Recently, Vacher et al. (2021) suggested that OT distances between smooth densities (like those of our convolved measures) may be computed more efficiently, but their algorithm is restricted to compactly supported distributions and leaves important hyperparameters unspecified.

Turning to $\mathsf{d}_2^{(\sigma)}$, the MMD form from (5) readily enables efficient computation. In Figure 1 (right), we plot the $n^{-1/2}$ upper bound it gives for $\mathsf{W}_2^{(\sigma)}$ empirical convergence, using a closed form for relevant expectations described in Appendix A.5.1. We emphasize that this bound is too loose for practical approximation and serves rather as a theoretical tool for obtaining correct rates. In Figure 2, we plot distributions of $\sqrt{n}\,\mathsf{d}_2^{(1)}(\hat{\mu}_n, \mu)$ as n increases, using $\mu = \mathcal{N}_s$ with varied s and $d = 5$. Distributions are computed using kernel density estimation over 50 trials and estimating μ by $\hat{\mu}_{1000}$. We see convergence to a limit distribution for small σ (estimating $\|G\|_{\infty,\mathcal{F}}$ from Theorem 3) and note the necessity of the sub-Gaussian condition (3), with convergence failing as σ surpasses $1/\sqrt{2}$, supporting Proposition 4.

Next, we examine M-SWE when $d = 1$, exploiting the fact that W_p can be expressed as an L^p distance between quantile functions (see, e.g., (Villani, 2008)). The considered task is fitting a two-parameter generative model to a Gaussian mixture (parameterized by the means of its two modes). Distance minimization is implemented via gradient descent. Plotted in Figure 3 are $\sqrt{n}$-scaled scatter plots of the estimation errors, with 40 trials for each σ and n pair. The consistent spread of the (scaled) estimation errors as n increases demonstrates the $n^{-1/2}$ convergence rate. The bottom-right subplot shows $\mathsf{W}_2^{(\sigma)}$ estimation errors that further support the fast convergence. In Appendix A.5, we provide additional results for a single Gaussian parameterized by mean and variance.

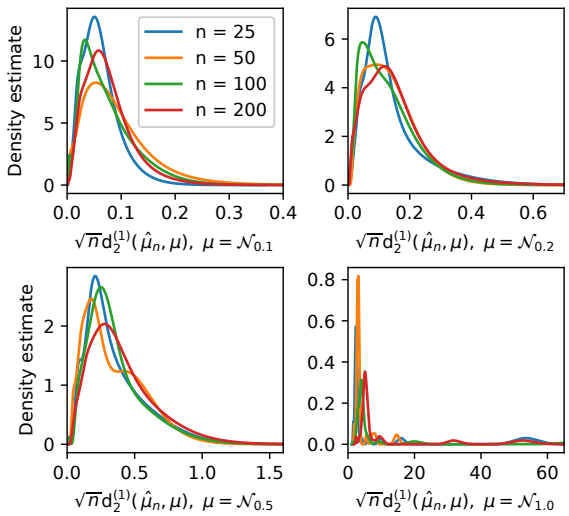

Figure 2. Empirical limiting behavior of $\sqrt{n}\,\mathsf{d}_2^{(1)}(\hat{\mu}_n, \mu)$ for $\mu = \mathcal{N}_s$ with $s \in \{0.1, 0.2, 0.5, 1.0\}$ and $d = 5$.

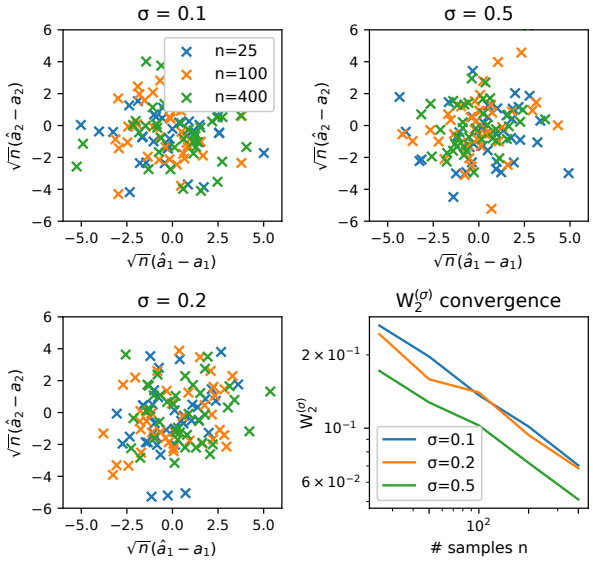

Figure 3. One-dimensional limiting behavior of M-SWE estimates for the two mean parameters of $\mu = \mathcal{N}(a_1, 1)/2 + \mathcal{N}(a_2, 1)/2$ with $a_1 = -1$ and $a_2 = 1$. Also shown is a log-log plot of $\mathsf{W}_2^{(\sigma)}$ convergence in n.

Finally, we provide two-sample testing results in Figure 4 for $p = 1$, leveraging the simplifications discussed at the end of Section 6.1. We approximate the convolved empirical measures by adding Gaussian noise samples and compute W_1 (exactly) for $d = 1$ via its representation as the L^1 distance between quantile functions. For $d = 2$, we estimate W_1 using a standard implementation of W-GAN (Gulrajani et al., 2017; Cao, 2017). For varied sample sizes $n = m$, the quantiles of $\sqrt{\frac{n^2}{N}}\mathsf{W}_1^{(\sigma)}(\hat{\mu}_n^B, \hat{\nu}_n^B)$ are estimated

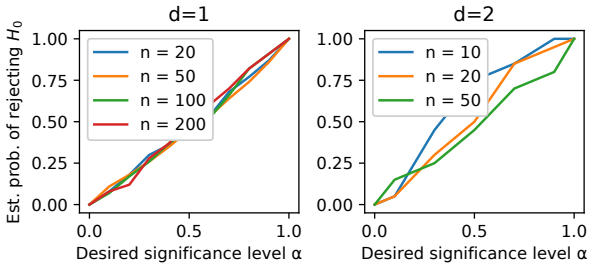

Figure 4. Estimated probability that $\mathsf{W}_1^{(0.1)}$ two-sample test rejects null hypothesis $H_0 : \mu = \nu$ given that $\mu = \nu = \mathsf{Unif}([0,1]^d)$.

using 1000 and 200 bootstrap samples for $d = 1$ and $d = 2$, respectively. The probability of rejecting the null hypothesis for varied significance levels and sample sizes is estimated by repeating the tests over 100 and 200 draws of the original samples, for $d = 1$ and $d = 2$ respectively. Figure 4 displays the probability of false alarm versus the significance level α. Evidently, the curves approximately fall along the diagonal $y = x$, supporting the consistency result.

8. Conclusions and Future Directions

This work provided a thorough analysis of structural and statistical properties of the Gaussian-smoothed Wasserstein distance $\mathsf{W}_p^{(\sigma)}$. While $\mathsf{W}_p^{(\sigma)}$ maintains many desirable properties of standard W_p, we have shown via comparison to the smooth Sobolev IPM $\mathsf{d}_p^{(\sigma)}$ that it admits a parametric empirical convergence rate, avoiding the curse of dimensionality that arises when estimating W_p from data. Using this fast rate and the associated limit distribution for $\mathsf{d}_p^{(\sigma)}$, we have explored new applications to two-sample testing and generative modeling.

An important direction for future research is efficient computation of $\mathsf{W}_p^{(\sigma)}$. While standard methods for computing W_p are applicable in the smooth case (by sampling the noise), it is desirable to find computational techniques that make use of structure induced by the convolution with a known smooth kernel. Furthermore, while $\mathsf{W}_p^{(\sigma)}$ exhibits an expected empirical convergence rate of $O(n^{-1/2})$ that is optimal in n, the prefactor scales exponentially with dimension and warrants additional study. We suspect that this scaling can be shown under a manifold hypothesis to depend only on the intrinsic dimension of the data distribution rather than that of the ambient space.

Finally, we are interested in the limiting behavior of $\mathsf{W}_p^{(\sigma)}$ as $\sigma \to 0$ and $p \to \infty$. The former case has implications for standard W_p and its dependence on intrinsic dimension, as well as for noise annealing that is common in machine learning practice. The latter may connect to differential privacy, where smoothing corresponds to (Gaussian) noise injection and W_∞ underlies the Wasserstein privacy mechanism (Song et al., 2017).

Acknowledgements

S. Nietert is supported by the National Science Foundation (NSF) Graduate Research Fellowship under Grant DGE-1650441. Z. Goldfeld is supported by the NSF CRII grant CCF-1947801, in part by the 2020 IBM Academic Award, and in part by the NSF CAREER Award CCF-2046018. K. Kato is partially supported by NSF grants DMS-1952306 and DMS-2014636.

References

Altschuler, J., Weed, J., and Rigollet, P. Near-linear time approximation algorithms for optimal transport via sinkhorn iteration. In *Advances in Neural Information Processing Systems (NeurIPS-2017)*, pp. 1964–1974, Long Beach, CA, USA, Dec. 2017.

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

Aronszajn, N. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68(3):337–404, 1950.

Arora, S., Ge, R., Liang, Y., Ma, T., and Zhang, Y. Generalization and equilibrium in generative adversarial nets (GANs). In *International Conference on Machine Learning (ICML-2017)*, Sydney, Australia, Aug. 2017.

Benamou, J.-D. and Brenier, Y. A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem. *Numerische Mathematik*, 84(3):375–393, 2000.

Bernton, E., Jacob, P. E., Gerber, M., and Robert, C. P. On parameter estimation with the Wasserstein distance. *Information and Inference. A Journal of the IMA*, 8(4): 657–676, 2019.

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

Bilodeau, G. G. The Weierstrass transform and Hermite polynomials. *Duke Mathematical Journal*, 29:293–308, 1962.

Blanchet, J., Murthy, K., and Zhang, F. Optimal transport based distributionally robust optimization: Structural properties and iterative schemes. *arXiv preprint arXiv:1810.02403*, 2018.

Bogachev, V. I. *Gaussian measures*. American Mathematical Society, 1998.

Boissard, E. and Le Gouic, T. On the mean speed of convergence of empirical and occupation measures in wasserstein distance. *Annales de l'IHP Probabilités et statistiques*, 50(2):539–563, 2014.

Bonnotte, N. *Unidimensional and Evolution Methods for Optimal Transportation*. PhD thesis, Paris-Sud University, 2013.

Borgwardt, K. M., Gretton, A., Rasch, M. J., Kriegel, H.-P., Schölkopf, B., and Smola, A. J. Integrating structured biological data by Kernel Maximum Mean Discrepancy. *Bioinformatics*, 22(14):e49–e57, 07 2006.

Boucheron, S., Lugosi, G., and Massart, P. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013.

Cao, M. WGAN-GP. https://github.com/caogang/wgan-gp, 2017.

Chen, H.-B. and Niles-Weed, J. Asymptotics of smoothed Wasserstein distances. *arXiv preprint arXiv:2005.00738*, 2020.

Courty, N., Flamary, R., and Tuia, D. Domain adaptation with regularized optimal transport. In *European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD 2014)*, Nancy, France, Sep. 2014.

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

Cuturi, M. Sinkhorn distances: Lightspeed computation of optimal transport. In Burges, C. J. C., Bottou, L., Ghahramani, Z., and Weinberger, K. Q. (eds.), *Advances in Neural Information Processing Systems (NIPS-2013)*, pp. 2292–2300, Lake Tahoe, Nevada, USA, Dec. 2013.

del Bariro, E. and Loubes, J.-M. Central limit theorems for empirical transportation cost in general dimension. *The Annals of Probability*, 47(2):926–951, 2019.

del Barrio, E., Giné, E., and Matrán, C. Central limit theorems for the Wasserstein distance between the empirical and the true distributions. *The Annals of Probability*, 27 (2):1009–1071, 1999.

del Barrio, E., Giné, E., and Utzet, F. Asymptotics for L_2 functionals of the empirical quantile process, with applications to tests of fit based on weighted Wasserstein distances. *Bernoulli*, 11(1):131–189, 2005.

Dereich, S., Scheutzow, M., and Schottstedt, R. Constructive quantization: Approximation by empirical measures. *Annales de l'Institut Henri Poincaré Probabilités et Statistiques*, 49(4):1183–1203, 2013.

Deshpande, I., Hu, Y., Sun, R., Pyrros, A., Siddiqui, N., Koyejo, S., Zhao, Z., Forsyth, D. A., and Schwing, A. G. Max-sliced Wasserstein distance and its use for GANs. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR-2019)*, pp. 10648–10656, Long Beach, CA, USA, Jun. 2019.

Dolbeault, J., Nazaret, B., and Savaré, G. A new class of transport distances between measures. *Calculus of Variations and Partial Differential Equations*, 34(2):193–231, 2009.

Feydy, J., Séjourné, T., Vialard, F., Amari, S., Trouvé, A., and Peyré, G. Interpolating between optimal transport and MMD using sinkhorn divergences. In Chaudhuri, K. and Sugiyama, M. (eds.), *The International Conference on Artificial Intelligence and Statistics (AISTATS-2019)*, pp. 2681–2690, Naha, Okinawa, Japan, Apr. 2019.

Fournier, N. and Guillin, A. On the rate of convergence in Wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, 162:707–738, 2015.

Gao, R. and Kleywegt, A. J. Distributionally robust stochastic optimization with Wasserstein distance. *arXiv preprint arXiv:1604.02199*, 2016.

Genevay, A., Cuturi, M., Peyré, G., and Bach, F. R. Stochastic optimization for large-scale optimal transport. In Lee, D. D., Sugiyama, M., von Luxburg, U., Guyon, I., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems (NeurIPS-2016)*, pp. 3432–3440, Barcelona, Spain, Dec. 2016.

Genevay, A., Chizat, L., Bach, F. R., Cuturi, M., and Peyré, G. Sample complexity of sinkhorn divergences. In *The International Conference on Artificial Intelligence and Statistics (AISTATS-2019)*, pp. 1574–1583, Naha, Okinawa, Japan, Apr. 2019.

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

Goldfeld, Z., Greenewald, K., and Kato, K. Asymptotic guarantees for generative modeling based on the smooth Wasserstein distance. In *Advances in Neural Information Processing Systems (NeurIPS-2020)*, 2020a.

Goldfeld, Z., Greenewald, K. H., Niles-Weed, J., and Polyanskiy, Y. Convergence of smoothed empirical measures with applications to entropy estimation. *IEEE Transactions on Information Theory*, 66(7):1489–1501, 2020b.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. A kernel two-sample test. *Journal of Machine Learning Research*, 13:723–773, 2012.

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., and Courville, A. C. Improved training of Wasserstein GANs. In *Advances in Neural Information Processing Systems (NeurIPS-2017)*, Long Beach, CA, USA, Dec. 2017.

Hallin, M., Mordant, G., and Segers, J. Multivariate goodness-of-fit tests based on wasserstein distance. *arXiv preprint arXiv:2003.06684*, 2020.

Ledoux, M. On optimal matching of Gaussian samples. *Journal of Mathematical Science*, 238:495–522, 2019.

Lei, J. Convergence and concentration of empirical measures under Wasserstein distance in unbounded functional spaces. *Bernoulli*, 26(1):767–798, 2020.

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

Mena, G. and Niles-Weed, J. Statistical bounds for entropic optimal transport: Sample complexity and the central limit theorem. In *Advances in Neural Information Processing Systems (NeurIPS-2019)*, pp. 4543–4553, Vancouver, BC, Canada, Dec. 2019.

Milman, E. On the role of convexity in isoperimetry, spectral gap and concentration. *Inventiones Mathematicae*, 177 (1):1–43, 2009.

Mohajerin Esfahani, P. and Kuhn, D. Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations. *Mathematical Programming*, 171:115–166, 2018.

Mroueh, Y., Li, C., Sercu, T., Raj, A., and Cheng, Y. Sobolev GAN. In *International Conference on Learning Representations (ICLR-2018)*, Vancouver, BC, Canada, May 2018.

Nadjahi, K., Durmus, A., Simsekli, U., and Badeau, R. Asymptotic guarantees for learning generative models with the sliced-Wasserstein distance. In *Advances in Neural Information Processing Systems (NeurIPS-2019)*, pp. 250–260, 2019.

Nadjahi, K., Durmus, A., Chizat, L., Kolouri, S., Shahrampour, S., and Simsekli, U. Statistical and topological properties of sliced probability divergences. In *Advances in Neural Information Processing Systems (NeurIPS-2020)*, Dec. 2020.

Oldham, K., Myland, J., and Spanier, J. *An Atlas of Functions*. Springer, second edition, 2009.

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

Peyre, G. Comparison between W_2 distance and $\dot{H}^{-1}$ norm, and localization of Wasserstein distance. *ESAIM: Control, Optimisation and Calculus of Variations*, 24(4):1489–1501, 2018.

Pollard, D. The minimum distance method of testing. *Metrika*, 27(1):43–70, 1980.

Rabin, J., Peyré, G., Delon, J., and Bernot, M. Wasserstein barycenter and its application to texture mixing. In *Scale Space and Variational Methods in Computer Vision (SSVM-2011)*, volume 6667, pp. 435–446, Ein-Gedi, Israel, May 2011.

Ramdas, A., García Trillos, N., and Cuturi, M. On Wasserstein two-sample testing and related families of nonparametric tests. *Entropy*, 19(2):Paper No. 47, 2017.

Rasala, R. The Rodrigues formula and polynomial differential operators. *Journal of Mathematical Analysis and Applications*, 84(2):443–482, 1981.

Rigollet, P. and Weed, J. Entropic optimal transport is maximum-likelihood deconvolution. *Comptes Rendus Mathématique. Académie des Sciences. Paris*, 356(11-12):1228–1235, 2018.

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

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

Santambrogio, F. *Optimal Transport for Applied Mathematicians*. Birkhäuser, 2015.

Schmuland, B. Dirichlet forms with polynomial domain. *Mathematica Japonica*, 37(6):1015–1024, 1992.

Si, N., Blanchet, J. H., Ghosh, S., and Squillante, M. S. Quantifying the empirical wasserstein distance to a set of measures: Beating the curse of dimensionality. In *Advances in Neural Information Processing Systems (NeurIPS-2020)*, pp. 21260–21270, virtual, Dec. 2020.

Singh, S. and Póczos, B. Minimax distribution estimation in wasserstein distance. *arXiv preprint arXiv:1802.08855*, 2018.

Sommerfeld, M. and Munk, A. Inference for empirical Wasserstein distances on finite spaces. *Journal of Royal Statistical Society Series B*, 80:219–238, 2018.

Song, S., Wang, Y., and Chaudhuri, K. Pufferfish privacy mechanisms for correlated data. In *ACM International Conference on Management of Data (SIGMOD-2017*, pp. 1291–1306, New York, NY, USA, May 2017.

Steinwart, I. and Christmann, A. *Support Vector Machines*. Information Science and Statistics. Springer, 2008.

Tameling, C., Sommerfeld, M., and Munk, A. Empirical optimal transport on countable metric spaces: Distributional limits and statistical applications. *Annals of Applied Probability*, 29:2744–2781, 2019.

Tolstikhin, I., Bousquet, O., Gelly, S., and Schölkopf, B. Wasserstein auto-encoders. In *International Conference on Learning Representations (ICLR-2018)*, Vancouver, Canada, Apr.–May 2018.

Vacher, A., Muzellec, B., Rudi, A., Bach, F., and Vialard, F.-X. A dimension-free computational upper-bound for smooth optimal transport estimation. *arXiv preprint arXiv:2101.05380*, 2021.

van der Vaart, A. *Asymptotic Statistics*, volume 3. Cambridge University Press, 1998.

van der Vaart, A. and Wellner, J. A. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer, 1996.

var der Vaart, A. New Donsker classes. *The Annals of Probability*, 24(4):2128–2124, 1996.

Villani, C. *Topics in Optimal Transportation*. American Mathematical Society, 2003.

Villani, C. *Optimal Transport: Old and New*. Springer, 2008.

Weed, J. and Bach, F. Sharp asymptotic and finite-sample rates of convergence of empirical measures in wasserstein distance. *Bernoulli*, 25(4A):2620–2648, 11 2019. doi: 10.3150/18-BEJ1065. URL https://doi.org/10.3150/18-BEJ1065.

Xu, M., Zhou, Z., Lu, G., Tang, J., Zhang, W., and Yu, Y. Towards generalized implementation of wasserstein distance in gans. *arXiv preprint arXiv:2012.03420v2*, 2020.

Zhang, P., Liu, Q., Zhou, D., Xu, T., and He, X. On the discrimination-generalization tradeoff in GANs. In *International Conference on Learning Representations (ICLR-2018)*, Vancouver, BC, Canada, Apr.–May 2018.