Asymptotic Guarantees for Generative Modeling Based on the Smooth Wasserstein Distance

Ziv Goldfeld
Cornell University
goldfeld@cornell.edu

Kristjan Greenewald
MIT-IBM Watson AI Lab
kristjan.h.greenewald@ibm.com

Kengo Kato
Cornell University
kk976@cornell.edu

Abstract

Minimum distance estimation (MDE) gained recent attention as a formulation of (implicit) generative modeling. It considers minimizing, over model parameters, a statistical distance between the empirical data distribution and the model. This formulation lends itself well to theoretical analysis, but typical results are hindered by the curse of dimensionality. To overcome this and devise a scalable finite-sample statistical MDE theory, we adopt the framework of smooth 1-Wasserstein distance (SWD) $\mathsf{W}_1^{(\sigma)}$. The SWD was recently shown to preserve the metric and topological structure of classic Wasserstein distances, while enjoying dimension-free empirical convergence rates. In this work, we conduct a thorough statistical study of the minimum smooth Wasserstein estimators (MSWEs), first proving the estimator's measurability and asymptotic consistency. We then characterize the limit distribution of the optimal model parameters and their associated minimal SWD. These results imply an $O(n^{-1/2})$ generalization bound for generative modeling based on MSWE, which holds in arbitrary dimension. Our main technical tool is a novel high-dimensional limit distribution result for empirical $\mathsf{W}_1^{(\sigma)}$. The characterization of a nondegenerate limit stands in sharp contrast with the classic empirical 1-Wasserstein distance, for which a similar result is known only in the one-dimensional case. The validity of our theory is supported by empirical results, posing the SWD as a potent tool for learning and inference in high dimensions.

1 Introduction

Minimum distance estimation (MDE) considers the minimization of a statistical distance (SD) between the empirical data distribution and a parametric model class. Given an identically and independently distributed (i.i.d.) dataset $X_1, \ldots, X_n$ sampled from P, the goal is to learn a model Q_θ, for $\theta \in \Theta$, that approximates the empirical measure $P_n := n^{-1}\sum_{i=1}^n \delta_{X_i}$ under a SD[1] δ, i.e., we aim to find $\widehat{\theta}_n \in \operatorname{argmin}_{\theta\in\Theta} \delta(P_n, Q_\theta)$. This classic mathematical statistics problem [1–3] was adopted in recent years as a formulation of generative modeling. Indeed, both generative adversarial networks (GANs) [4–11] and variational (or Wasserstein) autoencoders [12, 13] stem from different strategies for (approximately) solving MDE[2] for various choices of δ.

Beyond the practical effectiveness of MDE-based generative models, this formulation is well-suited for a theoretic analysis. This inspired a recent line of works studying GAN generalization in terms of MDEs [9, 14, 15]. Such sample-complexity results boil down to the rate of empirical approximation under the chosen SD, i.e., the speed at which $\delta(P_n, P)$ converges to zero. Unfortunately, popular SDs, such as Wasserstein distances [16], f-divergences [17], and integral probability metrics [18]

[1] Recall that δ is an SD if $\delta(P,Q) = 0 \iff P = Q$.

[2] or a variant thereof, where Q_θ is also estimated from samples.

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

(excluding maximum mean discrepancy [19]) suffer from the curse of dimensionality (CoD), converging as $\delta(P_n,P)\asymp n^{-1/d}$, with d being the data dimension [20–23].[3] This limits the practical usefulness of the devised results, which degrade exponentially fast with dimension.

1.1 MDE with Smooth Wasserstein Distance and Contributions

To circumvent the CoD impasse, we adopt the smooth 1-Wasserstein distance (SWD) [27,28] as our SD. Namely, for any $\sigma>0$, consider $\mathsf{W}_1^{(\sigma)}(P,Q):=\mathsf{W}_1(P*\mathcal{N}_\sigma,Q*\mathcal{N}_\sigma)$, where $\mathcal{N}_\sigma=\mathcal{N}(0,\sigma^2\mathrm{I}_d)$ is the d-dimensional isotropic Gaussian measure of parameter σ, $P*\mathcal{N}_\sigma$ is the convolution of P and $\mathcal{N}_\sigma$, and W_1 is the regular 1-Wasserstein distance (see Section 2 for details). The motivation for this choice is twofold. First, the 1-Wasserstein distance is widely used for generative modeling [8, 13, 29, 30] due to its beneficial attributes, such as metric structure, robustness to support mismatch, compatibility to gradient-based optimization, etc. As shown in [28], these properties are all preserved under Gaussian smoothing. Second, while W_1 suffers from the CoD, [27] showed that $\mathbb{E}\big[\mathsf{W}_1^{(\sigma)}(P_n,P)\big]\lesssim_{\sigma,d} n^{-1/2}$ in all dimensions, whenever P is sub-Gaussian.[4] The considered minimum smooth Wasserstein estimator (MSWE) is thus

$$\widehat{\theta}_n\in\operatorname*{argmin}_{\theta\in\Theta}\mathsf{W}_1^{(\sigma)}(P_n,Q_\theta). \tag{1}$$

We first prove measurability and strong consistency of $\widehat{\theta}_n$, along with almost sure convergence of the associated minimal distance. Moving to a limit distribution analysis, we characterize the high-dimensional limits of $\sqrt{n}(\widehat{\theta}_n-\theta^\star)$ and $\sqrt{n}\inf_{\theta\in\Theta}\mathsf{W}_1^{(\sigma)}(P_n,Q_\theta)$, thus establishing $n^{-1/2}$ convergence rates for both quantities in arbitrary dimension. Leveraging these results along with the framework from [14], we derive a high-dimensional generalization bound of order $n^{-1/2}$ on generative modeling with $\mathsf{W}_1^{(\sigma)}$. Empirical results to support our theory are provided. Using synthetic data we validate both the limiting distributions of parameter estimates and the convergence of the SWD as the number of samples increases.

Our main technical tool is a novel high-dimensional limit distribution result for scaled empirical SWD, i.e., $\sqrt{n}\mathsf{W}_1^{(\sigma)}(P_n,P)$, which may be of independent interest. Our analysis relies on the Kantorovich-Rubinstein (KR) duality for W_1 [16], which allows representing $\sqrt{n}\mathsf{W}_1^{(\sigma)}(P_n,P)$ as a supremum of an empirical process indexed by the class of 1-Lipschitz functions convolved with a Gaussian density. We then prove that this function class is Donsker (i.e., satisfies the uniform central limit theorem (CLT)) under a polynomial moment condition on P.[5] By the continuous mapping theorem, we conclude that $\sqrt{n}\mathsf{W}_1^{(\sigma)}(P_n,P)$ converges in distribution to the supremum of a tight Gaussian process. To enable evaluation of the distributional limit, we also prove that the nonparametric bootstrap is consistent. The characterization of a high-dimensional limit distribution for empirical SWD stands in sharp contrast to the classic W_1 case, for which such a result is known only when $d=1$ [35].

1.2 Comparisons and Related Works

MDE questions similar to those studied herein were addressed for classic W_1 in [36,37] (see also [38, 39]). They derived limit distribution results only for the one-dimensional case, essentially because it is unknown whether a properly scaled $\mathsf{W}_1(P_n,P)$ has a nondegenerate limit in general when $d>1$.

Sliced Wasserstein distance MDE was recently analyzed in [40], covering arbitrary dimension, as done herein. Indeed, both sliced and smooth Wasserstein distances employ different approaches for alleviating the CoD. The sliced version eliminates dependence on d by definition, as it is an average of one-dimensional W_1 distances (via random projections of d-dimensional distributions). SWD, on

[3] One might hope that using more sophisticated estimates of P (instead of the empirical measure) or avoiding plugin methods altogether may alleviate the CoD. However, recent minimax analyses for Wasserstein distances [24], f-divergences [25] and integral probability metrics [26] show that the $n^{-1/d}$ rate is generally unavoidable.

[4] The explicit bound from [31] is $\mathbb{E}\big[\delta^{(\sigma)}(P_n,P)\big]\lesssim\sigma^{-d/2}n^{-1/2}$. While the dependence on n is optimal and decoupled from d (unlike in CoD rates), the prefactor is exponential in d—a dependence that warrants further exploration. See discussion in Section 6.

[5] The reader is referred to, e.g., [32–34] as useful references on modern empirical process theory.

the other hand, does not entail dimensionality reduction, but leverages Gaussian smoothing to level out local irregularities in the high-dimensional distributions, which speeds up empirical convergence rates. This, in turn, enables a thorough MSWE asymptotic analysis for any d. Sliced and smooth Wasserstein distances are also similar in that they are both metrics and (topologically) equivalent to W_1, but there are some notable differences. While sliced W_1 is easily computable using the one-dimensional formula, computational aspects of SWD are still under exploration (see Section 6 for further discussion). SWD might be preferable as a proxy for regular W_1, as the two are within an additive $2\sigma\sqrt{d}$ gap from one another [28, Lemma 1]. Comparison results for sliced Wasserstein seem weaker, assuming compact support and involving implicit dimension-dependent constants (cf., e.g., [41, Lemma 5.1.4]).

Also related to our work is entropic optimal transport (EOT). Its popularity has been driven both by algorithmic advances [42,43] (the latter gives a near-linear-time algorithm) and some statistical properties it possesses [44–46]. Specifically, two-sample empirical estimation under EOT is known to converge as $n^{-1/2}$ for smooth costs (thus, in particular, excluding entropic W_1) with compactly supported distributions [47], or squared cost with subgassian distributions [48]. In comparison, SWD enjoys this fast convergence rate in the stronger one-sample setting and under milder conditions on the distribution. A CLT for empirical EOT under quadratic cost was also derived in [48]. This result is similar to that of [49] for the classic 2-Wasserstein distance, but is markedly different from ours. Notably, [48] derive the CLT with unknown centering constants given by the expected empirical EOT (which differs from the population one). Furthermore, unlike the SWD, EOT is not a metric, even when the underlying cost is [50, 51].[6] In conclusion, while EOT can be efficiently computed, several gaps are still present as far as its statistical properties, and perhaps more importantly, it surrenders some desirable structural properties of classic Wasserstein distances.

Notation. Let $\|\cdot\|$ denote the Euclidean norm, and $x\cdot y$, for $x, y \in \mathbb{R}^d$, designate the inner product. For any probability measure Q on a measurable space $(S, \mathcal{S})$ and any measurable real function f on S, we use the notation $Qf := \int_S f \, \mathsf{d}Q$ whenever the integral exists. We write $a \lesssim_x b$ when $a \le C_x b$ for a constant C_x that depends only on x ($a \lesssim b$ means $a \le Cb$ for an absolute constant C).

We denote by $(\Omega, \mathcal{A}, \mathbb{P})$ the underlying probability space on which all random variables are defined. The class of Borel probability measures on $\mathbb{R}^d$ is $\mathcal{P}(\mathbb{R}^d)$. The subset of measures with finite first moment is denoted by $\mathcal{P}_1(\mathbb{R}^d)$, i.e., $P \in \mathcal{P}_1(\mathbb{R}^d)$ whenever $\int \|x\| \, \mathsf{d}P(x) < \infty$. The convolution of $P, Q \in \mathcal{P}(\mathbb{R}^d)$ is $(P * Q)(\mathcal{A}) := \int\int \mathbb{1}_{\mathcal{A}}(x+y) \, \mathsf{d}P(x) \, \mathsf{d}Q(y)$, where $\mathbb{1}_{\mathcal{A}}$ is the indicator of $\mathcal{A}$. The convolution of measurable functions f, g on $\mathbb{R}^d$ is $f * g(x) = \int f(x-y)g(y) \, \mathsf{d}y$. We also recall that $\mathcal{N}_\sigma := \mathcal{N}(0, \sigma^2 \mathrm{I}_d)$, 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.

For a non-empty set $\mathcal{T}$, let $\ell^\infty(\mathcal{T})$ denote the space of all bounded functions $f : \mathcal{T} \to \mathbb{R}$, equipped with the uniform norm $\|f\|_{\mathcal{T}} := \sup_{t\in\mathcal{T}} \big|f(t)\big|$. We denote $\mathsf{Lip}_1(\mathbb{R}^d) := \{f : \mathbb{R}^d \to \mathbb{R} : |f(x) - f(y)| \le \|x - y\| \; \forall x, y \in \mathbb{R}^d\}$ for the set of Lipschitz continuous functions on $\mathbb{R}^d$ with Lipschitz constant bounded by 1. When d is clear from the context we use the shorthand Lip_1.

2 Background and preliminaries

We next provide a short background on the central technical ideas used in the paper.

1-Wasserstein distance. The 1-Wasserstein distance $\mathsf{W}_1(P, Q)$ between $P, Q \in \mathcal{P}_1(\mathbb{R}^d)$ is

$$\mathsf{W}_1(P, Q) := \inf_{\pi\in\Pi(P,Q)} \int_{\mathbb{R}^d\times\mathbb{R}^d} \|x - y\| \, \mathsf{d}\pi(x, y),$$

where $\Pi(P, Q)$is the set of all couplings of P and Q. The KR duality further implies $\mathsf{W}_1(P, Q) = \sup_{f\in\mathsf{Lip}_1} \int_{\mathbb{R}^d} f \, \mathsf{d}(P - Q)$. See [16] for additional background.

Empirical approximation. Fix $P \in \mathcal{P}_1(\mathbb{R}^d)$ and let $X_1, \dots, X_n \sim P$ be i.i.d. Let $P_n = n^{-1} \sum_{i=1}^n \delta_{X_i}$ be the empirical distribution of $X_1, \dots, X_n$, where δ_x is the Dirac measure at x. The convergence rate of $\mathsf{W}_1(P_n, P)$ received much attention in the literature; see, e.g., [20,52–58].[7]

[6] EOT can be transformed into a Sinkhorn divergence via a simple modification, but it is still is not a metric since it lacks the triangle inequality [51].

[7] Those references also contain results on the more general Wasserstein distance and non-Euclidean spaces.

Sharp rates are known in all dimensions;[8] $\mathbb{E}[\mathsf{W}_1(P_n,P)] = O(n^{-1/2})$ if $d=1$, $= O(n^{-1/2}\log n)$ if $d=2$, and $= O(n^{-1/d})$ for $d \geq 3$ provided that P has sufficiently many moments (cf. [20]).

Limit distribution. Despite the comprehensive account of the expected $\mathsf{W}_1(P_n,P)$, limiting distribution results for a scaled version thereof are known only for $d=1$. Indeed, Theorem 2 in [59] yields that $\mathsf{Lip}_1(\mathbb{R})$ is a P-Donsker class if (and only if) $\sum_j P\big([-j,j]^c\big)^{1/2} < \infty$. Combining with KR duality, we have $\sqrt{n}\mathsf{W}_1(P_n,P) \stackrel{d}{\to} \sup_{f\in\mathsf{Lip}_1(\mathbb{R})} G_P(f)$ for some tight Gaussian process G_P in $\ell^\infty(\mathsf{Lip}_1(\mathbb{R}))$. An alternative derivation of the limit distribution for $d=1$ is given in [35], based on the fact that W_1 equals the L^1 distance between distribution functions when $d=1$. The arguments in those papers, however, do not carry over to general d. For $d\geq 2$, in general, the function class $\mathsf{Lip}_1(\mathbb{R}^d)$ is not Donsker; if it was, then $\mathbb{E}[\mathsf{W}_1(P_n,P)]$ would be of order $O(n^{-1/2})$, contradicting existing results lower bounding the rate of convergence of $\mathsf{W}_1(P_n,P)$ [20].

Smooth Wasserstein distance. We are interested in $d \geq 2$, and instead of W_1 consider the SWD [27, 28] $\mathsf{W}_1^{(\sigma)}(P,Q) := \mathsf{W}_1(P*\mathcal{N}_\sigma, Q*\mathcal{N}_\sigma)$. [27] shows that $\mathsf{W}_1^{(\sigma)}(P_n,P) = O_P(n^{-1/2})$, for all d and any sub-Gaussian P. Herein, we characterize the limit distribution of $\sqrt{n}\mathsf{W}_1^{(\sigma)}(P_n,P)$, prove that this distribution can be accurately estimated via the bootstrap, and derive concentration inequalities (see Supplement A.1 for the latter). To simplify discussions, henceforth we assume $0 < \sigma \leq 1$.

Stochastic processes. A stochastic process $G := \big(G(t)\big)_{t\in\mathcal{T}}$ indexed by $\mathcal{T}$ is Gaussian if the $\big(G(t_i)\big)_{i=1}^k$ are jointly Gaussian for any finite collection $\{t_i\}_{i=1}^k \subset \mathcal{T}$. A Gaussian process G is tight in $\ell^\infty(\mathcal{T})$ if and only if $\mathcal{T}$ is totally bounded for the pseudometric $d_G(s,t) = \sqrt{\mathbb{E}\big[|G(s)-G(t)|^2\big]}$, and G has sample paths a.s. uniformly d_G-continuous [33, Section 1.5]. If G is sample bounded, we view it as a mapping from the sample space into $\ell^\infty(\mathcal{T})$. A version of a stochastic process is another stochastic process with the same finite dimensional distributions.

3 Limit distribution theory for smooth Wasserstein distance

The main technical tool for treating MSWE is a characterization of the limit distribution of $\sqrt{n}\mathsf{W}_1^{(\sigma)}(P_n,P)$ in all dimensions, which is the focus of this section. We also derive consistency of the bootstrap as a means for computing the limit distribution, and establish concentration inequalities for $\mathsf{W}_1^{(\sigma)}(P_n,P)$ (see Supplement A.1 for the latter).

Starting from the limit distribution of $\sqrt{n}\mathsf{W}_1^{(\sigma)}(P_n,P)$, some definitions are needed to describe the limit random variable. Denote $\mathsf{Lip}_{1,0} := \{f \in \mathsf{Lip}_1 : f(0) = 0\}$, assume that $P\|x\|^2 < \infty$, and let $G_P^{(\sigma)} = \big(G_P^{(\sigma)}(f)\big)_{f\in\mathsf{Lip}_{1,0}}$ be a centered Gaussian process with covariance function $\mathbb{E}\big[G_P^{(\sigma)}(f)G_P^{(\sigma)}(g)\big] = \mathsf{Cov}_P(f*\varphi_\sigma, g*\varphi_\sigma)$, where $f,g,\in \mathsf{Lip}_{1,0}$. One may verify that $|f*\varphi_\sigma(x)| \leq \|x\| + \sigma\sqrt{d}$ (cf. Section A.2), so that $P|f*\varphi_\sigma|^2 < \infty$, for all $f \in \mathsf{Lip}_{1,0}$ (which ensures that the covariance function above is well-defined). With that, we are ready to state the theorem.

Theorem 1 (SWD limit distribution). *Assume that $P\|x\|^2 < \infty$. Let $\mathbb{R}^d = \bigcup_{j=1}^\infty I_j$ be a partition of $\mathbb{R}^d$ into bounded convex sets with nonempty interior such that $K := \sup_j \mathsf{diam}(I_j) < \infty$. If*

$$\sum\nolimits_{j=1}^\infty M_j P(I_j)^{1/2} < \infty \quad \textit{with} \quad M_j := \sup_{I_j} \|x\|, \tag{2}$$

then there exists a version of $G_P^{(\sigma)}$ that is tight in $\ell^\infty(\mathsf{Lip}_{1,0})$, and denoting the tight version by the same symbol $G_P^{(\sigma)}$, we have $\sqrt{n}\mathsf{W}_1^{(\sigma)}(P_n,P) \stackrel{d}{\to} \sup_{f\in\mathsf{Lip}_{1,0}} G_P^{(\sigma)}(f) =: L_P^{(\sigma)}$. In addition, we have $\sqrt{n}\mathbb{E}\left[\mathsf{W}_1^{(\sigma)}(P_n,P)\right] \lesssim_{d,K} \sigma^{-\lfloor d/2\rfloor}\sum_{j=1}^\infty M_j P(I_j)^{1/2}$.

The proof is given in Supplement A.2. We use KR duality to translate the Gaussian convolution in the measure space to the convolution of Lipschitz functions with a Gaussian density. It is then shown that

[8] Except $d = 2$, where a log factor is possibly missing.

this class of Gaussian-smoothed Lipschitz functions is P-Donsker by bounding the metric entropy of the function class restricted to each I_j. The proof substantially relies on empirical process theory.

Remark 1 (Discussion on Condition (2))**.** *Let $\{I_j\}$ consist of cubes with side length 1 and integral lattice points as vertices. One may then obtain the bound*

$$\sum_{j=1}^{\infty} M_j P(I_j)^{1/2} \lesssim_d \sum_{k=1}^{\infty} k^d P\big(\|x\|_\infty > k\big)^{1/2} \lesssim \int_1^\infty t^d P\big(\|x\|_\infty > t\big)^{1/2}\, \mathsf{d}t,$$

which is finite (by Markov's inequality) if there exists $\epsilon > 0$ such that $P|x_j|^{2(d+1)+\epsilon} < \infty$ for all j.

Proposition 1 in [27] shows that $\mathbb{E}\big[\mathsf{W}_1^{(\sigma)}(P_n, P)\big] = O(n^{-1/2})$ whenever P is sub-Gaussian. Theorem 1 substantially relaxes this moment condition, in addition to deriving a limit distribution.

Remark 2 (Limit distribution for empirical W_p)**.** *The limit distribution of $\sqrt{n}\mathsf{W}_p(P_n, P)$, when P is supported on a finite or a countable set, was derived in [60] and [61], respectively. [62] show asymptotic normality of $\sqrt{n}(\mathsf{W}_2(P_n, Q) - \mathbb{E}[\mathsf{W}_2(P_n, Q)])$, in arbitrary dimension, but under the assumption that $Q \neq P$. The limit distribution for the empirical 2-Wasserstein distance with $Q = P$ is known only when $d = 1$ [63]. None of the techniques employed in these works are applicable in our case, which therefore requires a different analysis as described above.*

The proof of Theorem 1 along with Lemma 3 from Supplement A.3 implies that the distribution of $L_P^{(\sigma)}$ can be estimated via the bootstrap [33, Chapter 3.6]. Let $X_1^B, \ldots, X_n^B$ be i.i.d. from P_n conditioned on $X_1, \ldots, X_n$, and set $P_n^B := n^{-1}\sum_{i=1}^n \delta_{X_i^B}$ as the bootstrap empirical distribution. Let $\mathbb{P}^B$ be the probability measure induced by the bootstrap (i.e., the conditional probability given $X_1, X_2, \ldots$).

Corollary 1 (Bootstrap consistency)**.** *Assume the conditions of Theorem 1 and that P is not a point mass. Then, we have $\sup_{t\geq 0}\big|\mathbb{P}^B\big(\sqrt{n}\mathsf{W}_1^{(\sigma)}(P_n^B, P_n) \leq t\big) - \mathbb{P}\big(L_P^{(\sigma)} \leq t\big)\big| \to 0$ a.s.*

This corollary, together with continuity of the distribution function of $L_P^{(\sigma)}$ (cf. Lemma 3 in the Appendix), implies that for $\widehat{q}_{1-\alpha} := \inf\{t \geq 0 : \mathbb{P}^B\big(\sqrt{n}\mathsf{W}_1^{(\sigma)}(P_n^B, P_n) \leq t\big) \geq 1 - \alpha\}$ (which can be computed numerically), we have $\mathbb{P}\big(\sqrt{n}\mathsf{W}_1^{(\sigma)}(P_n, P) > \widehat{q}_{1-\alpha}\big) = \alpha + o(1)$.

Remark 3 (Two-sample setting)**.** *Theorem 1 and Corollary 1 can be extended to the two-sample case, i.e., accounting for $\mathsf{W}_1^{(\sigma)}(P_n, Q_m)$. The proof of Theorem 1 shows that the function class $\mathcal{F}_{\sigma,d} := \big\{f * \varphi_\sigma : f \in \mathsf{Lip}_{1,0}\big\}$ is Donsker, for all d. Consequently, $\sqrt{\frac{mn}{m+n}}\mathsf{W}_1^{(\sigma)}(P_n, Q_m)$ converges in distribution to the supremum of a tight Gaussian process, if the population distributions agree (cf. [64, Chapter 3.7]). By [64, Theorem 3.7.6], this limit process can be consistently estimated by the two-sample bootstrap. Two-sample testing with (unsmoothed) W_p is studied in [65], but they find critical values for the tests only for $d = 1$. This is due to lack of tractable distribution approximation results for W_p in high dimensions. Our theory shows that we can overcome this bottleneck by adopting $\mathsf{W}_1^{(\sigma)}$.*

4 Minimum Smooth Wasserstein Estimation

We study the statistical properties of the MSWE $\widehat{\theta}_n \in \operatorname{argmin}_{\theta\in\Theta} \mathsf{W}_1^{(\sigma)}(P_n, Q_\theta)$ in high dimensions. Here $P \in \mathcal{P}_1(\mathbb{R}^d)$, P_n is the associated empirical measure, and $Q_\theta \in \mathcal{P}_1(\mathbb{R}^d)$, where $\theta \in \Theta \subset \mathbb{R}^{d_0}$, is the model class. We henceforth assume (without further mentioning) that the parameter space $\Theta \subset \mathbb{R}^{d_0}$ is compact with nonempty interior. The boundedness assumption on Θ can be weakened with some adjustments to the proofs of Theorems 2 and 3 below; cf. [37, Assumption 2.3].

4.1 Measurability and Consistency

The following theorem states that the MSWE is measurable. The proof (given in Supplement B.2) relies on Corollary 1 in [66], which provides a sufficient condition for the desired measurability.

Theorem 2 (MSWE measurability)**.** *Assume that the map $\theta \mapsto Q_\theta$ is continuous relative to the weak topology,*[9] *i.e., $Q_\theta \rightharpoonup Q_{\overline{\theta}}$ whenever $\theta \to \overline{\theta}$ in Θ. Then, for every $n \in \mathbb{N}$, there exists a measurable function $\omega \mapsto \widehat{\theta}_n(\omega)$ such that $\widehat{\theta}_n(\omega) \in \operatorname{argmin}_{\theta\in\Theta} \mathsf{W}_1^{(\sigma)}\big(P_n(\omega), Q_\theta\big)$ for every $\omega \in \Omega$ (this also implies that $\operatorname{argmin}_{\theta\in\Theta} \mathsf{W}_1^{(\sigma)}\big(P_n(\omega), Q_\theta\big)$ is nonempty).*

Next, we establish consistency of the MSWE. The proof relies on [67, Theorem 7.33]. To apply it, we verify epi-convergence of the map $\theta \mapsto \mathsf{W}_1^{(\sigma)}(P_n, Q_\theta)$ towards $\theta \mapsto \mathsf{W}_1^{(\sigma)}(P, Q_\theta)$ (after proper extensions). See Supplement B.3 for details.

Theorem 3 (MSWE consistency)**.** *Assume that the map $\theta \mapsto Q_\theta$ is continuous relative to the weak topology. Then, we have $\inf_{\theta\in\Theta} \mathsf{W}_1^{(\sigma)}(P_n, Q_\theta) \to \inf_{\theta\in\Theta} \mathsf{W}_1^{(\sigma)}(P, Q_\theta)$ a.s. In addition, there exists an event with probability one on which the following holds: for any sequence $\{\widehat{\theta}_n\}_{n\in\mathbb{N}}$ of measurable estimators such that $\mathsf{W}_1^{(\sigma)}(P_n, Q_{\widehat{\theta}_n}) \le \inf_{\theta\in\Theta} \mathsf{W}_1^{(\sigma)}(P_n, Q_\theta) + o(1)$, the set of cluster points of $\{\widehat{\theta}_n\}_{n\in\mathbb{N}}$ is included in $\operatorname{argmin}_{\theta\in\Theta} \mathsf{W}_1^{(\sigma)}(P, Q_\theta)$. In particular, if $\operatorname{argmin}_{\theta\in\Theta} \mathsf{W}_1^{(\sigma)}(P, Q_\theta)$ is unique, i.e., $\operatorname{argmin}_{\theta\in\Theta} \mathsf{W}_1^{(\sigma)}(P, Q_\theta) = \{\theta^\star\}$, then $\widehat{\theta}_n \to \theta^\star$ a.s.*

4.2 Limit Distributions

We study the limit distributions of the MSWE and the associated SWD. Results are presented for the 'well-specified' setting, i.e., when $P = Q_{\theta^\star}$ for some $\theta^\star$ in the interior of $\Theta \subset \mathbb{R}^{d_0}$. Extensions to the 'misspecified' case are straightforward (cf. [37, Theorem B.8]). Our derivation leverages the method of [2] for MDE analysis over normed spaces. To make the connection, we need some definitions.

For any $G \in \ell^\infty(\mathsf{Lip}_{1,0})$, define $\|G\|_{\mathsf{Lip}_{1,0}} := \sup_{f\in\mathsf{Lip}_{1,0}} |G(f)|$. With any $Q \in \mathcal{P}_1(\mathbb{R}^d)$, associate the functional $Q^{(\sigma)} : \mathsf{Lip}_{1,0} \to \mathbb{R}$ defined by $Q^{(\sigma)}(f) := Q(f * \varphi_\sigma) = (Q * \mathcal{N}_\sigma)(f)$. Note that $\big\|Q^{(\sigma)}\big\|_{\mathsf{Lip}_{1,0}} := \sup_{f\in\mathsf{Lip}_{1,0}} \big|Q^{(\sigma)}(f)\big|$ is finite as $Q \in \mathcal{P}_1(\mathbb{R}^d)$ and $|(f * \varphi_\sigma)(x)| \le \|x\| + \sigma\sqrt{d}$ for any $f \in \mathsf{Lip}_{1,0}$. Consequently, $Q^{(\sigma)} \in \ell^\infty(\mathsf{Lip}_{1,0})$ for any $Q \in \mathcal{P}_1(\mathbb{R}^d)$. Finally, observe that $\mathsf{W}_1^{(\sigma)}(P, Q) = \big\|P^{(\sigma)} - Q^{(\sigma)}\big\|_{\mathsf{Lip}_{1,0}}$, for any $P, Q \in \mathcal{P}_1(\mathbb{R}^d)$ (cf. Supplement A.2).

SWD limit distribution. We start from the limit distribution of the (scaled) infimized SWD. This result is central for deriving the limiting MSWE distribution (see Theorem 5 and Corollary 2 below). Theorem 4 is proven in Supplement B.4 via an adaptation of the argument from [2, Theorem 4.2].

Theorem 4 (Minimal SWD limit distribution)**.** *Let P satisfy the conditions of Theorem 1. In addition, suppose that (i) the map $\theta \mapsto Q_\theta$ is continuous relative to the weak topology; (ii) $P \neq Q_\theta$ for any $\theta \neq \theta^\star$; (iii) there exists a vector-valued functional $D^{(\sigma)} \in (\ell^\infty(\mathsf{Lip}_{1,0}))^{d_0}$ such that $\big\|Q_\theta^{(\sigma)} - Q_{\theta^\star}^{(\sigma)} - \langle \theta - \theta^\star, D^{(\sigma)}\rangle\big\|_{\mathsf{Lip}_{1,0}} = o(\|\theta - \theta^\star\|)$ as $\theta \to \theta^\star$, where $\langle t, D^{(\sigma)}\rangle := \sum_{i=1}^{d_0} t_i D_i^{(\sigma)}$ for $t \in \mathbb{R}^{d_0}$; (iv) the derivative $D^{(\sigma)}$ is nonsingular in the sense that $\langle t, D^{(\sigma)}\rangle \neq 0$, i.e., $\langle t, D^{(\sigma)}\rangle \in \ell^\infty(\mathsf{Lip}_{1,0})$ is not the zero functional for all $0 \neq t \in \mathbb{R}^{d_0}$. Then, $\sqrt{n}\inf_{\theta\in\Theta} \mathsf{W}_1^{(\sigma)}(P_n, Q_\theta) \xrightarrow{d} \inf_{t\in\mathbb{R}^{d_0}} \big\|G_P^{(\sigma)} - \big\langle t, D^{(\sigma)}\big\rangle\big\|_{\mathsf{Lip}_{1,0}}$, where $G_P^{(\sigma)}$ is the Gaussian process from Theorem 1.*

Remark 4 (Norm differentiability)**.** *Condition (iii) in Theorem 4 is called 'norm differentiability' in [2]. In these terms, the theorem assumes that the map $\theta \mapsto Q_\theta^{(\sigma)}, \Theta \to \ell^\infty(\mathsf{Lip}_{1,0})$, is norm differentiable around $\theta^\star$ with derivative $D^{(\sigma)}$. This allows approximating the map $\theta \mapsto Q_\theta^{(\sigma)}$ by the affine function $Q_{\theta^\star}^{(\sigma)} + \big\langle \theta - \theta^\star, D^{(\sigma)}\big\rangle$ near $\theta^\star$. Together with the result of Theorem 1 and the right reparameterization, norm differentiability is key for establishing the theorem.*

Remark 5 (Primitive conditions for norm differentiability)**.** *Suppose that $\{Q_\theta\}_{\theta\in\Theta}$ is dominated by a common Borel measure ν on $\mathbb{R}^d$, and let q_θ denote the density of Q_θ with respect to ν, i.e., $\mathsf{d}Q_\theta = q_\theta\, \mathsf{d}\nu$. Then, $Q_\theta * \mathcal{N}_\sigma$ has Lebesgue density $x \mapsto \int \varphi_\sigma(x-t) q_\theta(x)\, \mathsf{d}\nu(t)$. Assume that q_θ admits the Taylor expansion $q_\theta(x) = q_{\theta^\star}(x) + \dot{q}_{\theta^\star}(x) \cdot (\theta - \theta^\star) + r_\theta(x) \cdot (\theta - \theta^\star)$ with $r_\theta(x) = o(1)$ as $\theta \to \theta^\star$.*

[9] The *weak topology* on $\mathcal{P}(\mathbb{R}^d)$ is induced by integration against the set $C_b(\mathbb{R}^d)$ of bounded and continuous functions, i.e., $(\mu_k)_{k\in\mathbb{N}}$ converges weakly to μ, denoted by $\mu_k \rightharpoonup \mu$, if $\mu_k(f) \to \mu(f)$, for all $f \in C_b(\mathbb{R}^d)$.

Then, one may verify that Condition (iii) holds with $D^{(\sigma)}(f) = \int f(x) \int \varphi_\sigma(x-t)\dot{q}_{\theta^\star}(t)\,\mathsf{d}\nu(t)\,\mathsf{d}x = \int (f*\varphi_\sigma)(t)\dot{q}_{\theta^\star}(t)\,\mathsf{d}\nu(t)$*, for* $f \in \mathsf{Lip}_{1,0}$*, provided that* $\int(1+\|t\|)\|\dot{q}_{\theta^\star}(t)\|\,\mathsf{d}\nu(t) < \infty$ *and* $\int(1+\|t\|)\|r_\theta(t)\|\,\mathsf{d}\nu(t) = o(1)$ *(use the fact that* $|f(t)| \leq \|t\|$*, for any* $f \in \mathsf{Lip}_{1,0}$*).*

MSWE limit distribution. We study convergence in distribution of the MSWE. Optimally, the limit distribution of $\sqrt{n}(\widehat{\theta}_n - \theta^\star)$, for some $\widehat{\theta}_n \in \operatorname{argmin}_{\theta\in\Theta} \mathsf{W}_1^{(\sigma)}(P_n, Q_\theta)$, is the object of interest. However, a limit is guaranteed to exist only when the (convex) function $t \mapsto \big\|G_P^{(\sigma)} - \langle t, D^{(\sigma)}\rangle\big\|_{\mathsf{Lip}_{1,0}}$ has a unique minimum a.s. (see Corollary 2 below). To avoid this stringent assumption, before treating $\sqrt{n}(\widehat{\theta}_n - \theta^\star)$, we first consider the set of approximate minimizers $\widehat{\Theta}_n := \big\{\theta \in \Theta : \mathsf{W}_1^{(\sigma)}(P_n, Q_\theta) \leq \inf_{\theta'\in\Theta} \mathsf{W}_1^{(\sigma)}(P_n, Q_{\theta'}) + \lambda_n/\sqrt{n}\big\}$, where $\{\lambda_n\}_{n\in\mathbb{N}}$ is an arbitrary $o_{\mathbb{P}}(1)$ sequence.

We show that $\widehat{\Theta}_n \subset \theta^\star + n^{-1/2}K_n$ for some (random) sequence of compact convex sets $\{K_n\}_{n\in\mathbb{N}}$ with inner probability approaching one. Resorting to inner probability seems inevitable since the event $\{\widehat{\Theta}_n \subset \theta^\star + n^{-1/2}K_n\}$ need not be measurable in general (see [2, Section 7]). To define such sequence $\{K_n\}_{n\in\mathbb{N}}$, for any $L \in \ell^\infty(\mathsf{Lip}_{1,0})$ and $\beta \geq 0$, let $K(L,\beta) := \Big\{t \in \mathbb{R}^{d_0} : \big\|L - \langle t, D^{(\sigma)}\rangle\big\|_{\mathsf{Lip}_{1,0}} \leq \inf_{t'\in\mathbb{R}^d} \big\|L - \langle t', D^{(\sigma)}\rangle\big\|_{\mathsf{Lip}_{1,0}} + \beta\Big\}$. Lemma 7.1 of [2] ensures that for any $\beta \geq 0$, $L \mapsto K(L,\beta)$ is a measurable map from $\ell^\infty(\mathsf{Lip}_{1,0})$ into $\mathfrak{K}$ – the class of all compact, convex, and nonempty subsets of $\mathbb{R}^{d_0}$ – endowed with the Hausdorff topology. That is, the topology induced by the metric $d_{\mathsf{H}}(K_1,K_2) := \inf\big\{\delta > 0 : K_1^\delta \supset K_2,\ K_2^\delta \supset K_1\big\}$, where $K^\delta := \bigcup_{x\in K}\big\{y \in \mathbb{R}^{d_0} : \|y - x\| \leq \delta\big\}$ is the δ-blowup of K.

Theorem 5 (MSWE limit distribution)**.** *Under the conditions of Theorem 4, there exists a sequence of nonnegative reals* $\beta_n \downarrow 0$ *such that (i)* $\mathbb{P}_*\big(\widehat{\Theta}_n \subset \theta^\star + n^{-1/2}K(\mathbb{G}_n^{(\sigma)}, \beta_n)\big) \to 1$*, where* $\mathbb{G}_n^{(\sigma)} := \sqrt{n}(P_n^{(\sigma)} - P^{(\sigma)})$ *is the (smooth) empirical process and* $\mathbb{P}_*$ *denotes inner probability; and (ii)* $K(\mathbb{G}_n^{(\sigma)}, \beta_n) \xrightarrow{d} K(G_P^{(\sigma)}, 0)$ *as* $\mathfrak{K}$*-valued random variables.*

Given Theorem 4, the proof of Theorem 5 follows by a verbatim repetition of the argument from [2, Section 7.2]. The details are therefore omitted. If $\operatorname{argmin}_{t\in\mathbb{R}^{d_0}} \big\|G_P^{(\sigma)} - \langle t, D^{(\sigma)}\rangle\big\|_{\mathsf{Lip}_{1,0}}$ is unique a.s. (a nontrivial assumption), then Theorem 5 simplifies as follows.[10]

Corollary 2 (Simplified MSWE limit distribution)**.** *Assume the conditions of Theorem 4. Let* $\{\widehat{\theta}_n\}_{n\in\mathbb{N}}$ *be a sequence measurable estimators such that* $\mathsf{W}_1^{(\sigma)}(P_n, Q_{\widehat{\theta}_n}) \leq \inf_{\theta\in\Theta} \mathsf{W}_1^{(\sigma)}(P_n, Q_\theta) + o_{\mathbb{P}}(n^{-1/2})$*. Then, provided that* $\operatorname{argmin}_{t\in\mathbb{R}^{d_0}} \big\|G_P^{(\sigma)} - \langle t, D^{(\sigma)}\rangle\big\|_{\mathsf{Lip}_{1,0}}$ *is unique a.s., we have* $\sqrt{n}(\widehat{\theta}_n - \theta^\star) \xrightarrow{d} \operatorname{argmin}_{t\in\mathbb{R}^{d_0}} \big\|G_P^{(\sigma)} - \langle t, D^{(\sigma)}\rangle\big\|_{\mathsf{Lip}_{1,0}}$*.*

Corollary 2 is proven in Supplement B.5 using weak convergence of argmin of random convex maps.

Generalization discussion. Theorem 5 and Corollary 2 imply a generalization bound for MSWE-based generative model. Focusing on the corollary, $\widehat{\theta}_n$ is an approximate optimizer (e.g., obtained via some suboptimal gradient-based optimization) of the empirical MSWE problem $\mathsf{W}_1^{(\sigma)}(P_n, Q_\theta)$. The goal is to obtain a $\widehat{\theta}_n$ that approximates as best as possible the minimizer $\theta^\star$ of the population loss $\mathsf{W}_1^{(\sigma)}(P, Q_\theta)$. Corollary 2 thus states that the MSWE $\widehat{\theta}_n$ converges to the true optimum $\theta^\star$ (in expectation, in distribution, and with high probability) at a dimension-free rate of $n^{-1/2}$, which corresponds to the notion of GAN generalization from [9, 14]. In fact, by Eq. (10) from [14], we further have that $\mathsf{W}_1^{(\sigma)}(P, Q_{\widehat{\theta}_n}) - \inf_{\theta\in\Theta} \mathsf{W}_1^{(\sigma)}(P, Q_\theta) \lesssim n^{-1/2}$, with high probability.

[10] Note that $\operatorname{argmin}_{t\in\mathbb{R}^{d_0}} \big\|G_P^{(\sigma)} - \langle t, D^{(\sigma)}\rangle\big\|_{\mathsf{Lip}_{1,0}} \neq \emptyset$ provided that $D^{(\sigma)}$ is nonsingular, since the latter guarantees that $\big\|G_P^{(\sigma)} - \langle t, D^{(\sigma)}\rangle\big\|_{\mathsf{Lip}_{1,0}} \to \infty$ as $\|t\| \to \infty$.

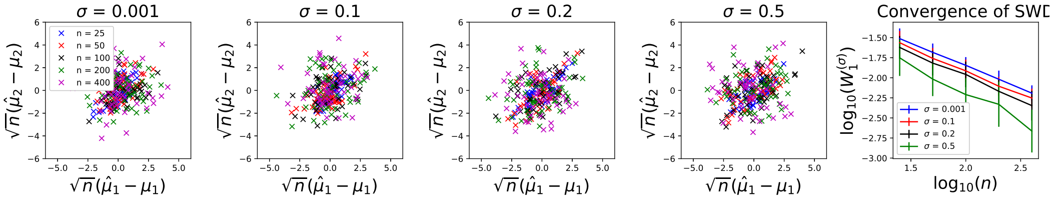
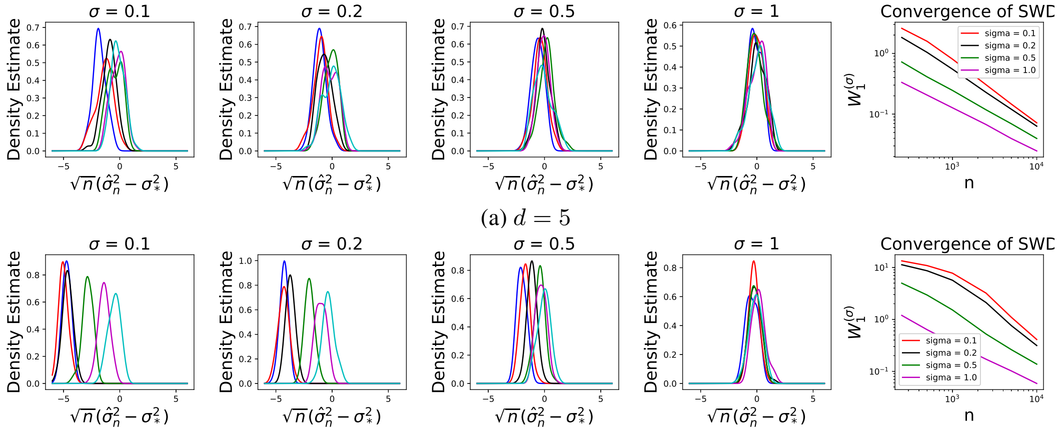


Figure 1: One-dimensional limiting distributions for the two mean parameters of the mixture $P = 0.5\mathcal{N}(\mu_1, 1) + 0.5\mathcal{N}(\mu_2, 1)$, for $\mu_1 = 0$ and $\mu_2 = 1$. Also shown on a log-log scale (with error bars) is the SWD convergence as a function of n.



(a) $d = 5$

(b) $d = 10$

Figure 2: Empirical limiting distributions for the variance parameter of an MSWE-based generative model fitted to $P = \mathcal{N}_1$. Also shown as a log-log plot is the SWD convergence as a function of n.

5 Empirical Results

We provide experiments on synthetic data validating our theory. We start with a one-dimensional setting since then an exact expression for the SWD is available (as the L^1 distance between cumulative distribution functions [68]). Afterwards, higher dimensional problems are explored using an estimator based on the neural network (NN) parameterized KR dual form of W_1.

Fig. 1 shows results for fitting two-parameter generative models in one dimension for a Gaussian mixture (parameterized by the two means, one from each mode). $\sqrt{n}$-scaled scatter plots of the estimation error are shown for various σ and n values, each formed from 50 estimation trials. Convergence of the corresponding SWD losses is shown on the right. Note that the point clouds closely overlap in each plot even as n increases, implying that indeed a limiting distribution is emerging, as predicted by the theory. In particular, the spread of the scatter plots does not increase despite a 16-fold increase in n, and the SWD loss converges at approximately an $n^{-1/2}$ rate. Supplement C gives additional results for a single Gaussian (parameterized by mean and variance).

In higher dimensions, the MSWE is computed as follows. We first draw n samples from P and obtain the empirical measure P_n. Sampling from P_n and $\mathcal{N}_\sigma$ and adding the obtained values produces samples from $P_n * \mathcal{N}_\sigma$. Applying similar steps to Q_θ, we may compute $\mathsf{W}_1^{(\sigma)}(P_n, Q_\theta)$ by applying standard W_1 estimators to samples from the convolved measures. We use the NN-based estimator for WGAN-GP discriminator from [29]. As a side note, we believe that more effective estimators that are tailored for the SWD structure are possible, but leave this exploration to future work.

Fig. 2 shows MSWE results in dimensions 5 and 10. The target distribution is a multivariate standard Gaussian $P = \mathcal{N}_{\sigma_\star}$ for $\sigma_\star = 1$. The model Q_θ is also an isotropic Gaussian, with a single (variance) parameter. The WGAN-GP discriminator has 3 hidden layers with 512 hidden units each. The resulting distribution of $\sqrt{n}(\hat{\sigma}_n^2 - \sigma_\star^2)$ is shown for various values of σ (the SWD smoothing parameter) and number of samples n. These distributions are computed using a kernel density estimate on 50 random trials. As seen in the figure, the distribution of $\sqrt{n}(\widehat{\sigma}_n^2 - \sigma_\star^2)$ converges to a clear limit as n increases, for all $\sigma > 0$ values (although convergence is slower for smaller σ). When σ is smaller, i.e., closer to the classic W_1 case, convergence is less pronounced, especially in higher dimensions. Finally, note that as predicted by our theory, the MSWE convergence rate is $n^{-1/2}$.

Lastly, we consider a more complex target P and parameterize Q_θ via a three-layer neural network with 256 hidden units per layer. Note that this corresponds to the generator in a GAN setup, where the parameters θ of the neural network are learned so that Q_θ matches a target distribution. We combine this with our neural SWD-based discriminator, effectively creating an SWD GAN that we train in a way similar to WGAN-GP. Setting $d = 10$, we take P as a 2^d-mode Gaussian mixture formed by equal-weighted isotropic Gaussians (with variance parameter 1) centered at each corner of the $[-1,1]^d$ hypercube. As there are too many parameters to visualize the limiting distribution, Fig. 3 instead shows the SWD convergence versus the number of samples n. As predicted, for $\sigma > 0$, the SWD asymptotically converges as approximately $n^{-1/2}$, though for smaller σ this rate only kicks in for higher values of n. This two-phase behavior is expected, since when n and σ are small, and d is large, the Gaussian convolution in the SWD is unlikely to result in smoothing different samples together.

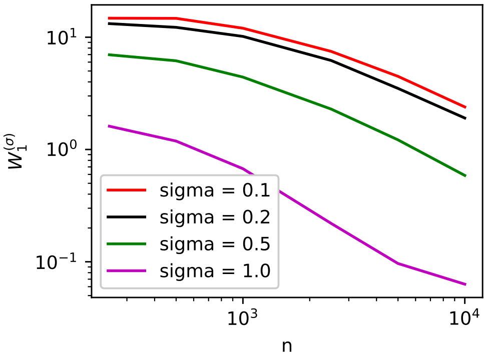


Figure 3: Convergence for fitting an NN generative model to a multivariate Gaussian mixture.

6 Summary and concluding remarks

We studied MDE with $\mathsf{W}_1^{(\sigma)}$ as the figure of merit. Measurability, strong consistency and limit distributions for MSWE in arbitrary dimensions were established. The characterization of high-dimensional distributional limits stands in sharp contrast with the classic W_1 MDE, where such a result is known only for $d = 1$ [37]. In particular, our results imply a uniform $n^{-1/2}$ convergence rate for MSWE for all d, highlighting the virtue of $\mathsf{W}_1^{(\sigma)}$ for high-dimensional generative modeling. Our ability to treat MSWE for arbitrary d relied on a novel limit distribution result for the empirical SWD. Under a polynomial moment condition on P we show that $\sqrt{n}\mathsf{W}_1^{(\sigma)}(P_n, P)$ converges in distribution to the supremum of a tight Gaussian process. This again contrasts the W_1 case, where a limit distribution is known only in $d = 1$. We have also established consistency of the bootstrap to enable evaluation of the limit distribution in practice.

This work focuses of statistical aspects of generative modeling with the SWE. A major goal going forward is to develop efficient algorithms for computing $\mathsf{W}_1^{(\sigma)}$, that are tailored to exploit the Gaussian convolution structure. We view the Monte Carlo algorithm employed herein merely as a placeholder. Gaussian smoothing significantly speeds up W_1 empirical convergence rates from $n^{-1/d}$ to $\sigma^{-d/2}n^{-1/2}$. While the latter is optimal in n, the exponential dependence of the prefactor on d calls for further exploration. We aim to relax this dependence under the manifold hypothesis, showing that the actual dependence is on the intrinsic dimension, and not the ambient one. Additional directions include an analysis for when $\sigma \downarrow 0$ is at a sufficiently slow rate (as a proxy for W_1). This is both theoretically challenging (calls for a finer analysis than the one presented herein) and practically relevant, as noise annealing is often used to stabilize training. SWDs of higher orders are also of interest.

Broader Impact

Our goal is to provide a stronger theoretical foundation for generative modeling based on the smoothed Wasserstein distance. We hope that this enables practitioners to build more robust, fair, and resilient generative models.

Funding Disclosure

The work of Z. Goldfeld was supported by the National Science Foundation Grant CCF-1947801 and the 2020 IBM Faculty Award. The work of K. Kato was supported by the National Science Foundation Grants DMS-1952306 and DMS-2014636.

References

[1] J. Wolfowitz. The minimum distance method. The Annals of Mathematical Statistics, pages 75–88, Mar. 1957.

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

[3] W. C. Parr and W. R. Schucany. Minimum distance and robust estimation. Journal of the American Statistical Association, 75(371):616–624, Sep. 1980.

[4] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In Advances in Neural Information Processing Systems (NeurIPS-2014), pages 2672–2680, 2014.

[5] Y. Li, K. Swersky, and R. Zemel. Generative moment matching networks. In Proceedings of the International Conference on Machine Learning (ICML-2015), pages 1718–1727, Lille, France, Jul. 2015.

[6] G. K. Dziugaite, D. M. Roy, and Z. Ghahramani. Training generative neural networks via maximum mean discrepancy optimization. In Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI-2015), pages 258–267, Amsterdam, Netherlands, Jul. 2015.

[7] Sebastian Nowozin, Botond Cseke, and Ryota Tomioka. f-GAN: Training generative neural samplers using variational divergence minimization. In Advances in Neural Information Processing Systems (NeurIPS-2016), pages 271–279, Barcelona, Spain, Dec. 2016.

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

[9] S. Arora, R. Ge, Y. Liang, T. Ma, and Y. Zhang. Generalization and equilibrium in generative adversarial nets (GANs). In Proceedings of the International Conference on Machine Learning (ICML-2017), pages 224–232, Sydney, Australia, Jul. 2017.

[10] Y. Mroueh and T. Sercu. Fisher GAN. In Proceedings of Advances in Neural Information Processing Systems (NeurIPS-2017), pages 2513–2523, Long Beach, California, US, Dec. 2017.

[11] Y. Mroueh, C.-L. Li, T. Sercu, A. Raj, and Y. Cheng. Sobolev GAN. In Proceedings of the International Conference on Learning Representations (ICLR-2018), Vancouver, Canada, Apr.-May 2018.

[12] D. P. Kingma and M. Welling. Auto-encoding variational bayes. In Proceedings of the International Conference on Learning Representations (ICLR-2014), Banff, Canada, Apr. 2014.

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

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

[15] B. Zhu, J. Jiao, and D. Tse. Deconstructing generative adversarial networks. IEEE Transactions on Information Theory, Mar. 2020.

[16] C. Villani. Optimal Transport: Old and New, volume 338. Springer Science & Business Media, 2008.

[17] I. Csiszár. Information-type measures of difference of probability distributions and indirect observation. Studia Ccientiarum Mathematicarum Hungarica, 2:229–318, 1967.

[18] A. Müller. Integral probability metrics and their generating classes of functions. Advances in Applied Probability, 29(2):429–443, 1997.

[19] I. O. Tolstikhin, B. K. Sriperumbudur, and B. Schölkopf. Minimax estimation of maximum mean discrepancy with radial kernels. In Proceedings of Advances in Neural Information Processing Systems (NeurIPS-2016), pages 1930–1938, Barcelona, Spain, Dec. 2016.

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

[21] A. B. Tsybakov. Introduction to nonparametric estimation. Springer Science & Business Media, 2008.

[22] X. Nguyen, M. J. Wainwright, and M. I. Jordan. Estimating divergence functionals and the likelihood ratio by convex risk minimization. IEEE Transactions on Information Theory, 56(11):5847–5861, Oct. 2010.

[23] B. K. Sriperumbudur, K. Fukumizu, A. Gretton, B. Schölkopf, and G. RG. Lanckriet. On integral probability metrics, ϕ-divergences and binary classification. arXiv preprint arXiv:0901.2698, 2009.

[24] S. Singh and B. Póczos. Minimax distribution estimation in Wasserstein distance. arXiv preprint arXiv:1802.08855, Feb. 2018.

[25] A. Krishnamurthy, K. Kandasamy, B. Póczos, and L. Wasserman. Nonparametric estimation of rényi divergence and friends. In Proceedings of the International Conference on Machine Learning (ICML-2014), pages 919–927, Beijing, China, Jun. 2014.

[26] T. Liang. Estimating certain integral probability metric (ipm) is as hard as estimating under the IPM. arXiv preprint arXiv:1911.00730, Nov. 2019.

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

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

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

[30] J. Adler and S. Lunz. Banach wasserstein GAN. In Advances in Neural Information Processing Systems (NeurIPS-2018), pages 6754–6763, Montréal, Canada, Dec. 2018.

[31] Z. Goldfeld, K. Greenewald, Y. Polyanskiy, and J. Weed. Convergence of smoothed empirical measures with applications to entropy estimation. Accepted to the IEEE Transactions on Information Theory, Feb. 2020.

[32] M. Ledoux and M. Talagrand. Probability in Banach Spaces: Isoperimetry and Processes. Springer. New York, 1991.

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

[34] E. Giné and R. Nickl. Mathematical Foundations of Infinite-Dimensional Statistical Models. Cambridge University Press, 2016.

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

[36] F. Bassetti, A. Bodini, and E. Regazzini. On minimum Kantorovich distance estimators. Statistics & Probability Letters, 76(12):1298–1302, 2006.

[37] E. Bernton, P. E. Jacob, M. Gerber, and C. P. Robert. On parameter estimation with the Wasserstein distance. Information and Inference: A Journal of the IMA, 2019.

[38] N. Belili, A. Bensaï, and H. Heinich. An estimate based Kantorovich functional and the Levy distance. Comptes Rendus de l'Academie des Sciences Series I Mathematics, 5(328):423–426, 1999.

[39] F. Bassetti and E. Regazzini. Asymptotic properties and robustness of minimum dissimilarity estimators of location-scale parameters. Theory of Probability & Its Applications, 50(2):171–186, 2006.

[40] K. Nadjahi, A. Durmus, U. Şimşekli, and R. Badeau. Asymptotic guarantees for learning generative models with the sliced-wasserstein distance. In Advances in Neural Information Processing Systems (NeurIPS-2019), Vancouver, Canada, Dec. 2019.

[41] N. Bonnotte. Unidimensional and evolution methods for optimal transportation. PhD thesis, Paris-Sud University, 2013.

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

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

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

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

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

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

[48] G. Mena and J. Niles-Weed. Statistical bounds for entropic optimal transport: sample complexity and the central limit theorem. In Proceedings of Advances in Neural Information Processing Systems (NeurIPS-2019), pages 4541–4551, Vancouver, Canada, Dec. 2019.

[49] E. del Barrio and J.-M. Loubes. Central limit theorems for empirical transportation cost in general dimension. Annals of Probability, 47:926–951, 2019.

[50] J. Feydy, T. Séjourné, F.-X. Vialard, S.-I. Amari, A. Trouvé, and G. Peyré. Interpolating between optimal transport and mmd using sinkhorn divergences. arXiv preprint arXiv:1810.08278, Oct. 2018.

[51] J. Bigot, E. Cazelles, and N. Papadakis. Central limit theorems for entropy-regularized optimal transport on finite spaces and statistical applications. arXiv preprint arXiv:1711.08947, 2019.

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

[53] F. Bolley, A. Guillin, and C. Villani. Quantitative concentration inequalities for empirical measures on non-compact spaces. Probab. Theory Related Fields, 137:541–593, 2007.

[54] E. Boissard. Simple bounds for the convergence of empirical and occupation measures in 1-Wasserstein distance. Electron. J. Probab., 16:2296–2333, 2011.

[55] S. Dereich, M. Scheutzow, and R. Schottstedt. Constructive quantization: Approximation by empirical measures. Ann. Inst. H. Poincaré Probab. Stat., 49(4):1183–1203, Nov. 2013.

[56] E. Boissard and T. Le Gouic. On the mean speed of convergence of empirical and occupation measures in Wasserstein distance. Ann. Inst. H. Poincaré Probab. Stat., 50(2):539–563, May 2014.

[57] J. Weed and F. Bach. Sharp asymptotic and finite-sample rates of convergence of empirical measures in Wasserstein distance. Bernoulli, 25(4A):2620–2648, Nov. 2019.

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

[59] E. Giné and J. Zinn. Empirical processes indexed by Lipschitz functions. Annals of Probability, 14:1329–1338, 1986.

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

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

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

[63] E. del Barrio, E. Giné, and F. Utzet. 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.

[64] A. var der Vaart. New Donsker classes. Annals of Probability, 24:2128–2124, 1996.

[65] A. Ramdas, N. G. Trillos, and M. Cuturi. On wasserstein two-sample testing and related families of nonparametric tests. Entropy, 19, 2017.

[66] L. D. Brown and R. Purves. Measurable selections of extrema. Annals of Statistics, 1(5):902–912, 1973.

[67] T. R. Rockafellar and R. J-B Wets. Variational Analysis, volume 317. Springer Science & Business Media, 2009.

[68] SS Vallender. Calculation of the wasserstein distance between probability distributions on the line. Theory of Probability & Its Applications, 18(4):784–786, 1974.

[69] Y. A. Davydov, M. A. Lifschits, and N. V. Smorodina. Local Properties of Distributions of Stochastic Functionals. Translation of Mathematical Monographs. American Mathematical Society, 1998.

[70] A. W. van der Vaart. Asymptotic Statistics. Cambridge University Press, Cambridge, UK, 1998.

[71] R. Adamczkak. A few remarks on the operator norm of random Toeplitz matrices. Journal of Theoretical Probability, 23:85–108, 2010.

[72] R. Adamczkak. A tail inequality for suprema of unbounded empirical processes with applications to Markov chains. Electronic Journal of Probability, 34:1000–1034, 2008.

[73] F. Santambrogio. Optimal Transport for Applied Mathematicians. Birkhäuser, 2015.

[74] C. D. Aliprantis and K. C. Border. Infinite Dimensional Analysis: A Hitchhiker's Guide. Springer, 2006.

[75] K. Kato. Asymptotics for argmin processes: Convexity arguments. Journal of Multivariate Analysis, 100:1816–1829, 2009.

[76] D. Pollard. Asymptotics for least absolute deviation regression estimators. Econometric Theory, 7:186–199, 1991.

[77] N. I. Hjort and D. Pollard. Asymptotics for minimizers of convex processes. Unpublished manuscript, 1993.