On Projection Robust Optimal Transport: Sample Complexity and Model Misspecification

Tianyi Lin[⋆] **Zeyu Zheng**[⋆] **Elynn Y. Chen**[⋆] **Marco Cuturi**[◁,▷] **Michael I. Jordan**[⋆]
University of California, Berkeley[⋆]
CREST - ENSAE[◁], Google Brain[▷]

Abstract

Optimal transport (OT) distances are increasingly used as loss functions for statistical inference, notably in the learning of generative models or supervised learning. Yet, the behavior of minimum Wasserstein estimators is poorly understood, notably in high-dimensional regimes or under model misspecification. In this work we adopt the viewpoint of projection robust (PR) OT, which seeks to maximize the OT cost between two measures by choosing a k-dimensional subspace onto which they can be projected. Our first contribution is to establish several fundamental statistical properties of PR Wasserstein distances, complementing and improving previous literature that has been restricted to one-dimensional and well-specified cases. Next, we propose the integral PR Wasserstein (IPRW) distance as an alternative to the PRW distance, by averaging rather than optimizing on subspaces. Our complexity bounds can help explain why both PRW and IPRW distances outperform Wasserstein distances empirically in high-dimensional inference tasks. Finally, we consider parametric inference using the PRW distance. We provide an asymptotic guarantee of two types of minimum PRW estimators and formulate a central limit theorem for max-sliced Wasserstein estimator under model misspecification. To enable our analysis on PRW with projection dimension larger than one, we devise a novel combination of variational analysis and statistical theory.

Proceedings of the 24[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130.

1 Introduction

Recent years have witnessed an ever-increasing role for ideas from optimal transport (OT) (Villani, 2008) in machine learning. Combining OT distances with the general principles of minimal distance estimation (MDE) (Wolfowitz, 1957; Basu et al., 2011) yields a powerful basis for various statistical inference problems, such as density estimation Bassetti et al. (2006), training of generative model (Arjovsky et al., 2017; Gulrajani et al., 2017; Montavon et al., 2016; Adler and Lunz, 2018; Cao et al., 2019), auto-encoders (Tolstikhin et al., 2018), clustering (Cuturi and Doucet, 2014; Bonneel et al., 2016; Ho et al., 2017; Ye et al., 2017), multitask regression (Janati et al., 2020), trajectory inference (Hashimoto et al., 2016; Schiebinger et al., 2017; Yang et al., 2020; Tong et al., 2020) or nonparametric testing (Ramdas et al., 2017); see Peyré and Cuturi (2019) and Panaretos and Zemel (2019) for reviews on these topics.

For OT ideas to continue to bear fruit in machine learning, it will be necessary to tackle two characteristic challenges: (1) high dimensionality and (2) model misspecification. Initial progress has been made on the latter problem by Bernton et al. (2019), who showed that in the misspecified case the minimum Wasserstein estimator (MWE) outputs the Wasserstein projection of the data-generating distribution onto the fitted model class. These authors also obtained results on robustness and the asymptotic distribution of the projection, while these results only apply to the one-dimensional setting. High-dimensional settings are challenging; indeed, it is known that the sample complexity of estimating the Wasserstein distance can grow exponentially in dimension (Dudley, 1969; Fournier and Guillin, 2015; Singh and Póczos, 2018; Weed and Bach, 2019; Lei, 2020).

We focus on a promising approach to treating high-dimensional problems: Compute the OT distance between low-dimensional projections of high-dimensional input measures. The simplest and most representa-

tive example of this approach is the sliced Wasserstein distance (Rabin et al., 2011; Bonnotte, 2013; Bonneel et al., 2015; Deshpande et al., 2019; Kolouri et al., 2019a; Nadjahi et al., 2020; Manole et al., 2019), which is defined as the average OT distance obtained between random 1-dimensional projections, and which is shown practical in real applications (Deshpande et al., 2018, 2019; Kolouri et al., 2016, 2019b; Carriere et al., 2017; Wu et al., 2019; Liutkus et al., 2019). In an important extension, Paty and Cuturi (2019) and Niles-Weed and Rigollet (2019) proposed very recently to seek the k-dimensional subspace ($k > 1$) that would maximize the OT distance between two measures after projection. The quantity is named as *projection robust Wasserstein* (PRW) distance[1], which is conceptually simple and does solve the curse of dimensionality in the so-called spiked model as proved in (Niles-Weed and Rigollet, 2019, Theorem 1) by recovering the $n^{-1/k}$ rate under the Talagrand transport inequality. This result suggests that PRW can be significantly more useful than the OT distance for inference tasks when the dimension is large. From a computational point of view, PRW becomes the max-sliced Wasserstein distance when the projection dimension is $k = 1$ and has an efficient implementation (Deshpande et al., 2019). For general $k \geq 1$, Lin et al. (2020) proposed to compute PRW using Riemannian optimization toolbox and provided theoretical guarantee and encouraging empirical results. However, it is desirable to understand its statistical behavior which mostly determines the practical performance of PRW.

Contributions. In this paper, we study the statistical properties of PRW and another so-called integrated PRW (IPRW), which replaces the maximum in the original PRW with an average of OT distance over k-dimensional projections. Our contributions can be summarized as follows.

1. We prove that the empirical measure $\widehat{\mu}_n$ converges to true measure $\mu_\star$ under both PRW and IPRW with different rates. These rates are new to our knowledge. For example, when the order $p = 3/2$ and the projected dimension $k \geq 3$, the rate is $n^{-1/k}$ for IPRW. For PRW, the rate is $(n^{-1/k}+n^{-1/6}\sqrt{dk\log(n)}+n^{-2/3}dk\log(n))$ when $\mu_\star$ satisfies a projection Bernstein tail condition and $(n^{-1/k} + n^{-1/2}\sqrt{dk\log(n)} + n^{-2/3}dk\log(n))$ when $\mu_\star$ satisfies a projection Poincaré inequality.

2. We derive the concentration results when $\mu_\star$ satisfies a Bernstein tail condition or a projection Poincaré inequality. In terms of tail conditions, our Bernstein condition and Poincare inequality handle subexponential tail while Talagrand inequality in Niles-Weed and Rigollet (2019) addresses subgaussian tail. Our assumptions are thus weaker than Niles-Weed and Rigollet (2019).

3. We establish asymptotic guarantees for the minimal PRW and expected minimal PRW estimators under model misspecification. For minimal PRW estimator with the order $p = 1$ and the projected dimension $k = 1$, we derive an asymptotic distribution for arbitrary dimension d with the $n^{-1/2}$ rate in the Hausdorff metric. Our assumptions are weaker than those used in Bernton et al. (2019), not requiring the nonsingularity of the Jacobian or the separability of the parameters. Our techniques for CLT in misspecified settings did not appear in Nadjahi et al. (2019) and complete the analysis in Bernton et al. (2019).

4. We conduct experiments on synthetic data and neural networks to validate our theory. We also present a simple optimization algorithm that can efficiently compute the PRW distance in practice even when $k \geq 2$; see Appendix F or Appendix B of the concurrent work byLin et al. (2020).

2 Preliminaries on Projected Optimal Transport

In this section, we provide some technical background materials on projection optimal transport. Throughout the paper, we denote $\|\cdot\|$ as the Euclidean norm (in the corresponding vector space) and $\Rightarrow$ as the convergence in the weak sense.

Wasserstein and sliced Wasserstein. Let $p \geq 1$ and define $\mathscr{P}(\mathbb{R}^d)$ and $\mathscr{P}_p(\mathbb{R}^d)$ as the set of all Borel measures on $\mathbb{R}^d$ and the subset that satisfies $M_p(\mu) := \int_{\mathbb{R}^d} \|x\|^p d\mu(x) < +\infty$. For two probability measures $\mu, \nu \in \mathscr{P}_p(\mathbb{R}^d)$, their Wasserstein distance of order p is defined as follows:

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

where the infimum is taken over $\Pi(\mu,\nu) \subseteq \mathscr{P}(\mathbb{R}^d \times \mathbb{R}^d)$—the set of probability measures with marginals μ and ν. In the 1D case, Rachev and Rüschendorf (1998, Theorem 3.1.2.(a)) have shown that $\mathcal{W}_p(\mu,\nu) = (\int_0^1 |F_\mu^{-1}(t) - F_\nu^{-1}(t)|^p dt)^{1/p}$, where F_μ^{-1} and F_ν^{-1} are the quantile functions of μ and ν. This 1D formula motivates the *sliced Wasserstein (SW) and max-sliced Wasserstein (max-SW)* distances (Bonnotte, 2013; Bonneel et al., 2015; Deshpande et al., 2019). In particular, the idea is to use as a proxy of (1) the aver-

[1]This quantity is also named as *Wasserstein Projection Pursuit* (WPP) (Niles-Weed and Rigollet, 2019). For simplicity, we refer from now on to PRW/WPP as PRW.

age or maximum of a set of 1D Wasserstein distances constructed by projecting d-dimensional measures to a random collection of 1D spaces. Computationally appealing, both SW and max-SW distances are widely used in practice, especially in generative modeling (Kolouri et al., 2019b; Deshpande et al., 2019; Liutkus et al., 2019). Practitioners observe that the SW distance only outputs a good Monte-Carlo approximation with a large number of projections, while the max-SW distance achieves similar results with fewer projections (Kolouri et al., 2019a; Nguyen et al., 2020).

Encouraged by the success of SW and max-SW, Paty and Cuturi (2019) asked whether we can gain more by using a subspace of dimension $k \geq 2$, define the *projection robust Wasserstein* (PRW) distance, and prove that this quantity is well posed if the order is $p \geq 1$. More specifically, let $\mathbb{S}_{d,k} = \{E \in \mathbb{R}^{d\times k} : E^\top E = I_k\}$ be the set of $d \times k$ orthogonal matrices and $E^\star$ be the linear transformation associated with E for any $x \in \mathbb{R}^d$ by $E^\star(x) = E^\top x$. For any measurable function f and $\mu \in \mathscr{P}(\mathbb{R}^d)$, we denote $f_\#\mu$ as the push-forward of μ by f, so that $f_\#\mu(A) = \mu(f^{-1}(A))$ where $f^{-1}(A) = \{x \in \mathbb{R}^d : f(x) \in A\}$ for any Borel set A. For any given subspace dimension K, the PRW distance of order p between μ and ν is defined by

$$\overline{\mathcal{PW}}_{p,k}(\mu,\nu) := \sup_{E\in\mathbb{S}_{d,k}} \mathcal{W}_p(E^\star_\#\mu, E^\star_\#\nu). \tag{2}$$

The PRW distance has better discriminative power than the SW or max-SW distances since it can extract more geometric information from high-dimensional projections than that from 1-dimensional projections; see Paty and Cuturi (2019) for more details.

As an alternative, we define the IPRW distance, which replaces the supremum in Eq. (2) with an average. The IPRW distance of order p between μ and ν is

$$\underline{\mathcal{PW}}_{p,k}(\mu,\nu) := \left(\int_{\mathbb{S}_{d,k}} \mathcal{W}_p^p(E^\star_\#\mu, E^\star_\#\nu)d\sigma(E)\right)^{1/p}, \tag{3}$$

where σ is the uniform distribution on $\mathbb{S}_{d,k}$. Note that IPRW is well defined for comparing two measures and match our intuition. For example, given three Gaussian distributions $\mu_i = \mathcal{N}(u_i, I_d)$ for $i = 1, 2, 3$, we have $\underline{\mathcal{PW}}_{p,2}(\mu_i, \mu_j) = c\|u_i - u_j\|$ where $c > 0$ only depend on p and the dimension d.

The IPRW and PRW distances generalize the SW and max-SW distances to the high-dimensional projection setting. Both PRW and and IPRW are distances and satisfy the triangle inequality: the proof for PRW is in Paty and Cuturi (2019, Proposition 1), while that for IPRW is the same as that for SW in Bonnotte (2013). Compared to the PRW distance, the IPRW distance performs better statistically but remains unfavorable in computational sense. Indeed, a large amount of projections from $\mathbb{S}_{d,k}$ are necessary to approximate the IPRW distance. However, if the intrinsic dimension of data distribution is small, the required number of random projections is small; see Nadjahi et al. (2019).

Let $X_{1:n} = (X_1, \ldots, X_n)$ be independent and identically distributed samples according to the true measure $\mu_\star \in \mathscr{P}_q(\mathbb{R}^d)$. The empirical measure of $X_{1:n}$ is defined by $\widehat{\mu}_n := (1/n)\sum_{i=1}^n \delta_{X_i}$. It is known that $\widehat{\mu}_n \Rightarrow \mu_\star$ almost surely, and $\mathcal{W}_p(\widehat{\mu}_n, \mu_\star) \to 0$ almost surely since Wasserstein distances metrizes weak convergence (Villani, 2008, Theorem 6.9) (note that $q \geq p \geq 1$). However, $\mathbb{E}[\mathcal{W}_p(\widehat{\mu}_n, \mu_\star)] \simeq n^{-1/d}$ whenever μ is absolutely continuous with respect to Lebesgue measure and $d > 2p$ (Dudley, 1969; Fournier and Guillin, 2015; Weed and Bach, 2019) ($\simeq$ means "equal to" with a constant independent of n). The convergence is slow when the dimension is high — an instance of the well-known curse-of-dimensionality phenomenon.

Due to the low-dimensional structure of the IPRW and PRW distances, the rate of IPRW and PRW distances is expected to be of $n^{-1/k}$ in the large-n limit. Similar rates have been derived for $\mathbb{E}[|\overline{\mathcal{PW}}_{k,p}(\widehat{\mu}_n, \widehat{\nu}_n) - \mathcal{W}_p(\mu,\nu)|]$ as a function of n under a spiked transport model for both μ and ν; see Niles-Weed and Rigollet (2019, Theorem 8). Their bound depends on problem dimension d and requires μ and ν to satisfy the Talagrand transport inequality (Talagrand, 1996). For the special case when $k = 1$, the rate for the IPRW distance was studied in (Nadjahi et al., 2020) and the minimax confidence intervals were established in Manole et al. (2019). To our knowledge, there has been no other paper on the statistical properties of IPRW and PRW distances for $k \geq 2$.

Parametric modeling and inference. A statistical model is a family of distributions, $\mathcal{M} = \{\mu_\theta \in \mathscr{P}(\mathbb{R}^d) \mid \theta \in \Theta\}$, where Θ is the parameter space. A minimal set of the conditions of a proper family of distribution are: (i) $(\Theta, \|\cdot\|_\Theta)$ is a Polish space, (ii) Θ is σ-compact, i.e., it is the union of countably many compact subspaces, and (iii) parameters are identifiable, i.e., $\mu_\theta = \mu_{\theta'}$ implies $\theta = \theta'$ for all $\theta, \theta' \in \Theta$. Since the space $\mathscr{P}_p(\mathbb{R}^d)$ endowed with the distance $\mathcal{W}_p$ is a Polish space, we estimate model coefficients using minimum distance estimation (MDE) (Wolfowitz, 1957; Basu et al., 2011), where the distance we consider here is PRW. The main reason why we do not choose IPRW in this setting is computational. The *minimum project robust Wasserstein* (MPRW) estimator is defined as follows:

$$\widehat{\theta}_n := \operatorname*{argmin}_{\theta\in\Theta} \overline{\mathcal{PW}}_{p,k}(\widehat{\mu}_n, \mu_\theta). \tag{4}$$

Note that the probability density function of μ_θ can be difficult to evaluate in practice, especially when μ_θ is a generative model. Nevertheless, in various settings, even if the density is not available, one can generate samples $Z_{1:m}$ from μ_θ and use them to approximate μ_θ. With this approximation, a natural alternative is the *minimum expected projection robust Wasserstein* (MEPRW) estimator, which is defined as follows (Bernton et al., 2019; Nadjahi et al., 2019):

$$\widehat{\theta}_{n,m} := \operatorname*{argmin}_{\theta\in\Theta} \mathbb{E}[\overline{\mathcal{PW}}_{p,k}(\widehat{\mu}_n, \widehat{\mu}_{\theta,m}) \mid X_{1:n}], \qquad (5)$$

where n is the number of samples from the data distribution $\mu_\star$, m is the number of samples from the parametric distribution μ_θ, and $\widehat{\mu}_{\theta,m}$ is an empirical version of μ_θ based on samples $Z_{1:m}$.

Existing works have established asymptotic guarantees for minimal Wasserstein and sliced Wasserstein estimators (Bernton et al., 2019; Nadjahi et al., 2019). Despite the similar proof paths, our results for the MPRW and MEPRW estimators are new and derived under weaker assumptions and more general settings than previous work; see Sections 3.3 and 3.4.

3 Main Results on Projection Robust Optimal Transport Estimation

Throughout this section, we assume $p \geq 1$ and $k \in [d] \triangleq \{1, 2, \ldots, d\}$ unless stated otherwise. Focusing on the IPRW and PRW distances, we prove that they are lower semi-continuous and metrize weak convergence. Through a new sample complexity analysis, we derive the convergence rate of empirical measures under both distances as well as an improved rate for the PRW distance when $\mu_\star$ satisfies either a Bernstein tail condition or the Poincaré inequality. For the generative models with the PRW distance, we study the misspecified setting where the limit $\theta_\star$ is not necessarily the limit of the maximum likelihood estimator. We establish the asymptotic properties of the MPRW and MEPRW estimators and formulate a central limit theorem when $p = 1$ and $k = 1$.

3.1 Topological properties

We begin with the results on the relationship between the IPRW, PRW and Wasserstein distances. The following lemma demonstrates their equivalence in a topological sense.

Lemma 3.1 *The IPRW, PRW and Wasserstein distances are equivalent. That is, for any sequence of probability measures $\{\mu_i\}_{i\in\mathbb{N}}$ and probability measure μ in $\mathscr{P}_p(\mathbb{R}^d)$, we have $\underline{\mathcal{PW}}_{p,k}(\mu_i, \mu) \to 0$ if and only if $\overline{\mathcal{PW}}_{p,k}(\mu_i, \mu) \to 0$ if and only if $\mathcal{W}_p(\mu_i, \mu) \to 0$.*

Lemma 3.1 is a generalization of Bayraktar and Guo (2019, Theorem 1) where the projection dimension is $k = 1$. By Lemma 3.1 and Villani (2008, Theorem 6.9), we obtain the following result regarding the topology induced by the IPRW and PRW distances of order p.

Theorem 3.2 *The IPRW and PRW distances both metrize weak convergence. In other words, for any sequence of probability measures $\{\mu_i\}_{i\in\mathbb{N}}$ and probability measure μ in $\mathscr{P}_p(\mathbb{R}^d)$, we have $\underline{\mathcal{PW}}_{p,k}(\mu_i, \mu) \to 0$ if and only if $\overline{\mathcal{PW}}_{p,k}(\mu_i, \mu) \to 0$ if and only if $\mu_i \Rightarrow \mu$.*

Theorem 3.2 generalizes Villani (2008, Theorem 6.9) since the PRW distance is the Wasserstein distance when the projection dimension $k = d$. When $k = 1$, Theorem 3.2 recovers the results presented by Bayraktar and Guo (2019) which implies that the SW and max-SW distances metrize weak convergence. It is worthy noting that this implication is stronger than Nadjahi et al. (2019, Theorem 1), which only provides a one-sided argument.

Theorem 3.3 *The IPRW and PRW distances are both lower semi-continuous in the usual weak topology. In other words, if the sequences of probability measures $\{\mu_i\}_{i\in\mathbb{N}}, \{\nu_i\}_{i\in\mathbb{N}} \subseteq \mathscr{P}(\mathbb{R}^d)$ satisfy $\mu_i \Rightarrow \mu$ and $\nu_i \Rightarrow \nu$ for probability measures $\mu, \nu \in \mathscr{P}(\mathbb{R}^d)$, then we have $\underline{\mathcal{PW}}_{p,k}(\mu, \nu) \leq \liminf_{i\to+\infty} \underline{\mathcal{PW}}_{p,k}(\mu_i, \nu_i)$ and $\overline{\mathcal{PW}}_{p,k}(\mu, \nu) \leq \liminf_{i\to+\infty} \overline{\mathcal{PW}}_{p,k}(\mu_i, \nu_i)$.*

The above theorem generalizes Nadjahi et al. (2019, Lemma S6) and is pivotal to our asymptotic analysis for the MPRW and MEPRW estimators.

3.2 Convergence and concentration of empirical measures

We provide the rate of the empirical measures under the IPRW and PRW distances of order p with the projection dimension k. We present our main result on convergence rates in the following theorem.

Theorem 3.4 *Let $\mu_\star \in \mathscr{P}_q(\mathbb{R}^d)$ and $M_q(\mu_\star) < +\infty$ for some $q \geq p \geq 1$. Then we have*[2]

$$\mathbb{E}[\underline{\mathcal{PW}}_{p,k}(\widehat{\mu}_n, \mu_\star)] \lesssim_{p,q} n^{-[\frac{1}{(2p)\vee k} \wedge (\frac{1}{p} - \frac{1}{q})]} (\log(n))^{\frac{\zeta_{p,q,k}}{p}},$$

where $\lesssim_{p,q}$ refers to "less than" with a constant depending only on (p, q) and

$$\zeta_{p,q,k} = \begin{cases} 2 & \text{if } k = q = 2p, \\ 1 & \text{if } (k \neq 2p \text{ and } q = \frac{kp}{k-p}) \text{ or } (q > k = 2p), \\ 0 & \text{otherwise.} \end{cases}$$

Remark 3.1 *Theorem 3.4 shows that our bound does not depend on d, while all bounds for the Wasserstein*

[2] $a \vee b = \max\{a, b\}$ and $a \wedge b = \min\{a, b\}$ here.

distance grow exponentially in d when $d \geq 2p$ (Lei, 2020, Theorem 3.1). This improvement shows that the PRW distance does not suffer from the curse of dimensionality while retaining flexibility via the choice of k. We are also aware of concurrent work (Nath and Jawanpuria, 2020) in which the sample complexity has no dependence on dimensionality.

Definition 3.1 *We say $\mu \in \mathscr{P}(\mathbb{R}^d)$ satisfies a* projection Bernstein tail condition *if there exist $\sigma, V > 0$ for all $E \in \mathbb{S}_{d,k}$ and $X \sim E^\star_\# \mu$ such that $\mathbb{E}[\|X\|^r] \leq (1/2)\sigma^2 r! V^{r-2}$ for all $r \geq 2$.*

Theorem 3.5 *Suppose $\mu_\star \in \mathscr{P}_q(\mathbb{R}^d)$ satisfies a projection Bernstein tail condition and assume the same setting as in Theorem 3.4. For all $n \geq 1$, the following inequality holds true:*

$$\mathbb{E}[\overline{\mathcal{PW}}_{p,k}(\widehat{\mu}_n, \mu_\star)] \lesssim_{p,q} n^{-[\frac{1}{(2p)\vee k} \wedge (\frac{1}{p} - \frac{1}{q})]} (\log(n))^{\frac{\zeta_{p,q,k}}{p}} + n^{\frac{1}{2} - \frac{1}{p}} \sqrt{dk \log(n)} + n^{-\frac{1}{p}} dk \log(n).$$

Definition 3.2 *We say $\mu \in \mathscr{P}(\mathbb{R}^d)$ satisfies a* projection Poincaré inequality *if there exists $M > 0$ for all $E \in \mathbb{S}_{d,k}$ and $X \sim E^\star_\# \mu$ such that $\mathrm{Var}(f(X)) \leq M\mathbb{E}[\|\nabla f(X)\|^2]$ for any $f : \mathbb{R}^d \to \mathbb{R}$ satisfying that $\mathbb{E}[f(X)^2] < +\infty$ and $\mathbb{E}[\|\nabla f(X)\|^2] < +\infty$.*

Theorem 3.6 *Suppose $\mu_\star \in \mathscr{P}_q(\mathbb{R}^d)$ satisfies a projection Poincaré inequality and assume the same setting as in Theorem 3.4. For all $n \geq 1$, the following inequality holds true:*

$$\mathbb{E}[\overline{\mathcal{PW}}_{p,k}(\widehat{\mu}_n, \mu_\star)] \lesssim_{p,q} n^{-[\frac{1}{(2p)\vee k} \wedge (\frac{1}{p} - \frac{1}{q})]} (\log(n))^{\frac{\zeta_{p,q,k}}{p}} + n^{-\frac{1}{2\vee p}} \sqrt{dk \log(n)} + n^{-\frac{1}{p}} dk \log(n).$$

We present concentration results when $\mu_\star$ satisfies stronger conditions than Definition 3.1 and 3.2.

Definition 3.3 *A measure $\mu \in \mathscr{P}(\mathbb{R}^d)$ satisfies a* Bernstein tail condition *if there exists $\sigma, V > 0$ such that $\mathbb{E}_{X\sim\mu}[\sup_{E\in\mathbb{S}_{d,k}} \|E^\top X\|^r] \leq (1/2)\sigma^2 r! V^{r-2}$ for all $i = 1, 2, \ldots, n$ and all $r \geq 2$.*

Theorem 3.7 *If $\mu_\star \in \mathscr{P}(\mathbb{R}^d)$ satisfies a Bernstein tail condition then the following statement holds true for both $W = \underline{\mathcal{PW}}_{p,k}$ and $W = \overline{\mathcal{PW}}_{p,k}$:*

$$\mathbb{P}(|W(\widehat{\mu}_n, \mu_\star) - \mathbb{E}[W(\widehat{\mu}_n, \mu_\star)]| \geq t) \leq 2\exp\left(-\frac{t^2}{8\sigma^2 n^{1-2/p} + 4tVn^{-1/p}}\right).$$

Definition 3.4 *$\mu \in \mathscr{P}(\mathbb{R}^d)$ satisfies a* Poincaré inequality *if there exists $M > 0$ for $X \sim \mu$ such that $\mathrm{Var}[f(X)] \leq M\mathbb{E}[\|\nabla f(X)\|^2]$ for any f satisfying $\mathbb{E}[f(X)^2] < +\infty$ and $\mathbb{E}[\|\nabla f(X)\|^2] < +\infty$.*

Theorem 3.8 *If $\mu_\star \in \mathscr{P}(\mathbb{R}^d)$ satisfies Poincaré inequality then the following statement holds true for both $W = \underline{\mathcal{PW}}_{p,k}$ and $W = \overline{\mathcal{PW}}_{p,k}$:*

$$\mathbb{P}(|W(\widehat{\mu}_n, \mu_\star) - \mathbb{E}[W(\widehat{\mu}_n, \mu_\star)]| \geq t) \leq 2\exp(-K^{-1}\min\{n^{\frac{1}{p}}t,\ n^{\frac{2}{2\vee p}}t^2\}),$$

where $K > 0$ only depends on M (cf. Definition 3.4).

Discussions. We demonstrate that the Bernstein-type tail conditions in Definition 3.1 and 3.3 are not strong enough to give an effective bound for all $p \geq 1$. The similar results for the Wasserstein distance have been recently derived by Lei (2020) and recognized as the standard limitation for the Bernstein-type tail conditions. This is also the motivation which drives us to consider a Poincaré inequality.

For Theorem 3.5 and 3.6, the first term matches that in Theorem 3.4 while the extra two terms come from bounding the gap $\mathbb{E}[\sup_{E\in\mathbb{S}_{d,k}}(\mathcal{W}_p(E^\star_\#\widehat{\mu}_n, E^\star_\#\mu_\star) - \mathbb{E}[\mathcal{W}_p(E^\star_\#\widehat{\mu}_n, E^\star_\#\mu_\star)])]$. Compared with Niles-Weed and Rigollet (2019, Theorem 8), where $\mu_\star$ satisfies the Talagrand transport inequality, our conditions are weaker but our rate matches their $n^{-1/k} + n^{-1/2}\sqrt{dk\log(n)}$ rate in the large-n limit when $p = 1$.

For Theorem 3.7 and 3.8, the latter bound is better than the former bound when $p > 1$. Moreover, the tail condition in Definition 3.3 is stronger than that in Definition 3.1 yet weaker than the standard Bernstein tail condition where $X \sim \mu$ inside the expectation without a sup; see Wainwright (2019). The Poincaré inequality is weaker than the log-Sobolev inequality and is satisfied by various exponential measures and the measures induced by Markov processes (Ledoux, 1999). Intutively, These two conditions handle subexponential tail while Talagrand inequality in Niles-Weed and Rigollet (2019) addresses subgaussian tail; see Ledoux (1999) and Talagrand (1996) for the details.

3.3 Properties of MPRW and MEPRW estimators

We derive the asymptotic properties of the MPRW and MEPRW estimators under model misspecification, which is common in practice. Our setting is more general than that considered in (Nadjahi et al., 2019) and our results support the applications in real-world scenario better. Specifically, while Nadjahi et al. (2019) focused on the well-specified setting, the statistical models can be misspecified in many real-world applications. We also use the Wasserstein distance in Assumptions 3.1 and 3.4 since these assumptions have been shown valid for many real-world application problems (Bernton et al., 2019).

Assumption 3.1 *There exists a probability measure $\mu_\star \in \mathscr{P}(\mathbb{R}^d)$ such that the data-generating process satisfies that $\lim_{n\to+\infty} \mathcal{W}_p(\widehat{\mu}_n, \mu_\star) = 0$ almost surely.*

Assumption 3.2 *The map $\theta \mapsto \mu_\theta$ is continuous: $\|\theta_n - \theta\|_\Theta \to 0$ implies $\mu_{\theta_n} \Rightarrow \mu_\theta$.*

Assumption 3.3 *There exists a constant $\tau > 0$ such that the set $\Theta_\star(\tau) \subseteq \Theta$ is bounded where $\Theta_\star(\tau) = \{\theta \in \Theta : \overline{\mathcal{PW}}_{p,k}(\mu_\star, \mu_\theta) \leq \inf_{\theta\in\Theta} \overline{\mathcal{PW}}_{p,k}(\mu_\star, \mu_\theta) + \tau\}$.*

Theorem 3.9 *Under Assumption 3.1-3.3, there exists a sample space Ω with $\mathbb{P}(\Omega) = 1$ such that, for all $\omega \in \Omega$, $\lim_{n\to+\infty} \inf_{\theta\in\Theta} \overline{\mathcal{PW}}_{p,k}(\widehat{\mu}_n(\omega), \mu_\theta) = \inf_{\theta\in\Theta} \overline{\mathcal{PW}}_{p,k}(\mu_\star, \mu_\theta)$, and $\limsup_{n\to+\infty} \operatorname{argmin}_{\theta\in\Theta} \overline{\mathcal{PW}}_{p,k}(\widehat{\mu}_n(\omega), \mu_\theta) \subseteq \operatorname{argmin}_{\theta\in\Theta} \overline{\mathcal{PW}}_{p,k}(\mu_\star, \mu_\theta)$. In addition, $\operatorname{argmin}_{\theta\in\Theta} \overline{\mathcal{PW}}_{p,k}(\widehat{\mu}_n(\omega), \mu_\theta) \neq \emptyset$ for all $n \geq n(\omega)$ with some $n(\omega) > 0$.*

Assumption 3.4 *If $\|\theta_n - \theta\|_\Theta \to 0$, then $\mathbb{E}[\mathcal{W}_p(\widehat{\mu}_{\theta_n,n}, \mu_{\theta_n})|X_{1:n}] \to 0$.*

In the next result, we present an analogous version of Theorem 3.9 for the MEPRW estimator as $\min\{n, m\} \to +\infty$. For the simplicity, we set $m := m(n)$ such that $m(n) \to +\infty$ as $n \to +\infty$.

Theorem 3.10 *Under Assumption 3.1-3.4, there exists a sample space Ω with $\mathbb{P}(\Omega) = 1$ such that, for all $\omega \in \Omega$, $\lim_{n\to+\infty} \inf_{\theta\in\Theta} \mathbb{E}[\overline{\mathcal{PW}}_{p,k}(\widehat{\mu}_n(\omega), \widehat{\mu}_{\theta,m(n)})|X_{1:n}] = \inf_{\theta\in\Theta} \overline{\mathcal{PW}}_{p,k}(\mu_\star, \mu_\theta)$ and $\limsup_{n\to+\infty} \operatorname{argmin}_{\theta\in\Theta} \mathbb{E}[\overline{\mathcal{PW}}_{p,k}(\widehat{\mu}_n(\omega), \widehat{\mu}_{\theta,m(n)}) | X_{1:n}] \subseteq \operatorname{argmin}_{\theta\in\Theta} \overline{\mathcal{PW}}_{p,k}(\mu_\star, \mu_\theta)$. In addition, $\operatorname{argmin}_{\theta\in\Theta} \mathbb{E}[\overline{\mathcal{PW}}_{p,k}(\widehat{\mu}_n(\omega), \widehat{\mu}_{\theta,m(n)})|X_{1:n}] \neq \emptyset$ for $n \geq n(\omega)$ with some $n(\omega) > 0$.*

Assumption 3.5 *There exists a constant $\tau > 0$ such that the set $\Theta_n(\tau) \subseteq \Theta$ is bounded where $\Theta_n(\tau) = \{\theta \in \Theta : \overline{\mathcal{PW}}_{p,k}(\widehat{\mu}_n, \mu_\theta) \leq \inf_{\theta\in\Theta} \overline{\mathcal{PW}}_{p,k}(\widehat{\mu}_n, \mu_\theta) + \tau\}$.*

Theorem 3.11 *Under Assumption 3.2, 3.4 and 3.5, it holds that $\lim_{m\to+\infty} \inf_{\theta\in\Theta} \mathbb{E}[\overline{\mathcal{PW}}_{p,k}(\widehat{\mu}_n, \widehat{\mu}_{\theta,m}) | X_{1:n}] = \inf_{\theta\in\Theta} \overline{\mathcal{PW}}_{p,k}(\widehat{\mu}_n, \mu_\theta)$ and $\limsup_{m\to+\infty} \operatorname{argmin}_{\theta\in\Theta} \mathbb{E}[\overline{\mathcal{PW}}_{p,k}(\widehat{\mu}_n, \widehat{\mu}_{\theta,m}) | X_{1:n}] \subseteq \operatorname{argmin}_{\theta\in\Theta} \overline{\mathcal{PW}}_{p,k}(\widehat{\mu}_n, \mu_\theta)$. In addition, $\operatorname{argmin}_{\theta\in\Theta} \mathbb{E}[\overline{\mathcal{PW}}_{p,k}(\widehat{\mu}_n, \widehat{\mu}_{\theta,m})|X_{1:n}] \neq \emptyset$ for $m \geq m_n$ with some $m_n > 0$.*

To this end, the MPRW and MEPRW estimators both asymptotically converge to $\theta_\star \in \Theta$, which is a minimizer of $\theta \to \overline{\mathcal{PW}}_{p,k}(\mu_\star, \mu_\theta)$, assuming its existence. Moreover, $\theta_\star$ is not the limit of maximum likelihood estimator and satisfies $\mu_{\theta_\star} = \mu_\star$ in a well-specified setting. Our consistency results support the success of generative modelling using the max-SW distance.

3.4 Rate of convergence and asymptotic distribution

We investigate the asymptotic distribution of the MPRW estimator under model misspecification and establish the rate of convergence when $k = p = 1$. For any $u \in \mathbb{S}^{d-1}$ and $t \in \mathbb{R}$, we define

$$F_\theta(u,t) = \int_{\mathbb{R}^d} \mathbf{1}_{(-\infty,t]}(\langle u, x\rangle)\, d\mu_\theta(x),$$
$$\widehat{F}_n(u,t) = (1/n)|\{i \in [n] : \langle u, X_i\rangle \leq t\}|.$$

The functions $F_\theta(u,\cdot)$ and $\widehat{F}_n(u,\cdot)$ are the cumulative distribution functions of $u^\star_\#\mu_\theta$ and $u^\star_\#\widehat{\mu}_n$ where $u \in \mathbb{S}^{d-1}$ is a unit vector. Let $L(\mathbb{S}^{d-1} \times \mathbb{R})$ be the class of functions on $\mathbb{S}^{d-1} \times \mathbb{R}$ such that $f(\cdot, t)$ is continuous and $f(u,\cdot)$ is absolutely integrable, with the norm $\|f\|_L = \sup_{u\in\mathbb{S}^{d-1}} \int_{\mathbb{R}} |f(u,t)|\, dt$.

Assumption 3.6 *There exists a measurable function $D_\star : \mathbb{S}^{d-1} \times \mathbb{R} \to \mathbb{R}^{d_\theta}$ such that $\|F_\theta(u,t) - F_{\theta_\star}(u,t) - \langle \theta - \theta_\star, D_\star(u,t)\rangle\|_L = o(\|\theta - \theta_\star\|_\Theta)$.*

Assumption 3.7 *There exists a random element $G_\star : \mathbb{S}^{d-1} \times \mathbb{R} \mapsto \mathbb{R}$ such that the stochastic process $\sqrt{n}(\widehat{F}_n - F_\star)$ converges weakly in $L(\mathbb{S}^{d-1} \times \mathbb{R})$ to $G_\star$[3].*

Assumption 3.8 *There exists a neighborhood $\mathcal{N}$ of $\theta_\star \in \Theta$ and a positive constant $c_\star$ such that $\overline{\mathcal{PW}}_{1,1}(\mu_\theta, \mu_\star) \geq \overline{\mathcal{PW}}_{1,1}(\mu_{\theta_\star}, \mu_\star) + c_\star\|\theta - \theta_\star\|_\Theta$ for all $\theta \in \mathcal{N}$.*

Remark 3.2 *Assumption 3.6 is strictly weaker than a norm-differentiation condition where $D_\star$ has to be nonsingular. Assumption 3.7 permits model misspecification where there is no $\theta_\star \in \Theta$ such that $F_{\theta_\star} = F_\star$ and thus is more general than Nadjahi et al. (2019, A8). Assumption 3.8 accounts for local strong identifiability for the model μ_θ around $\theta_\star$ and is necessary for the fast rate of $n^{-1/2}$ under model misspecification. (Bernton et al. (2019) assumes the analogous condition for the Wasserstein distance. However, their analysis depends on a much stronger version with $\mathcal{N} = \Theta$.) Thanks to Assumption 3.8, we do not require the condition that the parameters are weakly separable in the PRW sense.*

Remark 3.3 *In well-specified setting where there exists $\theta_\star \in \Theta$ such that $F_\star = F_{\theta_\star}$, it is straightforward to derive the norm-differentiation condition from Assumption 3.6 and 3.8. This is not true, however, under model misspecification. Moreover, there are minor*

[3] As pointed by Nadjahi et al. (2019), one can prove that Assumption 3.7 holds in general by extending (Dede, 2009, Proposition 3.5) and (del Barrio et al., 1999, Theorem 2.1(a)) with some mild conditions on the tails of $u^\star_\#\mu_\star$. Using the same argument, this extension can also be done for $\|\cdot\|_L$ in our paper.

technical issues in the proof of Bernton et al. (2019, Theorem B.8); see Appendix E.4. Fixing them would be straightforward but require additional assumptions. Fortunately, we can overcome this gap using some new techniques. Thus, with some refinement, our results can be interpreted as an improvement of Bernton et al. (2019) with fewer assumptions.

To study the asymptotic distributions in the misspecified setting, we employ definitions from Pollard (1980, Section 7). (Note, however, that our proof technique is different from Pollard (1980), which depends on the nonsingularity of $D_\star$ and requires $\mu_\star = \mu_{\theta_\star}$ for some $\theta_\star$ in the interior of Θ.)

Definition 3.5 (Hausdorff metric) *Let $\mathcal{S}$ be the class of convex and compact sets in $L(\mathbb{S}^{d-1} \times \mathbb{R})$ equipped with $\|\cdot\|_L$. The Hausdorff metric on $\mathcal{S}$ is defined by $d_H(S_1, S_2) = \inf\{\delta > 0 : S_1 \subseteq S_2^\delta, S_2 \subseteq S_1^\delta\}$, where $S^\delta = \cup_{x \in S}\{z \in L(\mathbb{S}^{d-1} \times \mathbb{R}) : \|z - x\|_L \leq \delta\}$.*

Definition 3.6 (Approximate MPRW estimators) *The set of approximate MPRW estimators is defined by $M_n = \{\theta \in \Theta : \overline{\mathcal{PW}}_{1,1}(\widehat{\mu}_n, \mu_\theta) \leq \inf_{\theta' \in \Theta} \overline{\mathcal{PW}}_{1,1}(\widehat{\mu}_n, \mu_{\theta'}) + \eta_n/\sqrt{n}\}$, where $\eta_n > 0$ such that $\mathbb{P}(\eta_n \to 0) = 1$ and M_n is nonempty.*

Theorem 3.12 *Suppose Assumption 3.1-3.3 and 3.6-3.8 hold for some $\theta_\star$ in the interior of Θ and let $G_n = \sqrt{n}(\widehat{F}_n - F_{\theta_\star})$ and $G_n^\star = G_\star + \sqrt{n}(F_\star - F_{\theta_\star})$. We also define $K(x, \beta) = \{\theta \in \mathcal{N}_1 : \|x - \sqrt{n}\langle\theta - \theta_\star, D_{\theta_\star}\rangle\|_L \leq \inf_{\theta' \in \mathcal{N}_1} \|x - \sqrt{n}\langle\theta' - \theta_\star, D_{\theta_\star}\rangle\|_L + \beta\}$ where*

$$\mathcal{N}_1 = \left\{\theta \in \mathcal{N} : \frac{\|F_\theta - F_{\theta_\star} - \langle\theta - \theta_\star, D_\star\rangle\|_L}{\|\theta - \theta_\star\|_\Theta} \leq \frac{c_\star}{2}\right\}.$$

Then there exists a sequence satisfying $\lim_{n\to+\infty} \beta_n = 0$ such that[4] $\mathbb{P}_\star(M_n \subseteq K(G_n, \beta_n)) \to 1$ as $n \to +\infty$. For any $\epsilon > 0$, we have $\mathbb{P}(d_H(K(G_n^\star, 0), K(G_n, \beta_n)) < \epsilon) \to 1$ as $n \to +\infty$.

Theorem 3.12 provides the theoretical guarantee for statistical inference with the max-SW distance under model misspecification. Indeed, since $K(G_n^\star, 0) = \operatorname{argmin}_{\theta \in \mathcal{N}_1} \|G_\star + \sqrt{n}(F_\star - F_{\theta_\star} - \langle\theta - \theta_\star, D_{\theta_\star}\rangle)\|_L$, the results indicate that the distributional limit of the approximate MPRW estimator set is close to the limit of the sets $\operatorname{argmin}_{\theta \in \mathcal{N}_1} \|G_\star + \sqrt{n}(F_\star - F_{\theta_\star} - \langle\theta - \theta_\star, D_{\theta_\star}\rangle)\|_L$ in the Hausdorff metric. Note that $d > 1$ is allowed but we need $k = 1$. This is necessary for our techniques since the current analysis heavily depends on the explicit form of PRW using cumulative distribution functions. Deriving CLT when $k > 1$ is important but out of the scope of this paper.

[4] $\mathbb{P}_\star$ denotes the (inner) probability; see Pollard (1980) for details.

Remark 3.4 *In the well-specified setting, Assumption 3.8 can be replaced by Assumption A.1-A.2. Under certain conditions, we derive the CLT (cf. Theorem A.3) which is analogous to Nadjahi et al. (2019, Theorem 6) for the minimum sliced Wasserstein estimators. We refer to Theorem A.3 in Appendix A for a simplified version in well-specified setting.*

Discussions. We make some additional remarks on the relationship between our work and the existing works by Bernton et al. (2019) and Nadjahi et al. (2019). Since PRW is a type of Wasserstein, the consistency proof roadmap is essentially similar to that in Bernton et al. (2019) and Nadjahi et al. (2019). However, we remark that (i) the sample complexity bounds of PRW are new; (ii) the techniques for CLT in misspecified settings did not appear in Nadjahi et al. (2019) and complete the analysis in Bernton et al. (2019). Remark 3.2 states that our Assumption 3.8 is weaker than that is used in Bernton et al. (2019). In particular, $\mathcal{N}$ is the neighborhood defined in Assumption 3.8 and accounts for a **local** strong identifiability. In contrast, Bernton et al. (2019) requires a **global** strong identifiability ($\mathcal{N} = \Theta$). Remark 3.3 states that our setting is more general than the well-specified setting which is discussed by Nadjahi et al. (2019) from a technical point of view.

4 Experiments

We empirically validate our theoretical findings through several experiments on synthetic and real data. Given the space limit, we present the experimental setup in Appendix G and explain an optimization algorithm for computing the PRW distance and estimators in Appendix F. We defer the additional results on other dataset to Appendix H.

We set $\mu = \nu = \mathcal{U}([-v, v]^d)$ as an uniform distribution over a hypercube and study the convergence and computation of $\mathcal{PW}_{2,k}(\widehat{\mu}_n, \widehat{\nu}_n)$ and $\overline{\mathcal{PW}}_{2,k}(\widehat{\mu}_n, \widehat{\nu}_n)$ for $n \in \{20, 100, 250, 500, 1000\}$. Figure 1 presents average distances and computational times for $(d, v) \in \{(10, 1), (30, 5), (50, 5)\}$, where the shaded areas show the max-min values over 100 runs. First, the IPRW distance is significantly smaller than the PRW distance for small n especially when d and v are large. This confirms Theorem 3.4 which shows that the IPRW distance is independent of d. Second, the PRW distance nearly matches the IPRW distance when n is large. This confirms Theorem 3.6 since the uniform distribution with its bounded domain satisfies the Poincaré inequality. Finally, the computation of the PRW distance is faster than that of the IPRW distance.

Consider the parametric inference using Gaussian

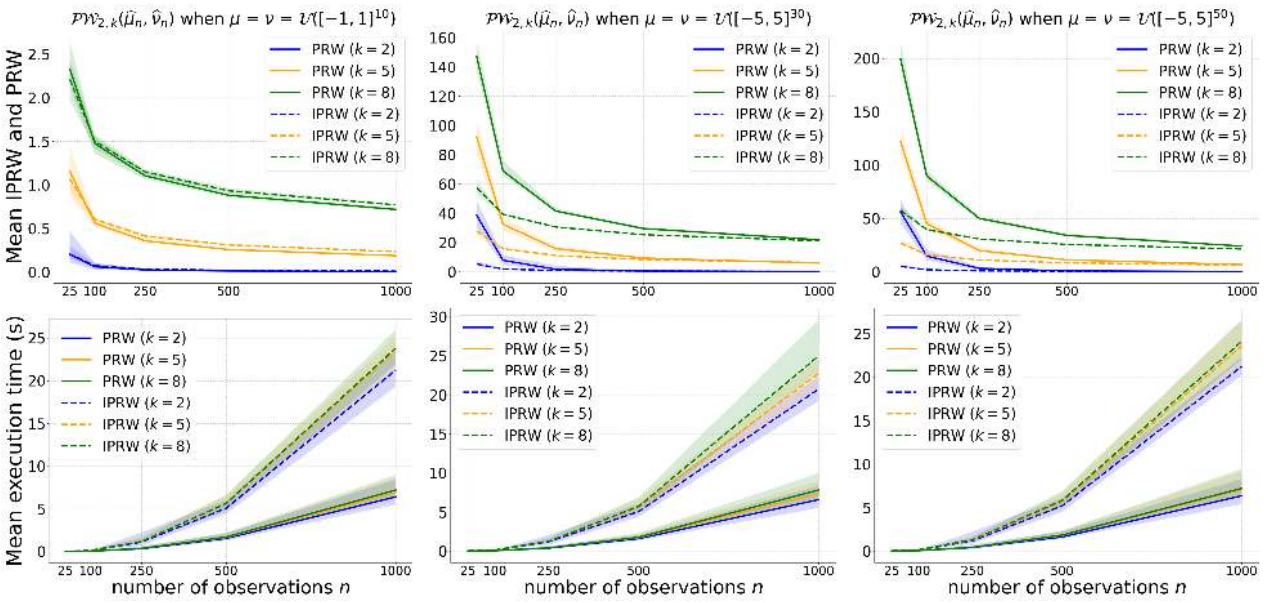


Figure 1: Mean values (Top) and mean computational time (Bottom) of the IPRW and PRW distances of order 2 between empirical measures $\widehat{\mu}_n$ and $\widehat{\nu}_n$ as the number of points n varies. Results are averaged over 100 runs.

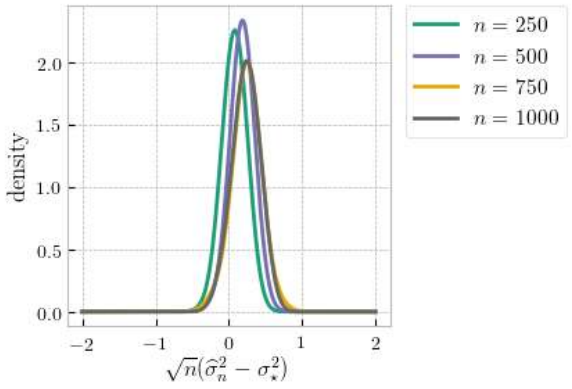


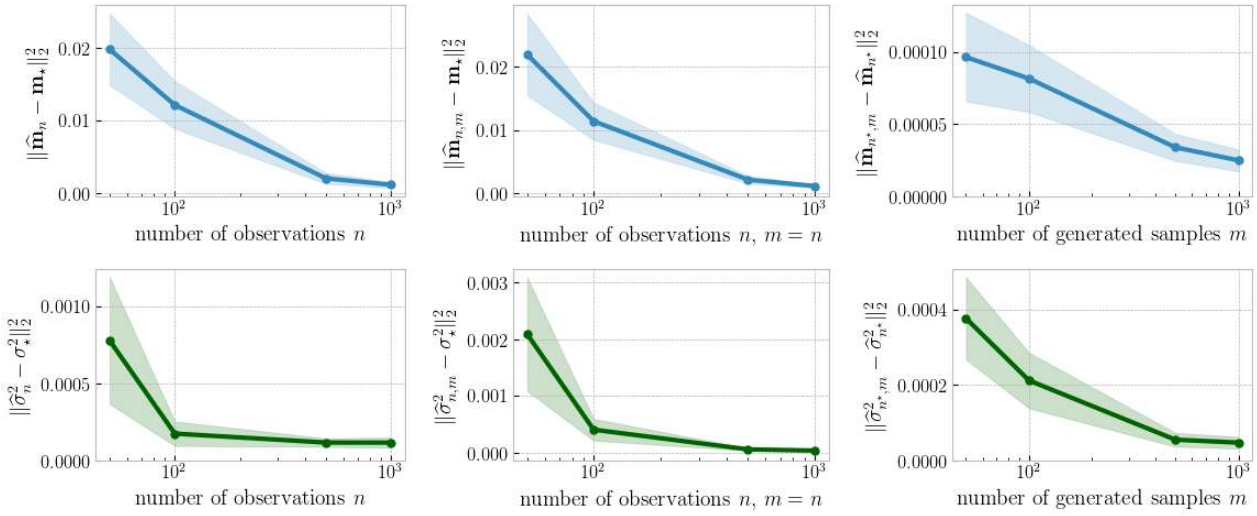
Figure 2: Probability density of estimation of centered and rescaled $\widehat{\sigma}_n$ on the Gaussian model.

models $\mathcal{M} = \{\mathcal{N}(\mathbf{m}, \sigma^2\mathbf{I}) : \mathbf{m}\in\mathbb{R}^2, \sigma^2>0\}$ and a collection of i.i.d. observations generated from a mixture of 8 Gaussian distributions in $\mathbb{R}^2$. This simple setting is useful since the closed-form expression of Gaussian density makes the computation of the MPRW estimator of order 1 tractable. Following the setup in Nadjahi et al. (2019, Section 4), we illustrate the consistency of the MPRW and MEPRW estimators of order 1 and the convergence of MEPRW estimator of order 1 to MPRW estimator of order 1. Results are shown in Figure 3; they are consistent with Theorem 3.9, 3.10 and 3.11, where $\mathbf{m}_\star = \widehat{\mathbf{m}}_{10^5}$. Despite the model misspecification, our estimators still converge as the number of observations increases and the MEPRW estimator converges to the MPRW estimator as we generate more samples. We also verify our central limit theorem by estimating the density of $\widehat{\sigma}_n^2$ with a kernel density estimator[5] over 100 runs. Figure 2 shows the distribution centered and rescaled by $\sqrt{n}$ for each n, where $\sigma_\star^2 = \widehat{\sigma}_{10^5}^2$, and it confirms the convergence rate we derived in Theorem 3.12; see Appendix H for the case with 12 or 25 distributions.

We conduct experiments on image generation using the PRW generator of order 2, as an alternative to the SW generator (Deshpande et al., 2018). Here we focus on the case of $k = 1$, where the PRW generator is exactly max-SW generator. We train the neural networks (NNs) with $(n, m) \in \{(100, 20), (1000, 40), (5000, 60), (10000, 100)\}$ where n is the number of training samples and m is the number of generated samples. We compare their testing losses to that of a NN trained using $n = 10^5$ (i.e. whole training dataset) and $m = 200$. All the testing losses are evaluated using the trained models on the the testing dataset ($n = 10^4$) with $m = 250$ generated samples. Figure 4 presents the mean testing loss on ImageNet200 over 10 runs, where the shaded areas show the max-min values over the runs.

Discussions. First, PRW has better discriminative power than max-SW or SW since it considers high-order summaries and extract more geometric information from two high-dimensional distributions, in order to distinguish them better; see Paty and Cuturi (2019) for the details. Moreover, we have presented in Figure 1 (top row) and Figure 5 (top row) that the PRW/IPRW value increase as k increases.

[5]The approach we apply here is the same as used by Nadjahi et al. (2019).



(a) MPRW vs. n (b) MEPRW vs. $n = m$ (c) MEPRW with $n = 2000$ vs. m

Figure 3: Minimal PRW and expected PRW estimations using Gaussian models and n samples from the mixture of 8 Gaussian distributions. Results are averaged over 100 runs and shaded areas represent standard deviation.

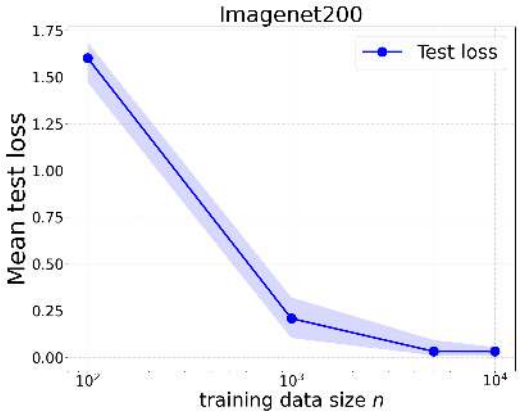


Figure 4: Mean test loss for different value of (n, m) on ImageNet200.

Thus, PRW/IPRW based on larger k-dimensional projections have better discriminative power.

Second, the IPRW computation generally requires many random projections and is thus more time-consuming than PRW for a desired accuracy when $k = 1$; see Kolouri et al. (2019a, Page 4). Fortunately, it may require much fewer for certain application problems when the intrinsic dimension of data distribution is small, and is easily amenable to parallel computation. Thus, IPRW can serve as a practical alternative to PRW. Moreover, the reported PRW and IPRW values in Figure 1 and Figure 5 (appendix) are computed by using 30 iterations for PRW and 100 projections for IPRW. Therefore, the statistical/simulation error contributes to the flip of order between IPRW and PRW when their true values are close.

Finally, our experimental results show that the max-sliced Wasserstein estimator works well in practice and converges to some point as the number of samples grow. This supports our consistency results since the max-sliced Wasserstein distance is PRW with $k = 1$. Note that there are many existing works on the empirical comparison between max-SW and SW using generative modeling and we refer the interested readers to Kolouri et al. (2019a) and the reference therein.

5 Conclusion

We study in this paper the statistical aspect of the projection robust Wasserstein (PRW) distance. Our work provides an enhanced understanding of two PRW distances and the associated minimal distance estimators under model misspecification, complementing the existing literature (Niles-Weed and Rigollet, 2019; Bernton et al., 2019; Nadjahi et al., 2019, 2020). Experiments on synthetic and real datasets highlight some aspects of our theoretical results. Future work includes theory for entropic PRW and the applications of PRW with $k \geq 2$ to deep generative models.

6 Acknowledgements

We would like to thank the area chair and four anonymous referees for constructive suggestions that improve the quality of this paper. Elynn Y. Chen is supported by National Science Foundation under the grant number DMS-1803241. This work was supported in part by the Mathematical Data Science program of the Office of Naval Research under grant number N00014-18-1-2764.

References

P-A. Absil, R. Mahony, and R. Sepulchre. *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, 2009.

J. Adler and S. Lunz. Banach Wasserstein GAN. In *NIPS*, pages 6754–6763, 2018.

C. D. Aliprantis and K. C. Border. *Infinite Dimensional Analysis: A Hitchhiker's Guide*. Springer Science & Business Media, 2006.

M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein generative adversarial networks. In *ICML*, pages 214–223, 2017.

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

A. Basu, H. Shioya, and C. Park. *Statistical Inference: The Minimum Distance Approach*. CRC Press, 2011.

E. Bayraktar and G. Guo. Strong equivalence between metrics of wasserstein type. *ArXiv Preprint: 1912.08247*, 2019.

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*, 8(4):657–676, 2019.

P. Billingsley. *Convergence of Probability Measures*. John Wiley & Sons, 2013.

N. Bonneel, J. Rabin, G. Peyré, and H. Pfister. Sliced and radon Wasserstein barycenters of measures. *Journal of Mathematical Imaging and Vision*, 51(1): 22–45, 2015.

N. Bonneel, G. Peyré, and M. Cuturi. Wasserstein barycentric coordinates: histogram regression using optimal transport. *ACM Transactions on Graphics*, 35(4):71:1–71:10, 2016.

N. Bonnotte. *Unidimensional and Evolution Methods for Optimal Transportation*. PhD thesis, Paris 11, 2013.

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

J. Cao, L. Mo, Y. Zhang, K. Jia, C. Shen, and M. Tan. Multi-marginal Wasserstein GAN. In *NeurIPS*, pages 1774–1784, 2019.

M. Carriere, M. Cuturi, and S. Oudot. Sliced Wasserstein kernel for persistence diagrams. In *ICML*, pages 664–673. JMLR. org, 2017.

M. Cuturi and A. Doucet. Fast computation of Wasserstein barycenters. In *ICML*, pages 685–693, 2014.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *NeurIPS*, pages 2292–2300, 2013.

S. Dede. An empirical central limit theorem in L1 for stationary sequences. *Stochastic Processes and Their Applications*, 119(10):3494–3515, 2009.

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

I. Deshpande, Z. Zhang, and A. G. Schwing. Generative modeling using the sliced Wasserstein distance. In *CVPR*, pages 3483–3491, 2018.

I. Deshpande, Y-T. Hu, R. Sun, A. Pyrros, N. Siddiqui, S. Koyejo, Z. Zhao, D. Forsyth, and A. G. Schwing. Max-sliced Wasserstein distance and its use for GANs. In *CVPR*, pages 10648–10656, 2019.

R. M. Dudley. The speed of mean Glivenko-Cantelli convergence. *The Annals of Mathematical Statistics*, 40(1):40–50, 1969.

R. Flamary and N. Courty. Pot python optimal transport library, 2017. URL `https://github.com/rflamary/POT`.

N. Fournier and A. Guillin. On the rate of convergence in Wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, 162(3-4):707–738, 2015.

I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville. Improved training of Wasserstein GANs. In *NeurIPS*, pages 5767–5777, 2017.

T. Hashimoto, D. Gifford, and T. Jaakkola. Learning population-level diffusions with generative RNNs. In *ICML*, pages 2417–2426, 2016.

N. Ho, X. Nguyen, M. Yurochkin, H. H. Bui, V. Huynh, and D. Phung. Multilevel clustering via Wasserstein means. In *ICML*, pages 1501–1509, 2017.

H. Janati, T. Bazeille, B. Thirion, M. Cuturi, and A. Gramfort. Multi-subject meg/eeg source imaging with sparse multi-task regression. *NeuroImage*, page 116847, 2020.

D. P. Kingma and J. Ba. ADAM: A method for stochastic optimization. In *ICLR*, 2015.

S. Kolouri, Y. Zou, and G. K. Rohde. Sliced Wasserstein kernels for probability distributions. In *CVPR*, pages 5258–5267, 2016.

S. Kolouri, K. Nadjahi, U. Simsekli, R. Badeau, and G. Rohde. Generalized sliced Wasserstein distances. In *NeurIPS*, pages 261–272, 2019a.

S. Kolouri, P. E. Pope, C. E. Martin, and G. K. Rohde. Sliced Wasserstein auto-encoders. In *ICLR*, 2019b.

M. Ledoux. Concentration of measure and logarithmic sobolev inequalities. In *Seminaire de probabilites XXXIII*, pages 120–216. Springer, 1999.

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

X. Li, S. Chen, Z. Deng, Q. Qu, Z. Zhu, and A. M-C. So. Nonsmooth optimization over Stiefel manifold: Riemannian subgradient methods. *ArXiv Preprint: 1911.05047*, 2019.

T. Lin, C. Fan, N. Ho, M. Cuturi, and M. I. Jordan. Projection robust Wasserstein distance and Riemannian optimization. *ArXiv Preprint: 2006.07458*, 2020.

H. Liu, A. M-C. So, and W. Wu. Quadratic optimization with orthogonality constraint: explicit łojasiewicz exponent and linear convergence of retraction-based line-search and stochastic variance-reduced gradient methods. *Mathematical Programming*, 178(1-2):215–262, 2019.

A. Liutkus, U. Simsekli, S. Majewski, A. Durmus, and F-R. Stöter. Sliced-Wasserstein flows: Nonparametric generative modeling via optimal transport and diffusions. In *ICML*, pages 4104–4113, 2019.

T. Manole, S. Balakrishnan, and Larry Wasserman. Minimax confidence intervals for the sliced wasserstein distance. *ArXiv Preprint: 1909.07862*, 2019.

G. Montavon, K-R. Müller, and M. Cuturi. Wasserstein training of restricted Boltzmann machines. In *NIPS*, pages 3718–3726. Curran Associates, Inc., 2016.

K. Nadjahi, A. Durmus, U. Simsekli, and R. Badeau. Asymptotic guarantees for learning generative models with the sliced-Wasserstein distance. In *NeurIPS*, pages 250–260, 2019.

K. Nadjahi, A. Durmus, L. Chizat, S. Kolouri, S. Shahrampour, and U. Şimşekli. Statistical and topological properties of sliced probability divergences. *ArXiv Preprint: 2003.05783*, 2020.

J. S. Nath and P. Jawanpuria. Statistical optimal transport posed as learning kernel embedding. *ArXiv Preprint: 2002.03179*, 2020.

K. Nguyen, N. Ho, T. Pham, and H. Bui. Distributional sliced-Wasserstein and applications to generative modeling. *ArXiv Preprint: 2002.07367*, 2020.

J. Niles-Weed and P. Rigollet. Estimation of Wasserstein distances in the spiked transport model. *ArXiv Preprint: 1909.07513*, 2019.

J. P. Nolan. Multivariate elliptically contoured stable distributions: theory and estimation. *Computational Statistics*, 28(5):2067–2089, 2013.

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

F-P. Paty and M. Cuturi. Subspace robust Wasserstein distances. In *ICML*, pages 5072–5081, 2019.

G. Peyré and M. Cuturi. Computational optimal transport. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

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

J. Rabin, G. Peyré, J. Delon, and M. Bernot. Wasserstein barycenter and its application to texture mixing. In *International Conference on Scale Space and Variational Methods in Computer Vision*, pages 435–446. Springer, 2011.

S. T. Rachev and L. Rüschendorf. *Mass Transportation Problems: Volume I: Theory*, volume 1. Springer Science & Business Media, 1998.

A. Ramdas, N. G. Trillos, and M. Cuturi. On Wasserstein two-sample testing and related families of nonparametric tests. *Entropy*, 19(2):47, 2017.

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

G. Samoradnitsky. *Stable Non-Gaussian Random Processes: Stochastic Models with Infinite Variance*. Routledge, 2017.

G. Schiebinger, J. Shu, M. Tabaka, B. Cleary, V. Subramanian, A. Solomon, S. Liu, S. Lin, P. Berube, L. Lee, et al. Reconstruction of developmental landscapes by optimal-transport analysis of single-cell gene expression sheds light on cellular reprogramming. *bioRxiv*, page 191056, 2017.

S. Singh and B. Póczos. Minimax distribution estimation in Wasserstein distance. *ArXiv Preprint: 1802.08855*, 2018.

M. Talagrand. Transportation cost for Gaussian and other product measures. *Geometric & Functional Analysis GAFA*, 6(3):587–600, 1996.

I. Tolstikhin, O. Bousquet, S. Gelly, and B. Schoelkopf. Wasserstein auto-encoders. In *ICLR*, 2018.

A. Tong, J. Huang, G. Wolf, D. van Dijk, and S. Krishnaswamy. Trajectorynet: A dynamic optimal transport network for modeling cellular dynamics. *ArXiv Preprint: 2002.04461*, 2020.

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

M. J. Wainwright. *High-dimensional Statistics: A Non-asymptotic Viewpoint*, volume 48. Cambridge University Press, 2019.

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

J. Wolfowitz. The minimum distance method. *The Annals of Mathematical Statistics*, pages 75–88, 1957.

J. Wu, Z. Huang, D. Acharya, W. Li, J. Thoma, D. P. Paudel, and L. V. Gool. Sliced Wasserstein generative models. In *CVPR*, pages 3713–3722, 2019.

K. D. Yang, K. Damodaran, S. Venkatachalapathy, A. C. Soylemezoglu, G. V. Shivashankar, and C. Uhler. Predicting cell lineages using autoencoders and optimal transport. *PLoS computational biology*, 16 (4):e1007828, 2020.

J. Ye, P. Wu, J. Z. Wang, and J. Li. Fast discrete distribution clustering using Wasserstein barycenter with sparse support. *IEEE Transactions on Signal Processing*, 65(9):2317–2332, 2017.