*Information and Inference: A Journal of the IMA* (2023) **00**, 1–29
https://doi.org/DOI HERE

# Statistical Inference with Regularized Optimal Transport

ZIV GOLDFELD*
*School of Electrical and Computer Engineering, Cornell University, Rhodes Hall, Ithaca, NY 14853, USA*

KENGO KATO
*Department of Statistics and Data Science, Cornell University, Comstock Hall, Ithaca, NY 14853, USA*

GABRIEL RIOUX
*Center for Applied Mathematics, Cornell University, Rhodes Hall, Ithaca, NY 14853, USA*

AND

RITWIK SADHU
*Department of Statistics and Data Science, Cornell University, Comstock Hall, Ithaca, NY 14853, USA*
*Corresponding author: goldfeld@cornell.edu

[Received on 16 June 2022; revised on 31 July 2023; accepted on 22 November 2023]

Optimal transport (OT) is a versatile framework for comparing probability measures, with many applications to statistics, machine learning, and applied mathematics. However, OT distances suffer from computational and statistical scalability issues to high dimensions, which motivated the study of regularized OT methods like slicing, smoothing, and entropic penalty. This work establishes a unified framework for deriving limit distributions of empirical regularized OT distances, semiparametric efficiency of the plug-in empirical estimator, and bootstrap consistency. We apply the unified framework to provide a comprehensive statistical treatment of: (i) average- and max-sliced $p$-Wasserstein distances, for which several gaps in existing literature are closed; (ii) smooth distances with compactly supported kernels, the analysis of which is motivated by computational considerations; and (iii) entropic OT, for which our method generalizes existing limit distribution results and establishes, for the first time, efficiency and bootstrap consistency. While our focus is on these three regularized OT distances as applications, the flexibility of the proposed framework renders it applicable to broad classes of functionals beyond these examples.

*Keywords:* bootstrap consistency; entropic optimal transport; limit distribution; semiparametric efficiency; sliced Wasserstein distance; smooth Wasserstein distance.

## 1. Introduction

Optimal transport (OT) theory [80, 96] provides a versatile framework for comparing probability distributions. Introduced by Monge [64] and later formulated by Kantorovich [49], the OT problem between two Borel probability measures $\mu,\nu$ on $\mathbb{R}^d$ is defined by

$$\mathsf{T}_c(\mu,\nu) := \inf_{\pi\in\Pi(\mu,\nu)} \int_{\mathbb{R}^d\times\mathbb{R}^d} c(x,y)d\pi(x,y), \tag{1.1}$$

where $\Pi(\mu,\nu)$ is the set of couplings between $\mu$ and $\nu$. The special case of the $p$-Wasserstein distance for $p\in[1,\infty)$ is given by $\mathsf{W}_p(\mu,\nu) := \big(\mathsf{T}_{\|\cdot\|^p}(\mu,\nu)\big)^{1/p}$. Thanks to an array of favorable properties, including the Wasserstein metric structure, a convenient duality theory, robustness to support mismatch, and the rich geometry induced on the space of probability measures, OT and the Wasserstein distance have seen a surge of applications in statistics, machine learning, and applied mathematics. These include generative modeling [4, 8, 18, 45, 88], robust/adversarial machine learning (ML) [10, 98], domain adaptation [21, 85], image recognition [54, 77, 79], vector quantile regression [16, 20, 38, 46], Bayesian estimation [7], and causal inference [89]. Unfortunately, OT distances are generally hard to compute and suffer from the curse of dimensionality in empirical estimation, whereby the number of samples needed for reliable estimation grows exponentially with dimension.

These deficits have motivated the introduction of regularized OT methods that aim to alleviate the said computational and statistical bottlenecks. Three prominent regularizations are: (1) slicing via lower-dimensional projections [6, 14, 66, 67, 75]; (2) smoothing via convolution with a chosen kernel [11, 17, 40, 41, 42, 43, 44, 47, 68, 78, 101]; and (3) convexification via entropic penalty [1, 22, 29, 37, 52, 60, 82]. These techniques preserve many properties of classic OT but avoid the curse of dimensionality, which enables a scalable statistical theory. As reviewed below[1], much effort was devoted to exploring dimension-free empirical convergence rates and limit distributions, bootstrapping, and other statistical aspects of regularized OT, although several notable gaps in the literature remain. Furthermore, proof techniques for such results are typically on a case-by-case basis and do not follow a unified approach, despite evident similarities between the three regularization methods as complexity reduction techniques of the classic OT framework.

The present paper develops a unified framework for deriving limit distributions, semiparametric efficiency bounds, and bootstrap consistency for a broad class of functionals that, in particular, encompasses the empirical regularized OT distances mentioned above (Section 3). As example applications of the general framework, we explore a comprehensive treatment of the following problems:

- **Average- and max-sliced $\mathsf{W}_p$ (Section 4):** Our limit distribution theory closes existing gaps in the literature (e.g., a limit distribution result for sliced $\mathsf{W}_1$ was assumed in [67] but left unproven), with the efficiency and bootstrap consistency results providing additional constituents for valid statistical inference.
- **Smooth $\mathsf{W}_p$ with compactly supported kernels (Section 5):** Gaussian-smoothed OT was previously shown to preserve the classic Wasserstein structure while alleviating the curse of dimensionality. Motivated by computational considerations, herein we study smoothing with compactly supported kernels. We explore the metric, topological, and statistical aspects previously derived under Gaussian smoothing.
- **Entropic OT (Section 6):** A central limit theorem (CLT) for empirical entropic OT (EOT) was derived [29, 60] for independent data via a markedly different proof technique than proposed herein. Revisiting this problem using our general machinery, we rederive this CLT allowing for dependent data, and also obtain new results on semiparametric efficiency and bootstrap consistency.

The unified limit distribution framework, stated in Proposition 1, relies on the extended functional delta method for Hadamard directionally differentiable functionals [76, 83]. To match the delta method with the regularized OT setup, we focus on a functional on a space of probability measures that is (a) locally Lipschitz with respect to (w.r.t.) the sup-norm for a Donsker function class and (b) Gâteaux directionally differentiable at the population distribution. To apply this framework, we seek to: (i) set up the regularized distance as a locally Lipschitz functional $\delta$ w.r.t. $\|\cdot\|_{\infty,\mathscr{F}} = \sup_{f\in\mathscr{F}}|\cdot|$; (ii) show $\mathscr{F}$ to be Donsker to obtain convergence of the empirical process in $\ell^\infty(\mathscr{F})$; (iii) characterize the Gâteaux

[1] We postpone the literature review on each regularization method to its respective section.

directional derivative of $\delta$ at $\mu$. For each regularized distance (sliced, smooth, and entropic), we identify the appropriate function class $\mathscr{F}$ and establish the desired Lipschitz continuity and differentiability, relying on OT duality theory. Regularization enforces the dual potentials to possess smoothness or low-dimensionality properties, which are leveraged to show that $\mathscr{F}$ is Donsker. Of note is that our framework does not require independent and identically distributed (i.i.d.) data and can be applied for any estimate (not only the empirical distribution) of the population distribution, so long as the uniform limit theorem mentioned in (ii) holds true.

As the general framework stems from the extended functional delta method, the limiting variable of the (scaled and centered) empirical regularized distance is given by the directional derivative of $\delta$ at the population distribution. Linearity of the derivative implies that the limit variable is centered Gaussian. In this case, it is natural to ask whether the empirical distance attains the semiparametric efficiency lower bound (cf. [93, Chapter 25]). Semiparametric efficiency bounds serve as analogs of Cramér-Rao lower bounds in semiparametric estimation and account for the fundamental difficulty of estimating functionals of interest. We show that the asymptotic variance of the empirical distance indeed agrees with the semiparametric efficiency bound, relative to a certain tangent space. Still, even when the limiting variable is Gaussian, direct analytic estimation of the asymptotic variance may be nontrivial. To account for that, we explore bootstrap consistency for empirical regularized OT distances. Altogether, the limit distribution theory, semiparametric efficiency, and bootstrap consistency provide a comprehensive statistical account of the considered regularized OT distances.

A unifying approach of a similar flavor to ours, but for classic OT distances, was proposed in [48]. Focusing solely on the supremum functional, they used the extended functional delta method to derive limit distributions for classic $\mathsf{W}_p$, with $p \geq 2$, for compactly supported distributions under the alternative in dimensions $d \leq 3$. In comparison, our approach is more general and can treat any functional that adheres to the aforementioned local Lipschitz continuity and differentiability. This is crucial for analyzing regularized OT distances as some instances do not amount to a supremum functional. For instance, average-sliced Wasserstein distances correspond to mixed $L^1$-$L^\infty$ functionals, which are not accounted for by the setup from [48]. The functional delta method was also used in [86, 87] to derive limit distributions for OT between discrete population distributions by parametrizing them using simplex vectors. This result was extended to semi-discrete OT in [30] by exploiting the fact that complexity of the optimal potentials class is reduced when one of the measures is supported on a discrete set. Another recent application can be found in [44], where this approach was leveraged for Gaussian-smoothed $\mathsf{W}_p$ by embedding the domain of the Wasserstein distance into a certain dual Sobolev space.

The paper is organized as follows. Section 2 presents notation used throughout the paper and background on Wasserstein distances and the extended functional delta method. Section 3 presents a unified framework for deriving limit distributions, bootstrap consistency, and semiparametric efficiency bounds for regularized OT distances. The tools developed therein will be applied to sliced Wasserstein distances in Section 4, smooth Wasserstein distances with compactly supported kernels in Section 5, and EOT in Section 6. Section 7 leaves some concluding remarks. Proofs for the results in Sections 2–6 are found in Appendices **??**–**??**.

## 2. Background and Preliminaries

This section collects notation used throughout the paper and sets up necessary background on Wasserstein distances and the extended functional delta method.

### 2.1. *Notation*

Let $\|\cdot\|$ denote the Euclidean norm and $B(x,r)$ be the open ball with center $x \in \mathbb{R}^d$ and radius $r > 0$. For a subset $A$ of a topological space $S$, let $\overline{A}^S$ denote the closure of $A$; if the space $S$ is clear from the context, then we simply write $\overline{A}$ for the closure. The space of Borel probability measures on $S$ is denoted by $\mathscr{P}(S)$. When $S$ is a normed space with norm $\|\cdot\|_S$, we denote $\mathscr{P}_p(S) := \{\mu \in \mathscr{P}(S) : \int \|x\|_S^p d\mu(x) < \infty\}$ for $1 \le p < \infty$. The (topological) support of $\mu \in \mathscr{P}(S)$ is denoted as $\mathrm{spt}(\mu)$. For any finite signed Borel measure $\gamma$ on $S$, we identify $\gamma$ with the linear functional $f \mapsto \gamma(f) = \int f d\gamma$. For $\mu \in \mathscr{P}(S)$ and a $\mu$-integrable function $h$ on $S$, $h\mu$ denotes the signed measure $hd\mu$. Let $\xrightarrow{w}, \xrightarrow{d}$, and $\xrightarrow{\mathbb{P}}$ denote weak convergence of probability measures, convergence in distribution of random variables, and convergence in probability, respectively. When necessary, convergence in distribution is understood in the sense of Hoffmann-Jørgensen (cf. Chapter 1 in [92]). For any nonempty set $S$, let $\ell^\infty(S)$ be the Banach space of bounded real functions on $S$ equipped with the sup-norm $\|\cdot\|_{\infty,S} = \sup_{x\in S}|\cdot|$. For any measure space $(S, \mathscr{S}, \mu)$ and $1 \le p < \infty$, let $L^p(\mu) = L^p(S, \mathscr{S}, \mu)$ denote the Banach space of measurable functions $f : S \to \mathbb{R}$ with $\|f\|_{L^p(\mu)} = (\int |f|^p d\mu)^{1/p} < \infty$. If $\mu$ is $\sigma$-finite and $\mathscr{S}$ is countably generated, then the space is separable. For two numbers $a$ and $b$, we use the notation $a \wedge b = \min\{a,b\}$ and $a \vee b = \max\{a,b\}$.

### 2.2. *Wasserstein distances*

The Wasserstein distance is a specific instance of the OT problem from (1.1), defined as follows.

**Definition 1** (Wasserstein distance) *Let $1 \le p < \infty$. The $p$-th Wasserstein distance between $\mu, \nu \in \mathscr{P}_p(\mathbb{R}^d)$ is defined as*

$$\mathsf{W}_p(\mu,\nu) := \inf_{\pi\in\Pi(\mu,\nu)} \left[\int_{\mathbb{R}^d\times\mathbb{R}^d} \|x-y\|^p \, d\pi(x,y)\right]^{1/p}, \tag{2.1}$$

*where $\Pi(\mu,\nu)$ is the set of couplings of $\mu$ and $\nu$.*

The $p$-Wasserstein distance is a metric on $\mathscr{P}_p(\mathbb{R}^d)$ and metrizes weak convergence plus convergence of $p$th moments, i.e., $\mathsf{W}_p(\mu_n,\mu) \to 0$ if and only if $\mu_n \xrightarrow{w} \mu$ and $\int \|x\|^p d\mu_n(x) \to \int \|x\|^p d\mu(x)$. Wasserstein distances admit the following dual form (cf. [96, Theorem 5.9]):

$$\mathsf{W}_p^p(\mu,\nu) = \sup_{\varphi\in L^1(\mu)} \left[\int_{\mathbb{R}^d} \varphi d\mu + \int_{\mathbb{R}^d} \varphi^c d\nu\right], \tag{2.2}$$

where $\varphi^c(y) = \inf_{x\in\mathbb{R}^d}\left[\|x-y\|^p - \varphi(x)\right]$ is the $c$-transform of $\varphi$ (for the cost $c(x,y) = \|x-y\|^p$). A function $f : \mathbb{R}^d \to [-\infty,\infty)$ is called *$c$-concave* if $f = g^c$ for some function $g : \mathbb{R}^d \to [-\infty,\infty)$. There is at least one $c$-concave $\varphi \in L^1(\mu)$ that attains the supremum in (2.2), and we call this $\varphi$ an *OT potential* from $\mu$ to $\nu$ for $\mathsf{W}_p$. Further, when $1 < p < \infty$ and $\mu$ is supported on a connected set with negligible boundary and has a (Lebesgue) density, then the OT potential from $\mu$ to $\nu$ is unique on $\mathrm{int}(\mathrm{spt}(\mu))$ up to additive constants [28, Corollary 2.7]. Various smoothness properties of the potentials can be established under appropriate regularity conditions on the cost and $\mu,\nu$—a fact that we shall leverage in our derivations.

**Remark 1** (Literature review on $\mathsf{W}_p$ limit distribution theory) *Distributional limits of $\sqrt{n}\big(\mathsf{W}_p^p(\hat{\mu}_n,\nu)-\mathsf{W}_p^p(\mu,\nu)\big)$ and its two-sample analogue for discrete $\mu,\nu$ under both the null $\mu=\nu$ and the alternative $\mu\neq\nu$ were derived in [86, 87]. Similar results for general distributions are known only in the one-dimensional case. Specifically, for $p=1,2$, [25, 26] leverage the representations of $\mathsf{W}_p$ in $d=1$ as the $L^p$ norm between distribution functions ($p=1$) and quantile functions ($p=2$) to derive distributional limits under the null. Limit distributions in $d=1$ for $p\geq 2$ under the alternative ($\mu\neq\nu$) were derived in [27]. In arbitrary dimension, [24] establish asymptotic normality of $\sqrt{n}\big(\mathsf{W}_2^2(\hat{\mu}_n,\nu)-\mathbb{E}\big[\mathsf{W}_2^2(\hat{\mu}_n,\nu)\big]\big)$ under the alternative $\mu\neq\nu$ by deriving an asymptotic linear representation using the Efron-Stein inequality. This was extended to general transportation costs satisfying certain regularity conditions in [28]. The main limitation of these results is the centering around the expected empirical distance (and not the population one), which does not enable performing inference for $\mathsf{W}_p$. This gap was addressed in [58], where a CLT for $\sqrt{n}\big(\mathsf{W}_2^2(\tilde{\mu}_n,\nu)-\mathsf{W}_2^2(\mu,\nu)\big)$ was established, but for a wavelet-based estimator $\tilde{\mu}_n$ of $\mu$ (as opposed to the empirical distribution), while assuming several technical conditions on the Lebesgue densities of $\mu,\nu$. As mentioned in the introduction, [48] leverage the extended functional delta method for the supremum functional to obtain limit distributions for $\mathsf{W}_p$, with $p\geq 2$, for compactly supported distributions under the alternative in dimensions $d\leq 3$.*

### 2.3. *Extended functional delta method*

Our unified framework for deriving limit distributions of empirical regularized OT distances relies on the extended functional delta method, which we set up next. Let $\mathfrak{D},\mathfrak{E}$ be normed spaces and $\phi:\Theta\subset\mathfrak{D}\to\mathfrak{E}$ be a map. Following [76, 83], we say that $\phi$ is *Hadamard directionally differentiable* at $\theta\in\Theta$ if there exists a map $\phi'_\theta:\mathscr{T}_\Theta(\theta)\to\mathfrak{E}$ such that

$$\lim_{n\to\infty}\frac{\phi(\theta+t_nh_n)-\phi(\theta)}{t_n}=\phi'_\theta(h) \tag{2.3}$$

for any $h\in\mathscr{T}_\Theta(\theta)$, $t_n\downarrow 0$, and $h_n\to h$ in $\mathfrak{D}$ such that $\theta+t_nh_n\in\Theta$. Here $\mathscr{T}_\Theta(\theta)$ is the *tangent cone* to $\Theta$ at $\theta$ defined as

$$\mathscr{T}_\Theta(\theta):=\left\{h\in\mathfrak{D}:h=\lim_{n\to\infty}\frac{\theta_n-\theta}{t_n}\text{ for some }\theta_n\to\theta\text{ in }\Theta\text{ and }t_n\downarrow 0\right\}.$$

The tangent cone $\mathscr{T}_\Theta(\theta)$ is closed, and if $\Theta$ is convex, then $\mathscr{T}_\Theta(\theta)$ coincides with the closure in $\mathfrak{D}$ of $\{(\vartheta-\theta)/t:\vartheta\in\Theta,t>0\}$. The derivative $\phi'_\theta$ is positively homogeneous and continuous but need not be linear.

**Lemma 1** (Extended functional delta method [34, 36, 76, 84]) *Let $\mathfrak{D},\mathfrak{E}$ be normed spaces and $\phi:\Theta\subset\mathfrak{D}\to\mathfrak{E}$ be a map that is Hadamard directionally differentiable at $\theta\in\Theta$ with derivative $\phi'_\theta:\mathscr{T}_\Theta(\theta)\to\mathfrak{E}$. Let $T_n:\Omega\to\Theta$ be maps such that $r_n(T_n-\theta)\xrightarrow{d}T$ for some $r_n\to\infty$ and Borel measurable map $T:\Omega\to\mathfrak{D}$ with values in $\mathscr{T}_\Theta(\theta)$. Then, $r_n\big(\phi(T_n)-\phi(\theta)\big)\xrightarrow{d}\phi'_\theta(T)$. Further, if $\Theta$ is convex, then we have $r_n\big(\phi(T_n)-\phi(\theta)\big)-\phi'_\theta\big(r_n(T_n-\theta)\big)\to 0$ in outer probability.*

Lemma 1 is at the core of our framework for deriving limit distributions. It is termed the "extended" functional delta method as it extends the (classical) functional delta method for Hadamard differentiable maps to directionally differentiable ones.

While Hadamard directional differentiability is sufficient to derive limit distributions, bootstrap consistency often requires (full) Hadamard differentiability. Recall that the map $\phi$ is *Hadamard differentiable* at $\theta$ *tangentially to* a vector subspace $\mathfrak{D}_0 \subset \mathfrak{D}$ if there exists a continuous *linear* map $\phi'_\theta : \mathfrak{D}_0 \to \mathfrak{E}$ satisfying (2.3) for any $h \in \mathfrak{D}_0$, $t_n \to 0$ ($t_n \neq 0$), and $h_n \to h$ in $\mathfrak{D}$ such that $\theta + t_n h_n \in \Theta$. The differences from Hadamard directional differentiability is that the derivative $\phi'_\theta$ must be linear and thus the domain must be a vector subspace of $\mathfrak{D}$, and the sequence $t_n \to 0$ must be a generic (nonzero) sequence converging to zero. The next lemma is useful for verifying Hadamard differentiability from the directional one.

**Lemma 2** *Let $\phi : \Theta \subset \mathfrak{D} \to \mathfrak{E}$ be Hadamard directionally differentiable at $\theta \in \Theta$ with derivative $\phi'_\theta : \mathscr{T}_\Theta(\theta) \to \mathfrak{E}$. If $\mathscr{T}_\Theta(\theta)$ contains a subspace $\mathfrak{D}_0$ on which $\phi'_\theta$ is linear, then $\phi$ is Hadamard differentiable at $\theta$ tangentially to $\mathfrak{D}_0$.*

## 3. Unified Framework for Statistical Inference

This section develops a general framework for deriving limit distributions, bootstrap consistency, and semiparametric efficiency bounds for regularized OT distances. We first treat the former two aspects together, and then move on to discuss efficiency. Throughout this section, $\mu_n$ designates an arbitrary random probability measure and not necessarily the empirical measure (unless explicitly stated otherwise).

### 3.1. *Limit distributions and bootstrap consistency*

The following result is an adaptation of the extended functional delta method from Lemma 1 to the space of probability measures, which enables directly applying it to empirical regularized OT.

**Proposition 1** (Limit distributions) *Consider the setting:*

> ***(Setting $\circledast$)*** *Let $\mathscr{F}$ be a class of Borel measurable functions on a topological space $S$ with a finite envelope $F$. For a given $\mu \in \mathscr{P}(S)$, let $\delta$ be a map from $\mathscr{P}_0 \subset \mathscr{P}(S)$ into a Banach space $(\mathfrak{E}, \|\cdot\|_{\mathfrak{E}})$, where $\mathscr{P}_0$ is a convex subset such that $\mu \in \mathscr{P}_0$ and $\int F d\nu < \infty$ for all $\nu \in \mathscr{P}_0$.*

*Further suppose that*

*(a)* *$\mu_n : \Omega \to \mathscr{P}_0$ are random probability measures with values in $\mathscr{P}_0$ for all $n \in \mathbb{N}$, such that there exists a tight random variable $G_\mu$ in $\ell^\infty(\mathscr{F})$ with $\sqrt{n}(\mu_n - \mu) \xrightarrow{d} G_\mu$ in $\ell^\infty(\mathscr{F})$;*

*(b)* *$\delta$ is locally Lipschitz continuous at $\mu$ with respect to $\|\cdot\|_{\infty,\mathscr{F}}$, in the sense that there exist constants $\varepsilon > 0$ and $C < \infty$ such that*

$$\|\nu - \mu\|_{\infty,\mathscr{F}} \vee \|\nu' - \mu\|_{\infty,\mathscr{F}} < \varepsilon \implies \|\delta(\nu) - \delta(\nu')\|_{\mathfrak{E}} \leq C\|\nu - \nu'\|_{\infty,\mathscr{F}};$$

*(c)* *For every $\nu \in \mathscr{P}_0$, the mapping $t \mapsto \delta\big(\mu + t(\nu - \mu)\big)$ is right differentiable at $t = 0$, and denote its right derivative by*

$$\delta'_\mu(\nu - \mu) = \lim_{t \downarrow 0} \frac{\delta\big(\mu + t(\nu - \mu)\big) - \delta(\mu)}{t}. \tag{3.1}$$

*Then (i) $\delta'_\mu$ uniquely extends to a continuous, positively homogeneous map on the tangent cone of $\mathscr{P}_0$ at $\mu$:*

$$\mathscr{T}_{\mathscr{P}_0}(\mu) := \overline{\{t(\nu-\mu) : \nu \in \mathscr{P}_0, t>0\}}^{\ell^\infty(\mathscr{F})};$$

*(ii) $G_\mu \in T_{\mathscr{P}_0}(\mu)$ almost surely (a.s.); and (iii) $\sqrt{n}\big(\delta(\mu_n)-\delta(\mu)\big) - \delta'_\mu\big(\sqrt{n}(\mu_n-\mu)\big) \to 0$ holds in outer probability. Consequently, we have the following convergence in distribution $\sqrt{n}\big(\delta(\mu_n) - \delta(\mu)\big) \xrightarrow{d} \delta'_\mu(G_\mu)$.*

The proof first identifies $\delta$ as a map defined on a subset of $\ell^\infty(\mathscr{F})$. Formally, let $\tau : \mathscr{P}_0 \ni \nu \mapsto (f \mapsto \nu(f)) \in \ell^\infty(\mathscr{F})$, and we identify $\delta$ with $\bar{\delta} : \tau\mathscr{P}_{0,\varepsilon} \to \mathfrak{E}$ defined by $\bar{\delta}(\tau\nu) = \delta(\nu)$, where $\mathscr{P}_{0,\varepsilon} = \{\nu \in \mathscr{P}_0 : \|\nu-\mu\|_{\infty,\mathscr{F}} < \varepsilon\}$. The local Lipschitz condition (b) guarantees that the map $\bar{\delta}$ is well-defined (indeed, without the local Lipschitz condition, $\bar{\delta}$ may not be well-defined as $\tau$ may fail to be one-to-one). With this identification, we apply the extended functional delta method, Lemma 1, by establishing Hadamard directional differentiability of $\delta$ at $\mu$. The latter essentially follows by local Lipschitz continuity (condition (b)) and Gâteaux directional differentiability (condition (c)). Since the derivative $\delta'_\mu$ is a priori defined only on $\mathscr{P}_{0,\varepsilon} - \mu$, we need to extend the derivative to the tangent cone $\mathscr{T}_{\mathscr{P}_0}(\mu)$, for which we need completeness of the space $\mathfrak{E}$; see the proof in **??** for details.

For i.i.d. data $X_1,\dots,X_n \sim \mu$ and $\mu_n = \hat{\mu}_n$ as the empirical measure, to apply Proposition 1 we will: (i) find a $\mu$-Donsker function class $\mathscr{F}$ such that the functional $\delta$ is locally Lipschitz w.r.t. $\|\cdot\|_{\infty,\mathscr{F}}$ at $\mu$; and (ii) find the Gâteaux directional derivative (3.1). In our applications to regularized OT, such a function class $\mathscr{F}$ will be chosen to contain dual potentials corresponding to a proper class of distributions. Regularization enforces dual potentials to possess certain smoothness or low-dimensionality properties, guaranteeing that $\mathscr{F}$ is indeed $\mu$-Donsker. The dual OT formulation also plays a crucial role in finding the Gâteaux directional derivative (3.1).

**Remark 2** (On Proposition 1) *We now clarify certain aspects of Proposition 1.*

***(Relaxed condition):*** *When $\delta(\mu_n)$ is well-defined, the condition that $\mu_n$ takes values in $\mathscr{P}_0$ can be relaxed to $\mu_n \in \mathscr{P}_0$ with inner probability approaching one.*

***(Data generating process):*** *Proposition 1 does not impose any dependence conditions on the data. In particular, it can be applied to dependent data as long as one can verify the uniform limit theorem in Condition (a). See, e.g., [2, 3, 5, 23, 33, 53, 70] on uniform CLTs for dependent data.*

***(Convexity of $\mathscr{P}_0$):*** *The assumption that $\mathscr{P}_0$ is convex can be replaced with the condition that $\mathscr{P}_0$ is a convex subset of $\ell^\infty(\mathscr{F})$. Namely, using the mapping $\tau : \mathscr{P}_0 \ni \nu \mapsto (f \mapsto \nu(f)) \in \ell^\infty(\mathscr{F})$, we only need that $\tau\mathscr{P}_0 = \{\tau\nu : \nu \in \mathscr{P}_0\} \subset \ell^\infty(\mathscr{F})$ is convex. Condition (c) then should read that $t \mapsto \bar{\delta}((1-t)\tau\mu + t\tau\nu)$ is differentiable from the right at $t=0$ with derivative $\delta'_\mu(\mu-\nu) = \lim_{t\downarrow 0} t^{-1}\big\{\bar{\delta}((1-t)\tau\mu + t\tau\nu) - \bar{\delta}(\tau\mu)\big\}$, where $\bar{\delta}(\tau\nu) = \delta(\nu)$. This modification is needed to cover the two-sample setting; see, e.g., the proof of Theorem 1 Part (ii).*

### 3.1.1. Bootstrap consistency

In applications of Proposition 1, the obtained limit distribution is often non-pivotal in the sense that it depends on the population distribution $\mu$, which is unknown in practice. To circumvent the difficulty of estimating the distribution of $\delta'_\mu(G_\mu)$ directly, one may apply the bootstrap. When $\mathscr{F}$ is $\mu$-Donsker and $\mu_n = \hat{\mu}_n$ is the empirical distribution of i.i.d. data from $\mu$, then the bootstrap (applied to the functional

$\delta$) is consistent for estimating the distribution of $\delta'_\mu(G_\mu)$ provided that the map $\nu \mapsto \delta(\nu)$ is Hadamard differentiable w.r.t. $\|\cdot\|_{\infty,\mathscr{F}}$ at $\nu = \mu$ tangentially to a subspace of $\ell^\infty(\mathscr{F})$ that contains the support of $G_\mu$; cf. Theorem 23.9 in [93] or Theorem 3.9.11 in [94]. The following corollary is useful for invoking such theorems under the setting of Proposition 1.

**Corollary 1** (Bootstrap consistency via Hadamard differentiability) *Consider the setting of Proposition 1. If, in addition, $G_\mu$ is a mean-zero Gaussian variable in $\ell^\infty(\mathscr{F})$, then* $\mathrm{spt}(G_\mu)$ *is a vector subspace of $\ell^\infty(\mathscr{F})$. If further $\delta'_\mu$ is linear on* $\mathrm{spt}(G_\mu)$*, then $\nu \mapsto \delta(\nu)$ is Hadamard differentiable w.r.t. $\|\cdot\|_{\infty,\mathscr{F}}$ at $\nu = \mu$ tangentially to* $\mathrm{spt}(G_\mu)$.

In general, when the functional is Hadamard directionally differentiable with a nonlinear derivative, the bootstrap fails to be consistent; cf. [34, 36]. An alternative way to estimate the limit distribution in such cases is to use subsampling or the "$m$-out-of-$n$" bootstrap [34, 74]; see Lemma 3 for the max-slicing case.

### 3.2. *Semiparametric efficiency*

In Proposition 1, if $\delta'_\mu$ is linear and $G_\mu$ is mean-zero Gaussian, then the limit distribution $\delta'_\mu(G_\mu)$ is mean-zero Gaussian as well. In such cases, it is natural to ask if the plug-in estimator $\delta(\mu_n)$ is asymptotically efficient in the sense of [93, p. 367], relative to a certain tangent space. Informally, the semiparametric efficiency bound at $\mu$ is computed as the largest Cramér-Rao lower bound among one-dimensional submodels passing through $\mu$.

Formally, consider estimating a functional $\kappa : \mathscr{P} \subset \mathscr{P}(S) \to \mathbb{R}$ at $\mu \in \mathscr{P}$ from i.i.d. data $X_1, \ldots, X_n \sim \mu$. We consider submodels $\{\mu_t : 0 \le t < \varepsilon'\}$ with $\mu_0 = \mu$ such that, for some measurable score function $h : S \to \mathbb{R}$, we have

$$\int \left[ \frac{d\mu_t^{1/2} - d\mu^{1/2}}{t} - \frac{1}{2} h d\mu^{1/2} \right]^2 \to 0,$$

where $d\mu_t$ and $d\mu$ are Radon-Nikodym densities w.r.t. a common dominating measure and the integration is taken w.r.t. the dominating measure. Score functions are square integrable w.r.t. $\mu$ and $\mu$-mean zero. A *tangent set* $\dot{\mathscr{P}}_\mu \subset L^2(\mu)$ of the model $\mathscr{P}$ at $\mu$ is the set of score functions corresponding to a collection of such submodels. If $\dot{\mathscr{P}}_\mu$ is a vector subspace of $L^2(\mu)$, then it is called a *tangent space*. Relative to a given tangent set $\dot{\mathscr{P}}_\mu$, the functional $\kappa : \mathscr{P} \to \mathbb{R}$ is called *differentiable* at $\mu$ if there exists a continuous linear functional $\dot{\kappa}_\mu : L^2(\mu) \to \mathbb{R}$ such that, for every $h \in \dot{\mathscr{P}}_\mu$ and a submodel $t \mapsto \mu_t$ with score function $h$,

$$\frac{\kappa(\mu_t) - \kappa(\mu)}{t} \to \dot{\kappa}_\mu h, \quad t \downarrow 0.$$

The semiparametric efficiency bound for estimating $\kappa$ at $\mu$, relative to $\dot{\mathscr{P}}_\mu$, is defined as

$$\sigma^2_{\kappa,\mu} = \sup_{h \in \mathrm{lin}(\dot{\mathscr{P}}_\mu)} \frac{(\dot{\kappa}_\mu h)^2}{\|h\|^2_{L^2(\mu)}},$$

where $\mathrm{lin}(\dot{\mathscr{P}}_\mu)$ is the linear span of $\dot{\mathscr{P}}_\mu$. In particular, the $N(0, \sigma^2_{\kappa,\mu})$ distribution serves as the "optimal" limit distribution for estimating $\kappa$ at $\mu$ in the sense of the Hájek-Le Cam convolution theorem and also in the local asymptotic minimax sense; see Chapter 25 in [93] for details.

The next proposition concerns the computation of the semiparametric efficiency bound.

**Proposition 2** (Semiparametric efficiency) *For Setting $\circledast$ from Proposition 1 with $\mathfrak{E} = \mathbb{R}$, consider estimating $\delta : \mathscr{P}_0 \to \mathbb{R}$ at $\mu$ from i.i.d. data $X_1, \dots, X_n \sim \mu$. Set*

$$\dot{\mathscr{P}}_{0,\mu} = \{h : h : S \to \mathbb{R} \text{ is bounded and measurable with } \mu\text{-mean zero}\}.$$

*Suppose that (a) the function class $\mathscr{F}$ is $\mu$-pre-Gaussian, i.e., there exists a tight mean-zero Gaussian process $G_\mu = \big(G_\mu(f)\big)_{f \in \mathscr{F}}$ in $\ell^\infty(\mathscr{F})$ with covariance function $\operatorname{Cov}\big(G_\mu(f), G_\mu(g)\big) = \operatorname{Cov}_\mu(f,g)$; (b) for every $h \in \dot{\mathscr{P}}_{0,\mu}$, $(1+th)\mu \in \mathscr{P}_0$ for sufficiently small $t > 0$; and (c) there exists a continuous linear functional $\delta'_\mu : \ell^\infty(\mathscr{F}) \to \mathbb{R}$ such that (3.1) holds for every $\nu \in \mathscr{P}_0$ of the form $\nu = (1+h)\mu$ for some $h \in \dot{\mathscr{P}}_{0,\mu}$. Then, the semiparametric efficiency bound for estimating $\delta$ at $\mu$ relative to the tangent space $\dot{\mathscr{P}}_{0,\mu}$ agrees with $\operatorname{Var}\big(\delta'_\mu(G_\mu)\big)$.*

Proposition 2 can be thought of as a variant of Theorem 3.1 in [91], which asserts that a Hadamard differentiable functional (tangentially to a sufficiently large subspace) of an asymptotically efficient estimator is again asymptotically efficient; see Remark **??** for more details. In **??**, we provide a direct and self-contained proof of Proposition 2. We note that Proposition 2 covers a slightly more general situation than [91, Theorem 3.1] since it only requires Gâteaux differentiability of the map $\delta$, and choosing a pre-Gaussian function class $\mathscr{F}$ in such a way that the derivative $\delta'_\mu$ extends to a continuous linear functional on $\ell^\infty(\mathscr{F})$. In particular, the efficiency bound computation in Proposition 2 is applicable even when Proposition 1 is difficult to apply. For instance, when the Gâteaux derivative $\delta'_\mu$ in (3.1) is a point evaluation, $\delta'_\mu(\nu - \mu) = (\nu - \mu)(f^\star)$ for some function $f^\star \in L^2(\mu)$, we can choose $\mathscr{F} = \{f^\star\}$ (singleton) and apply Proposition 2 to conclude that $\operatorname{Var}_\mu(f^\star)$ agrees with the semiparametric efficiency bound, relative to $\dot{\mathscr{P}}_{0,\mu}$ (note that the function class $\mathscr{F}$ in Proposition 2 need not be the same as the one in Proposition 1).

The following corollary covers the two-sample case. Define $\dot{\mathscr{P}}_{0,\nu}$ analogously to $\dot{\mathscr{P}}_{0,\mu}$ and set $\dot{\mathscr{P}}_{0,\mu} \oplus \dot{\mathscr{P}}_{0,\nu} = \{h_1 \oplus h_2 : h_1 \in \dot{\mathscr{P}}_{0,\mu}, h_2 \in \dot{\mathscr{P}}_{0,\nu}\}$.

**Corollary 2** (Semiparametric efficiency in two-sample setting) *Let $\mathscr{F}$ be a class of Borel measurable functions on a topological space $S$ with finite envelope $F$, and for given $\mu, \nu \in \mathscr{P}(S)$, let $\mathscr{P}_{0,\mu}, \mathscr{P}_{0,\nu}$ be subsets of $\mathscr{P}(S)$ containing $\mu, \nu$, respectively, such that $\int F d\rho < \infty$ for all $\rho \in \mathscr{P}_{0,\mu} \cup \mathscr{P}_{0,\nu}$. Let $\mathscr{P}_{0,\mu} \otimes \mathscr{P}_{0,\nu} = \{\rho_1 \otimes \rho_2 : \rho_1 \in \mathscr{P}_{0,\mu}, \rho_2 \in \mathscr{P}_{0,\nu}\}$. Consider estimating $\delta : \mathscr{P}_{0,\mu} \otimes \mathscr{P}_{0,\nu} \to \mathbb{R}$ at $\mu \otimes \nu$ from i.i.d. data $(X_1, Y_1), \dots, (X_n, Y_n) \sim \mu \otimes \nu$. Suppose that (a) the function class $\mathscr{F}$ is pre-Gaussian w.r.t. $\mu$ and $\nu$; (b) for every $h_1 \oplus h_2 \in \dot{\mathscr{P}}_{0,\mu} \oplus \dot{\mathscr{P}}_{0,\nu}$, $\big((1+th_1)\mu\big) \otimes \big((1+th_2)\nu\big) \in \mathscr{P}_{0,\mu} \otimes \mathscr{P}_{0,\nu}$ for sufficiently small $t > 0$; (c) there exist continuous linear functionals $\delta'_\mu : \ell^\infty(\mathscr{F}) \to \mathbb{R}$ and $\delta'_\nu : \ell^\infty(\mathscr{F}) \to \mathbb{R}$ such that $t^{-1}\big\{\delta\big(\big((1+th_1)\mu\big) \otimes \big((1+th_2)\nu\big)\big) - \delta(\mu \otimes \nu)\big\} \to \delta'_\mu(h_1\mu) + \delta'_\nu(h_2\nu)$ as $t \downarrow 0$ for every $h_1 \oplus h_2 \in \dot{\mathscr{P}}_{0,\mu} \oplus \dot{\mathscr{P}}_{0,\nu}$. Then, the semiparametric efficiency bound for estimating $\delta$ at $\mu \otimes \nu$ relative to the tangent space $\dot{\mathscr{P}}_{0,\mu} \oplus \dot{\mathscr{P}}_{0,\nu}$ agrees with $\operatorname{Var}\big(\delta'_\mu(G_\mu)\big) + \operatorname{Var}\big(\delta'_\mu(G_\nu)\big)$, where $G_\mu$ and $G_\nu$ are tight $\mu$- and $\nu$-Brownian bridges in $\ell^\infty(\mathscr{F})$, respectively.*

**Remark 3** (Efficiency of wavelet-based estimator of $\mathsf{W}_2$) *Theorem 18 in [58] establishes a CLT for a wavelet-based estimator $\mathsf{W}_2(\tilde{\mu}_n, \nu)$ for $\mathsf{W}_2(\mu, \nu)$ in the one-sample case under high-level assumptions that include global regularity of the OT potential $\varphi$. Their result reads as $\sqrt{n}\big(\mathsf{W}_2^2(\tilde{\mu}_n, \nu) - \mathsf{W}_2^2(\mu, \nu)\big) \xrightarrow{d} N\big(0, \operatorname{Var}_\mu(\varphi)\big)$. Their proof first establishes a CLT for the expectation centering (similarly to [24]) and* *then shows that the bias is negligible. To obtain the same result via Proposition 1, one would have to assume a uniform bound on the Hölder norm of OT potentials corresponding to a local neighborhood of $\mu$. Unfortunately, such uniform bounds on global regularity of OT potentials are currently unavailable, except for a few limited cases (cf. the discussion after Theorem 3 in [58]). For instance, when the marginals are defined on the flat torus and admit sufficiently smooth densities that are bounded away from* 0 *and* $\infty$*, Theorem 5 in [58] provides such a bound, rendering Proposition 1 applicable. Moreover, in the setting of Theorem 18 in [58], it is readily verified that* $\rho \mapsto \mathsf{W}_2(\rho,\nu)$ *is Gâteaux differentiable at* $\mu$ *with derivative* $(\rho-\mu)(\varphi)$ *(cf. the proof of Lemma* **??***), so* $\mathrm{Var}_\mu(\varphi)$ *indeed coincides with the semiparametric efficiency bound.*

## 4. Sliced Wasserstein distances

This section studies statistical aspects of sliced Wasserstein distances, deriving limit distributions, bootstrap consistency, and semiparametric efficiency.

### 4.1. *Background*

Average- and max-sliced Wasserstein distances are defined next.

**Definition 2** (Sliced Wasserstein distances) *Let* $1 \le p < \infty$*. The average-sliced and max-sliced p-Wasserstein distances between* $\mu,\nu \in \mathscr{P}_p(\mathbb{R}^d)$ *are defined, respectively, as*

$$\underline{\mathsf{W}}_p(\mu,\nu) := \left[\int_{\mathbb{S}^{d-1}} \mathsf{W}_p^p(\mathfrak{p}^\theta_\sharp\mu,\mathfrak{p}^\theta_\sharp\nu)d\sigma(\theta)\right]^{1/p} \text{ and } \overline{\mathsf{W}}_p(\mu,\nu) := \max_{\theta\in\mathbb{S}^{d-1}} \mathsf{W}_p(\mathfrak{p}^\theta_\sharp\mu,\mathfrak{p}^\theta_\sharp\nu),$$

*where* $\mathfrak{p}^\theta : \mathbb{R}^d \to \mathbb{R}$ *is the projection map* $x \mapsto \theta^\intercal x$*,* $\sigma$ *is the uniform distribution on the unit sphere* $\mathbb{S}^{d-1} := \{x \in \mathbb{R}^d : \|x\| = 1\}$*, and* $\mathfrak{p}^\theta_\sharp\mu := \mu\circ(\mathfrak{p}^\theta)^{-1}$ *is the pushforward of* $\mu$ *under* $\mathfrak{p}^\theta$*.*

The sliced distances $\underline{\mathsf{W}}_p$ and $\overline{\mathsf{W}}_p$ are metrics on $\mathscr{P}_p(\mathbb{R}^d)$ and, in fact, induce the same topology as $\mathsf{W}_p$ [6]. Sliced Wasserstein distances are efficiently computable using the closed-form expression for $\mathsf{W}_p$ between distributions on $\mathbb{R}$ using quantile functions. For $\mu \in \mathscr{P}(\mathbb{R}^d)$ and $\theta \in \mathbb{S}^{d-1}$, denote by $F_\mu(\cdot\,;\theta)$ and $F_\mu^{-1}(\cdot\,;\theta)$ the distribution and quantile functions of $\mathfrak{p}^\theta_\sharp\mu$, respectively, i.e.,

$$F_\mu(t;\theta) = \mu\big(\{x\in\mathbb{R}^d : \theta^\intercal x \le t\}\big) \quad \text{and} \quad F_\mu^{-1}(\tau;\theta) = \inf\{t\in\mathbb{R} : F_\mu(t;\theta)\ge\tau\}.$$

Then, $\mathsf{W}_p(\mathfrak{p}^\theta_\sharp\mu,\mathfrak{p}^\theta_\sharp\nu)$ equals the $L^p$-norm between the corresponding quantile functions,

$$\mathsf{W}_p^p(\mathfrak{p}^\theta_\sharp\mu,\mathfrak{p}^\theta_\sharp\nu) = \int_0^1 \left|F_\mu^{-1}(\tau;\theta) - F_\nu^{-1}(\tau;\theta)\right|^p d\tau,$$

which further simplifies for $p = 1$ to the $L^1$ distance between the corresponding distribution functions. Also, sliced Wasserstein distances between projected empirical distributions is readily computed using order statistics. Let $\hat\mu_n := n^{-1}\sum_{i=1}\delta_{X_i}$ and $\hat\nu_n := n^{-1}\sum_{i=1}\delta_{Y_i}$ be the empirical distributions of $X_1,\dots,X_n$ and $Y_1,\dots,Y_n$. For each $\theta \in \mathbb{S}^{d-1}$, let $X_i(\theta) = \theta^\intercal X_i$, and let $X_{(1)}(\theta) \le \cdots \le X_{(n)}(\theta)$ be the order statistics. Define $Y_{(1)}(\theta) \le \cdots \le Y_{(n)}(\theta)$ analogously. By Lemma 4.2 in [13], we have $\mathsf{W}_p^p(\mathfrak{p}^\theta_\sharp\hat\mu_n,\mathfrak{p}^\theta_\sharp\hat\nu_n) = \frac{1}{n}\sum_{i=1}^n \left|X_{(i)}(\theta) - Y_{(i)}(\theta)\right|^p$. The sliced distances $\underline{\mathsf{W}}_p$ and $\overline{\mathsf{W}}_p$ can be computed by integrating or maximizing the above over $\theta \in \mathbb{S}^{d-1}$.

#### 4.1.1. Literature review

Sliced Wasserstein distances have been applied to various statistical inference and machine learning tasks, including barycenter computation [75], generative modeling [31, 32, 66, 67], and autoencoders [51]. The statistical literature on sliced distances mostly focused on expected value analysis. Specifically, [69] show that if $\mu$ satisfies a $T_q(\sigma^2)$ inequality with $q \in [1,2]$, then $\mathbb{E}\big[\overline{\mathsf{W}}_p(\hat{\mu}_n,\mu)\big] \lesssim \sigma\big(n^{-1/(2p)} + n^{(1/q-1/p)_+}\sqrt{(d\log n)/n}\big)$ up to a constant that depends only on $p$. Further results on empirical convergence rates can be found in [55], where both $\underline{\mathsf{W}}_p$ and $\overline{\mathsf{W}}_p$ were treated, while replacing the transport inequality assumption of [69] with exponential moment bounds (via Bernstein's tail conditions) or Poincaré type inequalities. A limit distribution result for one-sample sliced $\mathsf{W}_1$ was mentioned in [67] but was left as an unproven assumption. Extensions to sliced $\mathsf{W}_p$ and two-sample results, all of which are crucial for principled statistical inference, are currently open. Consistency of the bootstrap and efficiency bounds are also unaccounted for by the existing literature.[2]

### 4.2. *Statistical analysis*

We move on to the statistical aspects of sliced $\mathsf{W}_p$, closing the aforementioned gaps. The $p>1$ case is treated under the general framework of Section 3 for compactly supported distributions. For $p=1$, we present a separate derivation that leverages its simplified form to obtain the results under mild moment assumptions.

#### 4.2.1. Order $p>1$

The next theorem characterizes limit distributions for average-sliced $p$-Wasserstein distances under both the one- and two-sample settings. It also states asymptotic efficiency of the empirical plug-in estimator, and consistency of the bootstrap. The latter facilitates statistical inference by providing a tractable estimate of the limiting distribution, and is set up as follows. Given the data $X_1,\dots,X_n$, let $X_1^B,\dots,X_n^B$ be an independent sample from $\hat{\mu}_n$, and set $\hat{\mu}_n^B := n^{-1}\sum_{i=1}^n \delta_{X_i^B}$ as the bootstrap empirical distribution. Define $\hat{\nu}_n^B$ analogously and let $\mathbb{P}^B$ denote the conditional probability given the data.

**Theorem 1** (Limit distribution, efficiency, and bootstrap consistency for $\underline{\mathsf{W}}_p^p$) *Let $1<p<\infty$, and suppose that $\mu,\nu$ are compactly supported, such that $\mu$ is absolutely continuous and* $\operatorname{spt}(\mu)$ *is convex. For every $\theta\in\mathbb{S}^{d-1}$, let $\varphi^\theta$ be an OT potential from $\mathfrak{p}^\theta_\sharp\mu$ to $\mathfrak{p}^\theta_\sharp\nu$ for $\mathsf{W}_p$, which is unique up to additive constants on* $\operatorname{int}(\operatorname{spt}(\mathfrak{p}^\theta_\sharp\mu))$. *Also, set $\psi^\theta=[\varphi^\theta]^c$ as the c-transform of $\varphi^\theta$ for $c(s,t)=|s-t|^p$. The following hold.*

*(i)* *We have*

$$\sqrt{n}\big(\underline{\mathsf{W}}_p^p(\hat{\mu}_n,\nu)-\underline{\mathsf{W}}_p^p(\mu,\nu)\big)\xrightarrow{d} N\left(0,v_p^2\right),$$

*where $v_p^2=\iint \operatorname{Cov}_\mu\big(\varphi^\theta\circ\mathfrak{p}^\theta,\varphi^\vartheta\circ\mathfrak{p}^\vartheta\big)d\sigma(\theta)d\sigma(\vartheta)$, which is well-defined under the current assumption. The asymptotic variance $v_p^2$ coincides with the semiparametric efficiency bound for*

[2] After the first version of the present paper was posted on the arXiv, we became aware that the latest update of [59] (arXiv update: April 4, 2022) contains limit distribution and bootstrap results for $\underline{\mathsf{W}}_p$ with $p>1$ under the alternative. Our work is independent of [59] and our approach is distinct; see Remark 5. After our paper was posted, [100] proved similar results to ours for $\underline{\mathsf{W}}_1$ and $\overline{\mathsf{W}}_1$ under a similar set of assumptions to us (see Remark 8 for details) and [99] derived limit distributions for $p>1$ uniformly in the slicing direction under the assumption of compact support and uniqueness of optimal potentials, but did not cover the max-slicing case in full generality. Neither of these works addressed asymptotic efficiency.

*estimating $\underline{\mathsf{W}}_p^p(\cdot,\nu)$ at $\mu$. Also, provided that $v_p^2 > 0$, we have*

$$\sup_{t\in\mathbb{R}}\left|\mathbb{P}^B\left(\sqrt{n}\big(\underline{\mathsf{W}}_p^p(\hat{\mu}_n^B,\nu)-\underline{\mathsf{W}}_p^p(\hat{\mu}_n,\nu)\big)\le t\right)-\mathbb{P}\big(N(0,v_p^2)\le t\big)\right| \xrightarrow{\mathbb{P}} 0.$$

*(ii) If in addition $\nu$ is absolutely continuous with convex support, then*

$$\sqrt{n}\big(\underline{\mathsf{W}}_p^p(\hat{\mu}_n,\hat{\nu}_n)-\underline{\mathsf{W}}_p^p(\mu,\nu)\big) \xrightarrow{d} N\left(0,v_p^2+w_p^2\right),$$

*where $v_p^2$ is given in (i) and $w_p^2=\iint \mathrm{Cov}_\nu\left(\psi^\theta\circ\mathfrak{p}^\theta,\psi^\vartheta\circ\mathfrak{p}^\vartheta\right)d\sigma(\theta)d\sigma(\vartheta)$. The asymptotic variance $v_p^2+w_p^2$ coincides with the semiparametric efficiency bound for estimating $\underline{\mathsf{W}}_p^p(\cdot,\cdot)$ at $(\mu,\nu)$. Also, provided that $v_p^2+w_p^2>0$, we have*

$$\sup_{t\in\mathbb{R}}\left|\mathbb{P}^B\left(\sqrt{n}\big(\underline{\mathsf{W}}_p^p(\hat{\mu}_n^B,\hat{\nu}_n^B)-\underline{\mathsf{W}}_p^p(\hat{\mu}_n,\hat{\nu}_n)\big)\le t\right)-\mathbb{P}\big(N(0,v_p^2+w_p^2)\le t\big)\right| \xrightarrow{\mathbb{P}} 0.$$

The derivation of the limit distributions in Theorem 1 follows from Proposition 1. We outline the main idea for the one-sample case. The functional of interest is set as the $p$th power of the average-sliced $p$-Wasserstein distance. Leveraging compactness of supports, we then show that $\underline{\mathsf{W}}_p^p$ is Lipschitz w.r.t. $\overline{\mathsf{W}}_1$ (cf. Lemma **??**). From the Kantorovich-Rubinstein duality, $\overline{\mathsf{W}}_1$ can be expressed as $\overline{\mathsf{W}}_1(\mu,\nu)=\|\mu-\nu\|_{\infty,\mathscr{F}}$ with $\mathscr{F}=\{\varphi\circ\mathfrak{p}^\theta:\theta\in\mathbb{S}^{d-1},\varphi\in\mathsf{Lip}_{1,0}(\mathbb{R})\}$, which is shown to be $\mu$-Donsker ($\mathsf{Lip}_{1,0}(\mathbb{R})$ denotes the class of 1-Lipschitz functions $\varphi$ on $\mathbb{R}$ with $\varphi(0)=0$). Evaluating the Gâteaux directional derivative of the sliced distance, we have all the conditions needed to invoke Proposition 1, which in turn yields the distributional limits. For $\underline{\mathsf{W}}_p$, the corresponding derivative turns out to be linear (in a suitable sense), so that asymptotic efficiency of the plug-in estimator and the bootstrap consistency follow from Proposition 2 and Corollary 1 combined with Theorem 23.9 in [93].

For the two-sample case, we think of $\underline{\mathsf{W}}_p(\mu,\nu)$ as a functional of the product measure $\mu\otimes\nu$, as the correspondence between $(\mu,\nu)$ and $\mu\otimes\nu$ is one-to-one. With this identification, the rest of the argument is analogous to the one-sample case. We note that in the two-sample case, the semiparametric efficiency bound is defined relative to the tangent space

$$\big\{h_1\oplus h_2 : h_1 \text{ and } h_2 \text{ are bounded measurable functions with } \mu(h_1)=\nu(h_2)=0\big\}.$$

This convention is adopted throughout when discussing semiparametric efficiency bounds in the two-sample case.

The asymptotic variances in Theorem 1 involve potentials between all slices of the marginal distributions, so direct estimation of the asymptotic variances seems highly nontrivial from a computational standpoint. Hence, the bootstrap offers a particularly appealing alternative for estimating the sampling distributions of empirical sliced Wasserstein distances.

**Remark 4** (Removing $p$th power) *While Theorem 1 states limit distributions for the pth power of $\underline{\mathsf{W}}_p$, we can readily obtain corresponding results for the average-sliced p-Wasserstein distance itself by invoking the delta method for the map $s\mapsto s^{1/p}$.*

**Remark 5** (Comparison with [59]) *Theorem 4 in [59] derives limit distributions and bootstrap consistency for empirical $\underline{\mathsf{W}}_p$ under the alternative, subject to the assumption that the projected densities $\{f_\mu(\cdot;\theta)\}_{\theta\in\mathbb{S}^{d-1}}$ and $\{f_\nu(\cdot;\theta)\}_{\theta\in\mathbb{S}^{d-1}}$ are uniformly integrable with*

$$\sup_{\theta\in\mathbb{S}^{d-1}} \operatorname{esssup}_{0<u<1} \frac{1}{f_\mu(F_\nu^{-1}(t;\theta);\theta)} \vee \frac{1}{f_\nu(F_\nu^{-1}(t;\theta);\theta)} < \infty.$$

*Here $f_\mu(\cdot;\theta)$ and $F_\mu^{-1}(\cdot;\theta)$ are the (Lebesgue) density and quantile function of $\mathfrak{p}^\theta_\sharp\mu$, and their composition is the so-called I-function; cf. [13, Equation (5.2)]. Verification of this condition for given distributions seems nontrivial. The proof of [59, Theorem 4] exploits the quantile function representation of $\mathsf{W}_p$ in $d=1$ along with a linearization step (of quantile functions). Our limit theorem, on the other hand, assumes that $\mu$ has a density with compact and convex support, and employs a markedly different proof via the general framework of Proposition 1.*

We next provide one- and two-sample limit distributions for the max-sliced Wasserstein distance. In this case, the Hadamard directional derivative is nonlinear and therefore the limit is non-Gaussian and the nonparametric bootstrap is inconsistent (cf. [34, 36]).

**Theorem 2** (Limit distribution for $\overline{\mathsf{W}}_p$) *Consider the assumption of Theorem 1.*

*(i)* *Setting $\mathfrak{S}_{\mu,\nu} := \{\theta\in\mathbb{S}^{d-1} : \mathsf{W}_p(\mathfrak{p}^\theta_\sharp\mu,\mathfrak{p}^\theta_\sharp\nu) = \overline{\mathsf{W}}_p(\mu,\nu)\}$, we have*

$$\sqrt{n}\big(\overline{\mathsf{W}}_p^p(\hat{\mu}_n,\nu) - \overline{\mathsf{W}}_p^p(\mu,\nu)\big) \xrightarrow{d} \sup_{\theta\in\mathfrak{S}_{\mu,\nu}} \mathbb{G}_\mu(\theta),$$

*where $\big(\mathbb{G}_\mu(\theta)\big)_{\theta\in\mathbb{S}^{d-1}}$ is a centered Gaussian process with continuous paths and covariance function $\operatorname{Cov}\big(\mathbb{G}_\mu(\theta),\mathbb{G}_\mu(\vartheta)\big) = \operatorname{Cov}_\mu\big(\varphi^\theta\circ\mathfrak{p}^\theta, \varphi^\vartheta\circ\mathfrak{p}^\vartheta\big)$, which is well-defined.*

*(ii)* *If in addition $\nu$ is also absolutely continuous with convex support, then*

$$\sqrt{n}\big(\overline{\mathsf{W}}_p^p(\hat{\mu}_n,\hat{\nu}_n) - \overline{\mathsf{W}}_p^p(\mu,\nu)\big) \xrightarrow{d} \sup_{\theta\in\mathfrak{S}_{\mu,\nu}} \big[\mathbb{G}_\mu(\theta) + \mathbb{G}'_\nu(\theta)\big], \tag{4.1}$$

*where $\big(\mathbb{G}'_\nu(\theta)\big)_{\theta\in\mathbb{S}^{d-1}}$ is independent of $\mathbb{G}_\mu$ given in (i) and defined analogously.*

Observe that, for $\mu,\nu\in\mathscr{P}_p(\mathbb{R}^d)$, the map $\theta\mapsto\mathsf{W}_p(\mathfrak{p}^\theta_\sharp\mu,\mathfrak{p}^\theta_\sharp\nu)$ is continuous, so the set $\mathfrak{S}_{\mu,\nu}$ is nonempty. The proof of Theorem 2 also relies on the general framework of Proposition 1. As in the average case, $\overline{\mathsf{W}}_p$ is Lipschitz w.r.t. $\overline{\mathsf{W}}_1$. This reduces the argument to characterizing the Gâteaux directional derivative, which requires extra work.

The nonlinearity of the Hadamard directional derivative means that the nonparametric bootstrap is inconsistent for $\overline{\mathsf{W}}_p$. Nevertheless, subsampling or the $m$-out-of-$n$ bootstrap can still consistently estimate the limit law. The next lemma deals with the $m$-out-of-$n$ bootstrap. Below, we say that the bootstrap quantity $S_n^B$ converges in distribution to a nonrandom law $\rho$ in probability if $\sup_{g\in\mathsf{BL}_1(\mathbb{R})}\big|\mathbb{E}^B[g(S_n^B)] - \mathbb{E}_{S\sim\rho}[g(S)]\big| \to 0$ in probability, where $\mathsf{BL}_1(\mathbb{R})$ denotes the class of (bounded) 1-Lipschitz functions $g:\mathbb{R}\to[-1,1]$ and $\mathbb{E}^B$ is the conditional expectation given the sample.

**Lemma 3** *Let $X_1^B,\dots,X_m^B$ and $Y_1^B,\dots,Y_m^B$ be independent samples from $\hat{\mu}_n$ and $\hat{\nu}_n$, respectively, where $m=m_n\to\infty$ with $m=o(n)$. Set $\hat{\mu}^B_{m,n}=m^{-1}\sum_{i=1}^m\delta_{X_i^B}$ and $\hat{\nu}^B_{m,n}=m^{-1}\sum_{i=1}^m\delta_{Y_i^B}$. Under the setting of Theorem 1 and conditionally on the data, the sequence $\sqrt{m}\big(\overline{\mathsf{W}}_p^p(\hat{\mu}^B_{m,n},\hat{\nu}^B_{m,n})-\overline{\mathsf{W}}_p^p(\hat{\mu}_n,\hat{\nu}_n)\big)$ converges in distribution to the limit in (4.1), in probability.*

**Remark 6** (Bias of plug-in estimator for $\overline{\mathsf{W}}_p$ and correction) *In general, the limit distributions for the max-sliced distance in Theorem 2 have positive means, which implies that empirical $\overline{\mathsf{W}}_p$ tends to be upward biased at the order of $n^{-1/2}$. Such an upward bias commonly appears in plug-in estimation of the maximum of a nonparametric function (cf. [19]). One may correct this bias using a precision correction similar to [19]. Namely, define $\widehat{\mathbb{W}}_p(\alpha):=\overline{\mathsf{W}}_p^p(\hat{\mu}_n,\hat{\nu}_n)-k_\alpha/\sqrt{n}$, where $k_\alpha$ is the $\alpha$-quantile of $\sup_{\theta\in\mathfrak{S}_{\mu,\nu}}\big[\mathbb{G}_\mu(\theta)+\mathbb{G}'_\nu(\theta)\big]$, which can be estimated via the subsampling or the m-out-of-n bootstrap. Provided that $k_\alpha$ is a continuity point of the distribution function of the limit variable, this estimator is upward $\alpha$-quantile unbiased, and in particular, median unbiased when $\alpha=1/2$, meaning that $\mathbb{P}\big(\widehat{\mathbb{W}}_p(\alpha)\le\overline{\mathsf{W}}_p^p(\mu,\nu)\big)=\alpha+o(1)$.*

**Remark 7** (Extensions) *We address two possible extensions of Theorems 1 and 2.*

***(Null case):*** *As $\varphi^\theta\circ\mathfrak{p}^\theta$ and $\psi^\theta\circ\mathfrak{p}^\theta$ are constant $\mu$- and $\nu$-a.e., respectively, for each $\theta\in\mathbb{S}^{d-1}$, the limit distributions in Theorems 1 and 2 degenerate to zero under the null, i.e., $\mu=\nu$. In $d=1$, [26] derived a limit distribution for $\mathsf{W}_2(\hat{\mu}_n,\mu)$ using the quantile function representation of $\mathsf{W}_2$, which requires several technical conditions concerning the tail of the Lebesgue density of $\mu$. Their argument hinges on approximating the (general) sample quantile process by the uniform quantile process, and applying results for the latter. This argument does not directly extend to the sliced distances, as the quantile process is indexed by the additional projection parameter $\theta\in\mathbb{S}^{d-1}$. Also, it seems nontrivial to find simple conditions on $\mu$ itself under which the projected distributions $\mathfrak{p}^\theta_\#\mu$ satisfy the conditions from [26] for all (or uniformly over) $\theta\in\mathbb{S}^{d-1}$. We leave null limit distributions for $\underline{\mathsf{W}}_p$ and $\overline{\mathsf{W}}_p$ for future research.*

***(Unbounded support):*** *For compactly supported distributions, dual potentials are Lipschitz continuous whose Lipschitz constants depend only on the order p and the radius of the support. This is a key result when we apply Proposition 1 to sliced Wasserstein distances; see also the discussion after Theorem 1. For distributions with unbounded supports, the recent work of [57] derives local Lipschitz estimates for dual potentials under a high-level anti-concentration assumption (see the discussion around their Lemma 11), but for empirical distributions, the local Lipschitz constants depend on the sample size n, which hinders the application of Proposition 1 (the function class being dependent on n does not bring any difficulty to finding error bounds but would require a very delicate argument for limit theorems). The extension of Theorems 1 and 2 to simple moment conditions would require highly technical arguments and hence is beyond the scope of the present paper.*

4.2.2. Order $p=1$

The analysis for $\underline{\mathsf{W}}_1$ relies on the explicit expression of $\mathsf{W}_1$ between distributions on $\mathbb{R}$ as the $L^1$ distance between distribution functions, whereby

$$\underline{\mathsf{W}}_1(\mu,\nu)=\int_{\mathbb{S}^{d-1}}\int_{\mathbb{R}}|F_\mu(t;\theta)-F_\nu(t;\theta)|\,dt\,d\sigma(\theta). \tag{4.2}$$

For $\overline{\mathsf{W}}_1$, the Kantorovich-Rubinstein duality yields

$$\overline{\mathsf{W}}_1(\mu,\nu) = \sup_{\theta\in\mathbb{S}^{d-1}} \sup_{\varphi\in\mathsf{Lip}_{1,0}(\mathbb{R})} \left[\int \varphi(\theta^\intercal x) d(\mu-\nu)(x)\right]. \tag{4.3}$$

Here $\mathsf{Lip}_{1,0}(\mathbb{R})$ denotes the class of 1-Lipschitz functions $\varphi$ on $\mathbb{R}$ with $\varphi(0)=0$. These explicit expressions enable us to derive a limit distribution theory for $\underline{\mathsf{W}}_1$ and $\overline{\mathsf{W}}_1$ under mild moment conditions, as presented next.

To state the result, set $\lambda$ as the Lebesgue measure on $\mathbb{R}$ and denote

$$\operatorname{sign}(t) = \mathbb{1}_{(0,\infty)}(t) - \mathbb{1}_{(-\infty,0)}(t).$$

Recall that a stochastic process $\big(Y(t)\big)_{t\in T}$ indexed by a measurable space $T$ is called measurable if $(t,\omega)\mapsto Y(t,\omega)$ is jointly measurable.

**Theorem 3** (Limit distribution for $\underline{\mathsf{W}}_1$ and $\overline{\mathsf{W}}_1$) *Let $\varepsilon>0$ be arbitrary.*

*(i) If $\mu\in\mathscr{P}_{2+\varepsilon}(\mathbb{R}^d)$ and $\nu\in\mathscr{P}_1(\mathbb{R}^d)$, then there exists a measurable, centered Gaussian process $\mathsf{G}_\mu = \big(\mathsf{G}_\mu(t,\theta)\big)_{(t,\theta)\in\mathbb{R}\times\mathbb{S}^{d-1}}$ with paths in $L^1(\lambda\otimes\sigma)$ and covariance function*

$$\operatorname{Cov}\big(\mathsf{G}_\mu(s,\theta),\mathsf{G}_\mu(t,\vartheta)\big) = \mu\big(\{x\in\mathbb{R}^d : \theta^\intercal x\le s, \vartheta^\intercal y\le t\}\big) - F_\mu(s;\theta)F_\mu(t;\vartheta), \tag{4.4}$$

*such that*

$$\sqrt{n}\big(\underline{\mathsf{W}}_1(\hat\mu_n,\nu)-\underline{\mathsf{W}}_1(\mu,\nu)\big) \xrightarrow{d} \iint \big[\operatorname{sign}(F_\mu-F_\nu)\big]\mathsf{G}_\mu d\lambda d\sigma + \iint_{F_\mu=F_\nu} |\mathsf{G}_\mu| d\lambda d\sigma. \tag{4.5}$$

*(ii) If $\mu,\nu\in\mathscr{P}_{2+\varepsilon}(\mathbb{R}^d)$, then*

$$\begin{aligned}&\sqrt{n}\big(\underline{\mathsf{W}}_1(\hat\mu_n,\hat\nu_n)-\underline{\mathsf{W}}_1(\mu,\nu)\big)\\ &\qquad \xrightarrow{d} \iint \big[\operatorname{sign}(F_\mu-F_\nu)\big](\mathsf{G}_\mu-\mathsf{G}'_\nu) d\lambda d\sigma + \iint_{F_\mu=F_\nu} |\mathsf{G}_\mu-\mathsf{G}'_\nu| d\lambda d\sigma,\end{aligned}$$

*where $\mathsf{G}'_\nu$ is independent of $\mathsf{G}_\mu$ given in $(i)$ and defined analogously.*

*(iii) Assume $\mu\in\mathscr{P}_{4+\varepsilon}(\mathbb{R}^d)$ and $\nu\in\mathscr{P}_1(\mathbb{R}^d)$. Consider the function class*

$$\mathscr{F} = \Big\{\varphi\circ\mathfrak{p}^\theta : \theta\in\mathbb{S}^{d-1}, \varphi\in\mathsf{Lip}_{1,0}(\mathbb{R})\Big\}.$$

*Then, there exists a tight $\mu$-Brownian bridge process $G_\mu$ in $\ell^\infty(\mathscr{F})$ such that*

$$\sqrt{n}\big(\overline{\mathsf{W}}_1(\hat\mu_n,\nu)-\overline{\mathsf{W}}_1(\mu,\nu)\big) \xrightarrow{d} \sup_{f\in M_{\mu,\nu}} G_\mu(f),$$

*where $M_{\mu,\nu} = \big\{f\in\overline{\mathscr{F}}^\mu : \mu(f-\nu(f)) = \overline{\mathsf{W}}_1(\mu,\nu)\big\}$ and $\overline{\mathscr{F}}^\mu$ is the completion of $\mathscr{F}$ for the standard deviation pseudometric $(f,g)\mapsto\sqrt{\operatorname{Var}_\mu(f-g)}$.*

*(iv)* *If $\mu,\nu \in \mathscr{P}_{4+\varepsilon}(\mathbb{R}^d)$, then there exists a tight $\nu$-Brownian bridge process $G'_\nu$ in $\ell^\infty(\mathscr{F})$ independent of $G_\mu$ above such that*

$$\sqrt{n}\big(\overline{\mathsf{W}}_1(\hat{\mu}_n,\hat{\nu}_n) - \overline{\mathsf{W}}_1(\mu,\nu)\big) \xrightarrow{d} \sup_{f \in M'_{\mu,\nu}} \big[G_\mu(f) - G'_\nu(f)\big],$$

*where $M'_{\mu,\nu} = \{f \in \overline{\mathscr{F}}^{\mu,\nu} : (\mu-\nu)(f) = \overline{\mathsf{W}}_1(\mu,\nu)\}$ and $\overline{\mathscr{F}}^{\mu,\nu}$ is the completion of $\mathscr{F}$ for the pseudometric $(f,g) \mapsto \sqrt{\mathrm{Var}_\mu(f-g)} + \sqrt{\mathrm{Var}_\nu(f-g)}$.*

While Theorem 1 requires distributions to have compact and convex support, Theorem 3 holds under mild moment assumptions. The derivation for $\underline{\mathsf{W}}_1$ uses the CLT in $L^1$ to deduce convergence of empirical projected distribution functions in $L^1(\lambda \otimes \sigma)$. The limit distribution is then obtained via the functional delta method by casting $\underline{\mathsf{W}}_1$ as the $L^1(\lambda \otimes \sigma)$ norm between distribution functions and characterizing the corresponding Hadamard directional derivative. For $\overline{\mathsf{W}}_1$, we use the Kantorovich-Rubinstein duality in conjunction with the fact that the class of projection 1-Lipschitz functions is Donsker under the said moment condition. In Part (iii), if $\mu = \nu$, then $M_{\mu,\nu} = \overline{\mathscr{F}}^{\mu}$, and since $G_\mu$ has uniformly continuous paths w.r.t. the standard deviation pseudometric, the limit variable becomes $\sup_{f \in \mathscr{F}} G_\mu(f)$. Likewise, in Part (iv), the limit variable becomes $\sup_{f \in \mathscr{F}}[G_\mu(f) - G'_\nu(f)]$ when $\mu = \nu$.

For $\underline{\mathsf{W}}_1$, if the second term on the right-hand side of (4.5) is zero, then the asymptotic normality holds. We state this result including its two-sample analogue next.

**Corollary 3** (Asymptotic normality for $\underline{\mathsf{W}}_1$) *For $\mu \in \mathscr{P}(\mathbb{R}^d)$ and $\theta \in \mathbb{S}^{d-1}$, define $\bar{l}^\theta_\mu = \sup \mathrm{spt}(\mathfrak{p}^\theta_\sharp \mu)$ and $\underline{l}^\theta_\mu = \inf \mathrm{spt}(\mathfrak{p}^\theta_\sharp \mu)$. The following hold.*

*(i)* *Under the assumption of Theorem 3 Part (i), if in addition $F_\mu(t;\theta) \neq F_\nu(t;\theta)$ for $(\lambda \otimes \sigma)$-almost all $(t,\theta) \in \big\{(s,\vartheta) : s \in [\underline{l}^\vartheta_\mu, \bar{l}^\vartheta_\mu], \vartheta \in \mathbb{S}^{d-1}\big\}$, then*

$$\sqrt{n}\big(\underline{\mathsf{W}}_1(\hat{\mu}_n,\nu) - \underline{\mathsf{W}}_1(\mu,\nu)\big) \xrightarrow{d} N(0,v_1^2),$$

*where $v_1^2$ is the variance of $\iint \big[\mathrm{sign}(F_\mu - F_\nu)\big]\mathsf{G}_\mu d\lambda d\sigma$. The asymptotic variance $v_1^2$ agrees with the semiparametric efficiency bound for estimating $\underline{\mathsf{W}}_1(\cdot,\nu)$ at $\mu$.*

*(ii)* *Under the assumption of Theorem 3 Part (ii), if in addition $F_\mu(t;\theta) \neq F_\nu(t;\theta)$ for $(\lambda \otimes \sigma)$-almost all $(t,\theta) \in \big\{(s,\vartheta) : s \in [\underline{l}^\vartheta_\mu \wedge \underline{l}^\vartheta_\nu, \bar{l}^\vartheta_\mu \vee \bar{l}^\vartheta_\nu], \vartheta \in \mathbb{S}^{d-1}\big\}$, then*

$$\sqrt{n}\big(\underline{\mathsf{W}}_1(\hat{\mu}_n,\hat{\nu}_n) - \underline{\mathsf{W}}_1(\mu,\nu)\big) \xrightarrow{d} N(0,v_1^2 + w_1^2),$$

*where $v_1^2$ is as above and $w_1^2$ is the variance of $\iint \big[\mathrm{sign}(F_\mu - F_\nu)\big]\mathsf{G}'_\nu d\lambda d\sigma$. The asymptotic variance $v_1^2 + w_1^2$ agrees with the semiparametric efficiency bound for estimating $\underline{\mathsf{W}}_1$ at $(\mu,\nu)$.*

*Finally, bootstrap consistency (as in Theorem 1) holds for both cases (i) and (ii).*

As an example, the above asymptotic normality holds when the population distributions are both Gaussian.

**Example 1** *Consider $\mu = N(\xi_1,\Sigma_1)$ and $\nu = N(\xi_2,\Sigma_2)$. Then $\mathfrak{p}^\theta_\sharp \mu = N(\theta^\intercal\xi_1, \theta^\intercal\Sigma_1\theta)$ and $\mathfrak{p}^\theta_\sharp \mu = N(\theta^\intercal\xi_2, \theta^\intercal\Sigma_2\theta)$. In this case, as long as $\xi_1 \neq \xi_2$ or $\Sigma_1 \neq \Sigma_2$, we have $\sigma(\{\theta : \theta^\intercal\xi_1 = \theta^\intercal\xi_2 \text{ and } \theta^\intercal\Sigma_1\theta = \theta^\intercal\Sigma_2\theta\}) = 0$, which follows from the fact that $\frac{Z}{\|Z\|} \sim \sigma$ for $Z \sim N(0,I_d)$ and Lemma 1 in [73]. Thus, $F_\mu(t;\theta) \neq F_\nu(t;\theta)$ for $(\lambda \otimes \sigma)$-almost all $(t,\theta) \in \mathbb{R} \times \mathbb{S}^{d-1}$ and the conclusion of Corollary 3 applies.*

**Remark 8** (Comparison with [100]) *The work [100], which appeared after our paper was posted on the arXiv, derived limit distribution results for $\underline{\mathsf{W}}_1$ and $\overline{\mathsf{W}}_1$ similar to the above. For the average-slicing in the one-sample case, they assume the slightly weaker moment condition $\int \sqrt{\mathbb{P}(\|X\| > t)}dt < \infty$, where $X \sim \mu$. Their proof strategy is essentially the same as ours and the above condition is imposed to verify a CLT for the empirical projected distribution function in $L^1(\lambda \otimes \sigma)$, as stated in the beginning of Step 1 in the proof of our Theorem 3(i). Our assumption, which requires finite $(2+\varepsilon)$-th moment, is meant to provide an elementary moment condition to guarantee the said CLT in $L^1(\lambda \otimes \sigma)$, and it is not difficult to see that the weaker (yet more high-level) condition $\int \sqrt{\mathbb{P}(\|X\| > t)}dt < \infty$ suffices for our proof to go through; see the discussion around equation (**??**). For the max-slicing case (with $p = 1$), [99] assume the exact same moment condition, although they do not cover the alternative case.*

## 5. Smooth Wasserstein Distance with Compactly Supported Kernels

We study smooth Wasserstein distances with compactly supported kernels, namely, when the considered distributions are convolved with a mollifier (also known as a bump function). Smoothing by means of the Gaussian kernel was extensively studied before for structural and statistical properties (see literature review below), but it remains unclear how to efficiently compute the Gaussian-smoothed Wasserstein distance.[3] Despite recent advancement in computation of continuous-to-continuous OT between distributions with smooth densities [65, 90], these approaches cannot handle the Gaussian-smoothed $\mathsf{W}_p$ since they assume compactly supported distributions.[4] Furnishing a smoothed framework that enjoys the structural and statistical virtues as the Gaussian-smoothed Wasserstein distance, while being amenable for efficient computation via the aforementioned approaches is the main motivation of this section. The main use case of compactly supported kernel paradigm is thus when the population distributions are also compactly supported. For the sake of generality, we next define the distance and provide structural properties for arbitrary $\mu,\nu \in \mathscr{P}(\mathbb{R}^d)$ distributions (possibly with unbounded support), but restrict to the compactly supported case for the statistical analysis. In **??**, we provide a thorough account of how to lift the algorithm from [90] to compute our smooth distance.

### 5.1. *Background*

To set up the smooth Wasserstein distance, we first define a smoothing kernel as follows. Let $\chi \in C^\infty(\mathbb{R}^d)$ be any non-negative function with $\int_{\mathbb{R}^d} \chi(x)dx = 1$ and $\int_{\mathbb{R}^d} \|x\|^p \chi(x)dx < \infty$, for all $1 \leq p < \infty$. Then, for any $\sigma > 0$, define $\chi_\sigma = \sigma^{-d}\chi(\cdot/\sigma) \in C^\infty(\mathbb{R}^d)$ and let $\eta_\sigma \in \mathscr{P}(\mathbb{R}^d)$ be a probability measure whose (Lebesgue) density is $\chi_\sigma$. We call $\eta_\sigma$ a *smoothing kernel* of parameter $\sigma$, and define the corresponding smooth Wasserstein distance as follows.

[3] While the empirical Gaussian-smoothed Wasserstein distance can be evaluated by sampling the kernel and applying computational methods for classic $\mathsf{W}_p$, this approach fails to exploit the smoothness of this framework and sacrifices the statistical advantages pertaining to estimation and inference.

[4] Convolution with a Gaussian kernel does not preserve compact support. In applications where compact support of the convolved distributions is immaterial, the Gaussian kernel is, however, a natural choice.

**Definition 3** (Smooth Wasserstein distances) *Let $1 \leq p < \infty$ and $\eta_\sigma$ be a smoothing kernel. The associated smooth p-Wasserstein distance between $\mu, \nu \in \mathscr{P}_p(\mathbb{R}^d)$ is*

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

**Example 2** (Standard mollifier) *A canonical example of a smooth compactly supported function is the standard mollifier*

$$\chi(x) = \begin{cases} \frac{1}{C_\chi}\exp\left(-\frac{1}{1-\|x\|^2}\right) & \text{if } \|x\| < 1 \\ 0 & \text{otherwise} \end{cases}, \tag{5.1}$$

*where $C_\chi = \int_{\mathbb{R}^d} \chi d\lambda$, from which a compactly supported kernel is readily constructed. Our results, however, are not specialized to the mollifier kernel and hold for any $\eta_\sigma$ as described above.*

As reviewed next, Gaussian-smoothed Wasserstein distances, i.e., when $\eta_\sigma = \gamma_\sigma := N(0,\sigma^2 I_d)$, have been extensively studied for their structural and statistical properties.

### 5.1.1. Literature review

Gaussian-smoothed Wasserstein distances were introduced in [43] as a means to mitigate the curse of dimensionality in empirical estimation. Indeed, [43] demonstrated that $\mathbb{E}\big[\mathsf{W}_p^{\gamma_\sigma}(\hat{\mu}_n,\mu)\big] = O(n^{-1/2})$, for $p = 1,2$, in arbitrary dimension provided that $\mu$ is sufficiently sub-Gaussian (cf. the recent preprint [11] for sharp bounds on the sub-Gaussian constant for which the rate is parametric when $p = 2$). Structural properties of $\mathsf{W}_1^{\gamma_\sigma}$ were explored in [40], showing that it metrizes the classic Wasserstein topology and establishing regularity in $\sigma$. These structural and statistical results were later generalized to $\mathsf{W}_p^{\gamma_\sigma}$ for any $p > 1$ [68], and asymptotics of the smooth distance as $\sigma \to \infty$ were explored [17]. Relations between $\mathsf{W}_p^{\gamma_\sigma}$ and maximum mean discrepancies were studies in [101], and nonparametric mixture model estimation under $\mathsf{W}_p^{\gamma_\sigma}$ was considered [47], again demonstrating scalability of error bounds with dimension. The study of limit distributions for empirical $\mathsf{W}_p^{\gamma_\sigma}$ was initiated in [42] for $p = 1$ in the one-sample case, extended to the two-sample setting in [78], and generalized to arbitrary $p > 1$ via a non-trivial application of the functional delta method in [44]. These works also considered bootstrap consistency and applications to minimum distance estimation and homogeneity testing. To date, a relatively complete limit distribution theory of $\mathsf{W}_p^{\gamma_\sigma}$ in arbitrary dimension is available, as opposed to the rather limited account of classic $\mathsf{W}_p$.

### 5.2. *Structural properties*

We henceforth consider a compactly supported smoothing kernel $\eta_\sigma$ and adopt the shorthand $\mathsf{W}_p^\sigma := \mathsf{W}_p^{\eta_\sigma}$. We start by revisiting structural properties previously established for the Gaussian-smoothed case and demonstrate that they remain valid for $\mathsf{W}_p^\sigma$.

**Proposition 3** (Stability of $\mathsf{W}_p^\sigma$) *For any $1 \leq p < \infty$, $\sigma > 0$, and $\mu, \nu \in \mathscr{P}(\mathbb{R}^d)$, we have*

$$\mathsf{W}_p^\sigma(\mu,\nu) \leq \mathsf{W}_p(\mu,\nu) \leq \mathsf{W}_p^\sigma(\mu,\nu) + 2\sigma(\mathbb{E}_{\eta_1}[\|X\|^p])^{1/p}.$$

*In particular,* $\lim_{\sigma\downarrow 0}\mathsf{W}_p^\sigma(\mu,\nu) = \mathsf{W}_p(\mu,\nu)$.

The first bound is due to contractivity of $\mathsf{W}_p$ w.r.t. convolution. Constructing a coupling between $\rho \in \mathscr{P}(\mathbb{R}^d)$ and $\rho * \eta_\sigma$ with total cost $\sigma(\mathbb{E}_{\eta_1}[\|X\|^p])^{1/p}$ proves the second. Since $\eta_\sigma$ is compactly supported, $\mathrm{spt}(\eta_1)$ is contained in a ball of radius $r > 0$, whereby $(\mathbb{E}_{\eta_1}[\|X\|^p])^{1/p} \le r$ which contrasts the dimension dependent gap for Gaussian kernels; cf. [40, 68].

As the smooth distance converges to the standard distance as $\sigma \to 0$, it is natural to expect that optimal couplings converge as well. This is stated in the next proposition.

**Proposition 4** (Stability of transport plans) *For $1 \le p < \infty$, $\mu, \nu \in \mathscr{P}_p(\mathbb{R}^d)$, and $\sigma_k \downarrow 0$. Let $\pi_k$ be an optimal coupling for $\mathsf{W}_p^{\sigma_k}(\mu,\nu)$ for each $k \in \mathbb{N}$. Then, there exists an optimal coupling $\pi$ for $\mathsf{W}_p(\mu,\nu)$ for which $\pi_k \xrightarrow{w} \pi$ along a subsequence.*

The proof of this result follows that of Theorem 4 in [40] and [44] with only minor changes and is hence omitted. Note that when the limiting $\pi$ is unique (e.g., when $p > 1$ and $\mu$ has a density), then extraction of a subsequence is not needed.

We next show that $\mathsf{W}_p^\sigma$ is indeed a metric on $\mathscr{P}_p(\mathbb{R}^d)$ that induces the Wasserstein topology.

**Proposition 5** (Metric and topological structure) *For $1 \le p < \infty$ and $\sigma > 0$, $\mathsf{W}_p^\sigma$ is a metric on $\mathscr{P}_p(\mathbb{R}^d)$ inducing the same topology as $\mathsf{W}_p$.*

The proof of Proposition 5 follows by observing that the characteristic function of $\eta_\sigma$ vanishes on at most a null set.

### 5.3. *Statistical analysis*

This section studies empirical convergence rates and limit distributions for the smooth Wasserstein distances with compactly supported kernels and population distributions. Let $\mathscr{X} \subset \mathbb{R}^d$ be compact, set $\mathscr{X}_\sigma := \mathscr{X} + \overline{B(0,\sigma)}$, and assume for simplicity that the density of $\eta_\sigma$ is positive on $B(0,\sigma)$, and identically zero on $\mathbb{R}^d \setminus B(0,\sigma)$. For any $\mu \in \mathscr{P}(\mathscr{X})$, the set $\mathscr{X}_\sigma$ contains the support of the convolved measure $\mu * \eta_\sigma$.

#### 5.3.1. Limit distributions for $p > 1$ under the alternative

Building on the unified framework from Proposition 1, the next theorem establishes asymptotic normality of empirical $\mathsf{W}_p^\sigma$ under the alternative. The null case and the $p = 1$ setting are treated in the sequel.

**Theorem 4** (Limit distributions for $\mathsf{W}_p^\sigma$ under the alternative) *Set $1 < p < \infty$, $\sigma > 0$, $\mathsf{V}_p^\sigma := \left[\mathsf{W}_p^\sigma\right]^p$, and let $\mu, \nu \in \mathscr{P}(\mathscr{X})$ be such that $\mathrm{int}(\mathrm{spt}(\mu * \eta_\sigma))$ is connected. Let $\varphi$ be an OT potential from $\mu * \eta_\sigma$ to $\nu * \eta_\sigma$ for $\mathsf{W}_p$, which is unique on $\mathrm{int}(\mathrm{spt}(\mu * \eta_\sigma))$ up to additive constants. The following hold.*

*(i) We have*

$$\sqrt{n}\left(\mathsf{V}_p^\sigma(\hat{\mu}_n,\nu) - \mathsf{V}_p^\sigma(\mu,\nu)\right) \xrightarrow{d} N\left(0, v_p^2\right)$$

*where $v_p^2 := \mathrm{Var}_\mu(\varphi * \chi_\sigma)$. The asymptotic variance $v_p^2$ coincides with the semiparametric effiency bound for estimating $\mathsf{V}_p^\sigma(\cdot,\nu)$ at $\mu$. Also, provided that $v_p^2 > 0$, we have*

$$\sup_{t\in\mathbb{R}}\left|\mathbb{P}^B\left(\sqrt{n}\big(\mathsf{V}_p^\sigma(\hat{\mu}_n^B,\nu) - \mathsf{V}_p^\sigma(\hat{\mu}_n,\nu)\big) \le t\right) - \mathbb{P}\big(N(0,v_p^2) \le t\big)\right| \xrightarrow{\mathbb{P}} 0.$$

*(ii) If in addition $\nu * \eta_\sigma$ has connected support, then*

$$\sqrt{n}\left(\mathsf{V}_p^\sigma(\hat{\mu}_n, \hat{\nu}_n) - \mathsf{V}_p^\sigma(\mu, \nu)\right) \xrightarrow{d} N\left(0, v_p^2 + w_p^2\right),$$

*where $v_p^2$ is as in (i) and $w_p^2 := \mathrm{Var}_\nu(\varphi^c * \chi_\sigma)$. The asymptotic variance $v_p^2 + w_p^2$ coincides with the semiparametric efficiency bound for estimating $\mathsf{V}_p^\sigma$ at $(\mu, \nu)$. Also, provided that $v_p^2 + w_p^2 > 0$, we have*

$$\sup_{t \in \mathbb{R}} \left| \mathbb{P}^B\left(\sqrt{n}\big(\mathsf{V}_p^\sigma(\hat{\mu}_n^B, \hat{\nu}_n^B) - \mathsf{V}_p^\sigma(\hat{\mu}_n, \hat{\nu}_n)\big) \le t\right) - \mathbb{P}\big(N(0, v_p^2 + w_p^2) \le t\big) \right| \xrightarrow{\mathbb{P}} 0.$$

The proof of Theorem 4 applies Proposition 1 to the functional $\rho * \eta_\sigma \mapsto \mathsf{W}_p^p(\rho * \eta_\sigma, \nu * \eta_\sigma)$ for $\rho \in \mathscr{P}(\mathscr{X})$ with $\mathrm{spt}(\rho) \subset \mathrm{spt}(\mu)$. To this end, we show that this functional is Lipschitz continuous w.r.t. $\|\cdot\|_{\infty,B}$ for the unit ball $B$ in $L^2(\mathscr{X}_\sigma)$, which follows by duality (2.2) and uniform bounds on the OT potentials (cf. Remark 1.13 in [95]). The differentiability result follows by adapting the Gaussian kernel case (cf. Lemma 3.3 of [44]). To prove weak convergence of the smoothed empirical process $\sqrt{n}(\hat{\mu}_n - \mu) * \eta_\sigma$ in $\ell^\infty(B)$, we employ the CLT in $L^2(\mathscr{X}_\sigma)$ and use a linear isometry from $L^2(\mathscr{X}_\sigma)$ into $\ell^\infty(B)$. Linearity of the derivative yields asymptotic efficiency and bootstrap consistency.

**Remark 9** (Connectedness assumption) *By* **??** *ahead, the condition from Theorem 4 that* $\mathrm{int}(\mathrm{spt}(\mu * \eta_\sigma))$ *is connected holds whenever $\mu$ itself has connected support.*

The ideas from the proof of Theorem 4 coupled with Hilbertian structure of $L^2(\mathscr{X}_\sigma)$ yield rates of convergence in expectation for empirical $\mathsf{W}_p^\sigma$.

**Proposition 6** (Parametric rate) *For $1 < p < \infty$, $\sigma > 0$, and $\mu, \nu \in \mathscr{P}(\mathscr{X})$ with $\mu \neq \nu$, we have*

$$\mathbb{E}\left[\left|\mathsf{W}_p^\sigma(\hat{\mu}_n, \nu) - \mathsf{W}_p^\sigma(\mu, \nu)\right|\right] \le 2\|\chi_\sigma\|_\infty \sqrt{\lambda(B(0,\sigma))\lambda(\mathscr{X}_\sigma)}\, \mathrm{diam}(\mathscr{X}_\sigma)^p \left[\mathsf{W}_p^\sigma(\mu, \nu)\right]^{1-p} n^{-1/2}.$$

5.3.2. Limit distributions for $p = 2$ under the null

We derive limit distributions for $\mathsf{W}_2^\sigma$ under the null. Our approach relies on the CLT in Hilbert spaces and is thus limited to $p = 2$. Let $C_0^\infty$ denote the space of infinitely differentiable, compactly supported real functions on $\mathbb{R}^d$.

**Definition 4** (Sobolev spaces and their duals) *The Sobolev seminorm of a differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ w.r.t. a reference measure $\mu \in \mathscr{P}(\mathbb{R}^d)$ is denoted by $\|f\|_{\dot{H}^{1,2}(\mu)} := \|\nabla f\|_{L^2(\mu)}$. The homogeneous Sobolev space is defined as the completion of $C_0^\infty + \mathbb{R}$ with respect to $\|\cdot\|_{\dot{H}^{1,2}(\mu)}$. The dual Sobolev space $\dot{H}^{-1,2}(\mu)$ is the topological dual of $\dot{H}^{1,2}(\mu)$.*

**Definition 5** (2-Poincaré inequality) *A probability measure $\mu \in \mathscr{P}(\mathbb{R}^d)$ is said to satisfy the* 2*-Poincaré inequality if there exists $C < \infty$, such that*

$$\|f - \mu(f)\|_{L^2(\mu)} \le C\|\nabla f\|_{L^2(\mu;\mathbb{R}^d)}, \quad f \in C_0^\infty,$$

*where $L^2(\mu;\mathbb{R}^k)$ is the space of Borel maps $f : \mathbb{R}^d \to \mathbb{R}^k$ with $\|f\|^2_{L^2(\mu;\mathbb{R}^k)} := \int_{\mathbb{R}^d} \|f\|^2 d\mu < \infty$.*

With these definitions in place, we state the limit distribution for $\mathsf{W}_2^\sigma$.

**Theorem 5** (Limit distributions for $\mathsf{W}_2^\sigma$ under the null) *Let $\mu \in \mathscr{P}(\mathscr{X})$ be such that $\mu * \eta_\sigma$ satisfies the 2-Poincaré inequality and set $\sigma > 0$. The following hold.*

*(i)* *We have*

$$\sqrt{n}\mathsf{W}_2^\sigma(\hat{\mu}_n, \mu) \xrightarrow{d} \|\mathbb{G}_\mu\|_{\dot{H}^{-1,2}(\mu*\eta_\sigma)},$$

*where $\left(\mathbb{G}_\mu(f)\right)_{f\in\dot{H}^{1,2}(\mu*\eta_\sigma)}$ is a centered Gaussian process with paths in $\dot{H}^{-1,2}(\mu * \eta_\sigma)$ a.s. and covariance function $\operatorname{Cov}\left(\mathbb{G}_\mu(f), \mathbb{G}_\mu(g)\right) = \operatorname{Cov}_\mu(f * \chi_\sigma, g * \chi_\sigma)$.*

*(ii)* *Additionally, if $\mu = \nu$, then*

$$\sqrt{n}\mathsf{W}_2^\sigma(\hat{\mu}_n, \hat{\nu}_n) \xrightarrow{d} \|\mathbb{G}_\mu - \mathbb{G}'_\mu\|_{\dot{H}^{-1,2}(\mu*\eta_\sigma)},$$

*where $\mathbb{G}'_\mu$ is an independent copy of $\mathbb{G}_\mu$.*

The proof of Theorem 5 follows a similar approach to the Gaussian kernel case. In contrast to the proof of Proposition 3.1 in [44], to show weak convergence of the smoothed empirical process in $\dot{H}^{-1,2}(\mu * \eta_\sigma)$, we apply the CLT in Hilbert spaces. To this end, we first verify that the smoothed empirical process has paths in $\dot{H}^{-1,2}(\mu * \eta_\sigma)$. This step requires control of the inverse of the density of $\mu * \eta_\sigma$, which decays to zero near the boundary of its support; see the proof of **??**. The extension to general $1 < p < \infty$ requires a much finer analysis than provided in the proof of **??**, and thus we focus on the $p = 2$ case.

**Remark 10** (Poincaré inequality) *In Theorem 5, a sufficient condition for $\mu * \eta_\sigma$ to satisfy the 2-Poincaré inequality is that both $\mu$ and $\eta_\sigma$ satisfy 2-Poincaré inequalities (see Proposition 1.1 in [97]). The kernel can always be chosen to satisfy 2-Poincaré by simply constructing it from the standard mollifier from Example 2. In that case, $\eta_\sigma$ is a log-concave measure (cf. [56, 81]) and hence satisfies the 2-Poincaré inequality [12, 63].*

Analogously to Proposition 6, parametric rates for empirical $\mathsf{W}_2^\sigma$ under the null follow from Hilbertian structure of $\dot{H}^{-1,2}(\mu * \eta_\sigma)$ and the ideas from the proof of Theorem 5.

**Proposition 7** (Parametric rate) *For $\sigma > 0$ and $\mu \in \mathscr{P}(\mathscr{X})$ for which $\mu * \eta_\sigma$ satisfies the 2-Poincaré inequality with constant $C_{\mu,\sigma}$, we have*

$$\mathbb{E}\left[\mathsf{W}_2^\sigma(\hat{\mu}_n, \mu)\right] \le 2C_{\mu,\sigma}\sqrt{(1 \vee \|\chi_\sigma^2\|_\infty)\lambda(\mathscr{X}_\sigma)}n^{-1/2}.$$

### 5.3.3. Limit distributions for $p = 1$

We now treat the limit distributions for $\mathsf{W}_1^\sigma$ under both the null and the alternative. The Kantorovich-Rubinstein duality for $\mathsf{W}_1$ enables us to do so in the absence of the additional assumptions required when $p > 1$. In what follows, let $\mathsf{Lip}_{1,0}$ denote the set of 1-Lipschitz functions $f$ on $\mathbb{R}^d$ with $f(0) = 0$ and $\mathscr{F}_\sigma = \{f * \chi_\sigma : f \in \mathsf{Lip}_{1,0}\}$. Observe that $\mathsf{W}_1^\sigma(\mu, \nu) = \sup_{f\in\mathscr{F}_\sigma}(\mu - \nu)(f)$ by the Kantorovich-Rubinstein duality.

**Theorem 6** (Limit distributions for $\mathsf{W}_1^\sigma$) *Let $\sigma > 0$ and $\mu, \nu \in \mathscr{P}(\mathscr{X})$. There exist independent, tight $\mu$- and $\nu$-Brownian bridge process $G_\mu$ and $G'_\nu$ in $\ell^\infty(\mathscr{F}_\sigma)$, respectively, such that:*

*(i)* *We have*

$$\sqrt{n}\big(\mathsf{W}_1^\sigma(\hat{\mu}_n,\nu)-\mathsf{W}_1^\sigma(\mu,\nu)\big)\xrightarrow{d}\sup_{f\in M_\sigma} G_\mu(f),$$

*where* $M_\sigma=\big\{f\in\overline{\mathscr{F}}_\sigma^{\mu}:\mu\big(f-\nu(f)\big)=\mathsf{W}_1^\sigma(\mu,\nu)\big\}$ *and* $\overline{\mathscr{F}}_\sigma^{\mu}$ *is the completion of* $\mathscr{F}_\sigma$ *for the pseudometric* $(f,g)\mapsto\sqrt{\mathrm{Var}_\mu(f-g)}$.

*(ii)* *We have*

$$\sqrt{n}\big(\mathsf{W}_1^\sigma(\hat{\mu}_n,\hat{\nu}_n)-\mathsf{W}_1^\sigma(\mu,\nu)\big)\xrightarrow{d}\sup_{f\in M'_\sigma}[G_\mu(f)-G'_\nu(f)],$$

*where* $M'_\sigma=\big\{f\in\overline{\mathscr{F}}_\sigma^{\mu,\nu}:(\mu-\nu)(f)=\mathsf{W}_1^\sigma(\mu,\nu)\big\}$ *and* $\overline{\mathscr{F}}_\sigma^{\mu,\nu}$ *is the completion of* $\mathscr{F}_\sigma$ *for the pseudometric* $(f,g)\mapsto\sqrt{\mathrm{Var}_\mu(f-g)}+\sqrt{\mathrm{Var}_\nu(f-g)}$.

Theorem 6 follows by showing that the function class $\mathscr{F}_\sigma$ is Donsker combined with the extended functional delta method for the supremum functional. The proof of Theorem 6 also establishes parametric rates for empirical $\mathsf{W}_1^\sigma$.

**Corollary 4** (Parametric rate) *For* $\sigma>0$, $\mu\in\mathscr{P}(\mathscr{X})$, *we have* $\mathbb{E}\big[\mathsf{W}_1^\sigma(\hat{\mu}_n,\mu)\big]=O(n^{-1/2})$.

We conclude this section by referring the reader to Appendix **??** for an account of computational aspects for smooth Wasserstein distances. There, we outline the algorithm from [90], show how to lift it to compute $\mathsf{W}_2^\sigma$, and discuss limitations of that method.

## 6. Entropic Optimal Transport

EOT is an efficiently-computable convexification of the OT problem. The general machinery of Proposition 1 enables deriving limit theorems for empirical EOT, generalizing previously available statements to allow for dependent data. Our theory also provides new results on semiparametric efficiency of empirical EOT and consistency of the bootstrap estimate.

### 6.1. *Background*

EOT regularizes OT by the Kullback-Leibler (KL) divergence as

$$\mathsf{S}_c^\varepsilon(\mu,\nu):=\inf_{\pi\in\Pi(\mu,\nu)}\int_{\mathbb{R}^d\times\mathbb{R}^d}c(x,y)d\pi(x,y)+\varepsilon\,\mathsf{D}_{\mathsf{KL}}(\pi\|\mu\otimes\nu),\tag{6.1}$$

where $\varepsilon>0$ and $\mathsf{D}_{\mathsf{KL}}(\mu\|\nu):=\int\log(d\mu/d\nu)d\mu$ if $\mu\ll\nu$ and $+\infty$ otherwise [52, 82]. We consider the quadratic cost $c(x,y)=\|x-y\|^2/2$, assume that $\varepsilon=1$, and use the shorthand $\mathsf{S}(\mu,\nu)=\mathsf{S}^1_{\|\cdot\|^2/2}(\mu,\nu)$. The assumption that $\varepsilon=1$ comes without loss of generality by a rescaling argument, since $\mathsf{S}^\varepsilon_{\|\cdot\|^2/2}(\mu,\nu)=\varepsilon\mathsf{S}(\mu_\varepsilon,\nu_\varepsilon)$, where $\mu_\varepsilon=f_{\varepsilon\sharp}\mu$ for $f_\varepsilon(x)=\varepsilon^{-1/2}x$.

To apply Proposition 1 to empirical EOT, we rely on the duality theory for EOT, whereby

$$\mathsf{S}(\mu,\nu)=\sup_{(\varphi,\psi)\in L^1(\mu)\times L^1(\nu)}\int_{\mathbb{R}^d}\varphi d\mu+\int_{\mathbb{R}^d}\psi d\nu-\int_{\mathbb{R}^d\times\mathbb{R}^d}e^{\varphi\oplus\psi-c}d\mu\otimes\nu+1,$$

with $(\varphi\oplus\psi)(x,y)=\varphi(x)+\psi(y)$. Assuming $\mu,\nu\in\mathscr{P}_2(\mathbb{R}^d)$, the supremum is attained by a pair $(\varphi,\psi)\in L^1(\mu)\times L^1(\nu)$ satisfying the so-called Schrödinger system

$$\begin{aligned}\int_{\mathbb{R}^d}e^{\varphi(x)+\psi(y')-c(x,y')}d\nu(y')&=1\quad\mu\text{-a.e. }x\in\mathbb{R}^d,\\ \int_{\mathbb{R}^d}e^{\varphi(x')+\psi(y)-c(x',y)}d\mu(x')&=1\quad\nu\text{-a.e. }y\in\mathbb{R}^d.\end{aligned}\tag{6.2}$$

We refer to such $(\varphi,\psi)$ as *optimal EOT potentials* (from $\mu$ to $\nu$ for $\varphi$ and vice versa for $\psi$). Optimal EOT potentials are unique $(\mu\otimes\nu)$-almost everywhere up to additive constants. Conversely, any $(\varphi,\psi)\in L^1(\mu)\times L^1(\nu)$ that admit (6.2) are optimal EOT potentials. See Section 1 in [71] and the references therein for details of the duality results for EOT.

### 6.1.1. Literature review

The entropic penalty transforms the OT linear optimization problem into a strongly convex one, allowing efficient computation via the Sinkhorn algorithm [1, 22]. While EOT forfeits the metric and topological structure of $\mathsf{W}_p$,[5] it attains fast empirical convergence in certain cases. Specifically, empirical EOT converges as $n^{-1/2}$ for smooth costs and compactly supported distributions [37], or for the squared cost with sub-Gaussian distributions [60].

Limit distributions for EOT (and the Sinkhorn divergence) for $c(x,y)=\|x-y\|^p$ in the discrete support case were provided in [9, 50]. Their approach is to parameterize each marginal by a finite-dimensional simplex vector and find the derivative of the EOT cost w.r.t. the simplex vector to apply the standard delta method; arguably, this approach does not directly extend to general distributions. A CLT for EOT between sub-Gaussian distribution was first derived in [60], showing asymptotic normality of $\sqrt{n}\big(\mathsf{S}(\hat{\mu}_n,\nu)-\mathbb{E}\big[\mathsf{S}(\hat{\mu}_n,\nu)\big]\big)$ and its two-sample analog using the Efron-Stein inequality similar to [24]. The main limitation of this result is that the centering term is the expected empirical EOT, which is undesirable because it does not enable performing inference for $\mathsf{S}(\mu,\nu)$. This limitation was addressed in the recent preprint [29], see the discussion in Remark 11. We provide here an alternative derivation of the CLT that relies on establishing the Hadamard derivatives of the EOT cost w.r.t. the marginals following the unified framework from Proposition 1, which automatically leads to asymptotic efficiency of empirical EOT and consistency of the bootstrap estimate, as well as the extension for dependent data. The Hadamard differentiability result (implicit in the proof) may be of independent interest as it pertains to stability analysis of EOT, which has attracted growing interest in the mathematics literature [35, 39, 61, 62, 72].

## 6.2. *Statistical analysis*

We next state the CLT, asymptotic efficiency, and bootstrap consistency for empirical EOT.

[5] Indeed, e.g., $\mathsf{S}_c^{\varepsilon}(\mu,\mu)\neq 0$; while this can be fixed via centering EOT to obtain the so-called Sinkhorn divergence, it is still not a metric since it lacks the triangle inequality [9].

**Theorem 7** (CLT, efficiency, and bootstrap consistency for EOT) *Suppose that $\mu,\nu \in \mathscr{P}(\mathbb{R}^d)$ are sub-Gaussian. Let $(\varphi,\psi)$ be optimal EOT potentials for $(\mu,\nu)$. Then, the following hold.*

*(i) We have $\sqrt{n}\big(\mathsf{S}(\hat{\mu}_n,\nu)-\mathsf{S}(\mu,\nu)\big) \xrightarrow{d} N\big(0,\mathfrak{v}_1^2\big)$ with $\mathfrak{v}_1^2 = \mathrm{Var}_\mu(\varphi)$. The asymptotic variance $\mathfrak{v}_1^2$ coincides with the semiparametric efficiency bound for estimating $\mathsf{S}(\cdot,\nu)$ at $\mu$. Finally, provided that $\mathfrak{v}_1^2 > 0$, we have*

$$\sup_{t\in\mathbb{R}}\left|\mathbb{P}^B\Big(\sqrt{n}\big(\mathsf{S}(\hat{\mu}_n^B,\nu)-\mathsf{S}(\hat{\mu}_n,\nu)\big)\le t\Big)-\mathbb{P}\big(N(0,\mathfrak{v}_1^2)\big)\le t\big)\right| \xrightarrow{\mathbb{P}} 0.$$

*(ii) We have $\sqrt{n}\big(\mathsf{S}(\hat{\mu}_n,\hat{\nu}_n)-\mathsf{S}(\mu,\nu)\big) \xrightarrow{d} N\big(0,\mathfrak{v}_1^2+\mathfrak{v}_2^2\big)$ where $\mathfrak{v}_1^2$ is as in (i) and $\mathfrak{v}_2^2 = \mathrm{Var}_\nu(\psi)$. The asymptotic variance $\mathfrak{v}_1^2+\mathfrak{v}_2^2$ coincides with the semiparametric efficiency bound for estimating $\mathsf{S}(\cdot,\cdot)$ at $(\mu,\nu)$. Finally, provided that $\mathfrak{v}_1^2+\mathfrak{v}_2^2 > 0$, we have*

$$\sup_{t\in\mathbb{R}}\left|\mathbb{P}^B\Big(\sqrt{n}\big(\mathsf{S}(\hat{\mu}_n^B,\hat{\nu}_n^B)-\mathsf{S}(\hat{\mu}_n,\hat{\nu}_n)\big)\le t\Big)-\mathbb{P}\big(N(0,\mathfrak{v}_1^2+\mathfrak{v}_2^2)\le t\big)\right| \xrightarrow{\mathbb{P}} 0.$$

**Remark 11** (Comparison with [29]) *As mentioned in Section 6.1.1, a CLT for one- and two-sample EOT was derived in Theorem 3.6 of [29], whose proof first expands the empirical EOT cost around its expectation and then shows that the bias is negligible. We rederive this result via a markedly different proof technique, relying on the unified framework from Proposition 1, which automatically also implies bootstrap consistency and asymptotic efficiency via Corollary 1 and Proposition 2, both of which were not addressed in [29]. In addition, as Proposition 1 does not assume i.i.d. data, the above result readily extends to dependent data, which falls outside the framework of [29]. For instance, suppose that $\{X_t\}_{t\in\mathbb{Z}}$ is a stationary $\beta$-mixing process with compactly supported marginal distribution $\mu$. Then, by Theorem 1 in [33],*

$$\sqrt{n}(\hat{\mu}_n-\mu) \xrightarrow{d} G \quad \text{in } \ell^\infty(\mathscr{F}_\sigma),$$

*where $\mathscr{F}_\sigma$ is the function class given in (**??**) ahead with $s = \max\{\lfloor d/2\rfloor+1,2\}$ and sufficiently large $\sigma > 0$, while $G$ is a tight centered Gaussian process in $\ell^\infty(\mathscr{F}_\sigma)$ with covariance function $\mathrm{Cov}\big(G(f),G(g)\big) = \sum_{t\in\mathbb{Z}}\mathrm{Cov}\big(f(X_0),g(X_t)\big)$. Conclude from the proof of Theorem 7 that*

$$\sqrt{n}\big(\mathsf{S}(\hat{\mu}_n,\nu)-\mathsf{S}(\mu,\nu)\big) \xrightarrow{d} G(\varphi) \sim N\Big(0,\textstyle\sum_{t\in\mathbb{Z}}\mathrm{Cov}\big(\varphi(X_0),\varphi(X_t)\big)\Big).$$

*Likewise, a CLT result holds for other forms of dependent data, such as exchangeable arrays [23]. Furthermore, for the EOT case, the corresponding Hadamard derivative is linear, so suitable dependent bootstrap methods, such as the block bootstrap for mixing data [15] and the (extended) pigeonhole bootstrap for exchangeable arrays [23], are consistent for the empirical EOT cost, provided that the bootstrap processes satisfy a uniform CLT for $\mathscr{F}_\sigma$.*

## 7. Concluding Remarks

This work developed a unified framework for proving limit distribution results for empirical regularized OT distances, semiparametric efficiency of the plug-in empirical estimator, and consistency of the bootstrap. As applications, we focused on three prominent OT regularization methods—smoothing, slicing, and entropic penalty—and provided a comprehensive statistical treatment thereof. We closed existing gaps in the literature (e.g., a limit distribution theory for sliced $\mathsf{W}_p$) and provided several new results concerning empirical convergence rates, asymptotic efficiency, and bootstrap consistency. In particular, for the smooth Wasserstein distance, we explored compactly supported smoothing kernels, which were shown to inherit the structural and statistical properties of the well-studied Gaussian-smoothed framework. The analysis of compactly supported kernels is motivated by computational considerations, as we demonstrated how to lift the efficient algorithm from [90] for computing $\mathsf{W}_2^2$ between smooth densities to the considered smooth OT distance.

Our framework is flexible and can treat a broad class of functionals, potentially well beyond the three examples considered herein. For instance, straightforward adaptations of our arguments for sliced $\mathsf{W}_p$ would yield limit distributions, efficiency, and bootstrap consistency of the projection-robust Wasserstein distance from [55], when the projected subspace is of dimension $k \leq 3$ (indeed, the class of projected OT potentials is still Donsker in that case). Going forward, we also plan to explore applicability of the unified framework to empirical OT maps or certain functionals thereof (e.g., inner product with a smooth test function).

### Acknowledgments

The authors would like to thank Sloan Nietert for fruitful discussions on smooth Wasserstein distances with compactly supported kernels, and Tudor Manole for bringing their attention to Theorem 4 in [59]. They would also like to thank the anonymous associate editor and reviewers, whose comments helped up improve the paper.

### Funding

Z. Goldfeld is supported by the NSF CRII grant CCF-1947801, in part by the 2020 IBM Academic Award, and in part by the NSF CAREER Award CCF-2046018. K. Kato is partially supported by the NSF grants DMS-1952306 and DMS-2014636. G. Rioux is partially supported by the NSERC postgraduate fellowship PGSD-567921-2022.

### Data availability

No new data were generated or analysed in support of this review.

## References

1. J. Altschuler, J. Niles-Weed, and P. Rigollet. Near-linear time approximation algorithms for optimal transport via Sinkhorn iteration. In *Proceedings of the International Conference on Neural Information Processing Systems*, volume 30, 2017.
2. D. W. Andrews and D. Pollard. An introduction to functional central limit theorems for dependent stochastic processes. *Int. Stat. Rev.*, pages 119–132, 1994.
3. M. A. Arcones and B. Yu. Central limit theorems for empirical and u-processes of stationary mixing sequences. *J. Theoret. Probab.*, 7(1):47–71, 1994.
4. M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein generative adversarial networks. In *Proceedings of the International Conference on Machine Learning*, pages 214–223, 2017.
5. J. Bae and S. Levental. Uniform CLT for Markov chains and its invariance principle: a martingale approach. *J. Theoret. Probab.*, 8(3):549–570, 1995.

6. E. Bayraktar and G. Guo. Strong equivalence between metrics of Wasserstein type. *Electron. Commun. Probab.*, 26:1–13, 2021.
7. E. Bernton, P. E. Jacob, M. Gerber, and C. P. Robert. Approximate bayesian computation with the Wasserstein distance. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 81(2):235–269, 2019.
8. E. Bernton, P. E. Jacob, M. Gerber, and C. P. Robert. On parameter estimation with the Wasserstein distance. *Inf. Inference*, 8(4):657–676, 2019.
9. J. Bigot, E. Cazelles, and N. Papadakis. Central limit theorems for entropy-regularized optimal transport on finite spaces and statistical applications. *Electron. J. Stat.*, 13(2):5120–5150, 2019.
10. J. Blanchet and K. Murthy. Quantifying distributional model risk via optimal transport. *Math. Oper. Res.*, 44 (2):565–600, 2019.
11. A. Block, Z. Jia, Y. Polyanskiy, and A. Rakhlin. Rate of convergence of the smoothed empirical Wasserstein distance. *arXiv preprint arXiv:2205.02128*, 2022.
12. S. G. Bobkov. Isoperimetric and analytic inequalities for log-concave probability measures. *Ann. Probab.*, 27(4):1903–1921, 1999.
13. S. G. Bobkov and M. Ledoux. *One-dimensional empirical measures, order statistics, and Kantorovich transport distances*. American Mathematical Society, 2019.
14. N. Bonnotte. *Unidimensional and evolution methods for optimal transportation*. PhD thesis, Paris 11, 2013.
15. P. Bühlmann. The blockwise bootstrap for general empirical processes of stationary sequences. *Stochastic Process. Appl.*, 58(2):247–265, 1995.
16. G. Carlier, V. Chernozhukov, and A. Galichon. Vector quantile regression: an optimal transport approach. *Ann. Statist.*, 44(3):1165–1192, 2016.
17. H.-B. Chen and J. Niles-Weed. Asymptotics of smoothed Wasserstein distances. *Potential Anal.*, pages 1–25, 2021.
18. Y. Chen, Q. Gao, and X. Wang. Inferential Wasserstein generative adversarial networks. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 84(1):83–113, 11 2021.
19. V. Chernozhukov, S. Lee, and A. M. Rosen. Intersection bounds: estimation and inference. *Econometrica*, 81(2):667–737, 2013.
20. V. Chernozhukov, A. Galichon, M. Hallin, and M. Henry. Monge–Kantorovich depth, quantiles, ranks and signs. *Ann. Statist.*, 45(1):223–256, 2017.
21. N. Courty, R. Flamary, D. Tuia, and A. Rakotomamonjy. Optimal transport for domain adaptation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(9):1853–1865, 2016.
22. M. Cuturi. Sinkhorn distances: lightspeed computation of optimal transport. In *Proceedings of the International Conference on Neural Information Processing Systems*, pages 2292–2300, 2013.
23. L. Davezies, X. D’Haultfœuille, and Y. Guyonvarch. Empirical process results for exchangeable arrays. *Ann. Statist.*, 49(2):845–862, 2021.
24. E. del Barrio and J.-M. Loubes. Central limit theorems for empirical transportation cost in general dimension. *Ann. Probab.*, 47(2):926–951, 2019.
25. E. del Barrio, E. Giné, and C. Matrán. Central limit theorems for the Wasserstein distance between the empirical and the true distributions. *Ann. Probab.*, 27(2):1009–1071, 1999.
26. 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.
27. E. del Barrio, P. Gordaliza, and J.-M. Loubes. A central limit theorem for $L^p$ transportation cost on the real line with application to fairness assessment in machine learning. *Inf. Inference*, 8(4):817–849, 2019.
28. E. del Barrio, A. González-Sanz, and J.-M. Loubes. Central limit theorems for general transportation costs. *arXiv preprint: arXiv:2102.06379*, 2021.
29. E. del Barrio, A. G. Sanz, J.-M. Loubes, and J. Niles-Weed. An improved central limit theorem and fast convergence rates for entropic transportation costs. *SIAM J. Math. Data Sci.*, 5(3):639–669, 2023.
30. E. del Barrio, A. González Sanz, and J.-M. Loubes. Central limit theorems for semi-discrete wasserstein distances. *Bernoulli*, 30(1):554–580, 2024.

31. I. Deshpande, Z. Zhang, and A. G. Schwing. Generative modeling using the sliced Wasserstein distance. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3483–3491, 2018.
32. 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10648–10656, 2019.
33. P. Doukhan, P. Massart, and E. Rio. Invariance principles for absolutely regular empirical processes. *Ann. Inst. Henri Poincaré Probab. Stat.*, 31(2):393–427, 1995.
34. L. Dümbgen. On nondifferentiable functions and the bootstrap. *Probab. Theory Related Fields*, 95:125–140, 1993.
35. S. Eckstein and M. Nutz. Quantitative stability of regularized optimal transport and convergence of Sinkhorn's algorithm. *SIAM J. Math. Anal.*, 54(6):5922–5948, 2022.
36. Z. Fang and A. Santos. Inference on directionally differentiable functions. *Rev. Econ. Stud.*, 86:377–412, 2019.
37. A. Genevay, L. Chizat, F. Bach, M. Cuturi, and G. Peyré. Sample complexity of Sinkhorn divergences. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, pages 1574–1583, 2019.
38. P. Ghosal and B. Sen. Multivariate ranks and quantiles using optimal transport: consistency, rates, and nonparametric testing. *Ann. Statist., to appear*, 2021.
39. P. Ghosal, M. Nutz, and E. Bernton. Stability of entropic optimal transport and Schrödinger bridges. *J. Funct. Anal.*, 283(9):109622, 2022.
40. Z. Goldfeld and K. Greenewald. Gaussian-smoothed optimal transport: Metric structure and statistical efficiency. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, pages 3327–3337, 2020.
41. Z. Goldfeld and K. Kato. Limit distribution for smooth total variation and $\chi^2$-divergence in high dimensions. In *Proceedings of the IEEE International Symposium on Information Theory*, 2020.
42. Z. Goldfeld, K. Greenewald, and K. Kato. Asymptotic guarantees for generative modeling based on the smooth Wasserstein distance. In *Proceedings of the International Conference on Neural Information Processing Systems*, pages 2527–2539, 2020.
43. Z. Goldfeld, K. Greenewald, J. Niles-Weed, and Y. Polyanskiy. Convergence of smoothed empirical measures with applications to entropy estimation. *IEEE Trans. Inform. Theory*, 66(7):4368–4391, 2020.
44. Z. Goldfeld, K. Kato, S. Nietert, and G. Rioux. Limit distribution theory for smooth $p$-Wasserstein distances. *Ann. Appl. Probab., to appear*, 2022.
45. I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville. Improved training of Wasserstein GANs. In *Proceedings of the International Conference on Neural Information Processing Systems*, pages 5769–5779, 2017.
46. M. Hallin, E. del Barrio, J. Cuesta-Albertos, and C. Matrán. Distribution and quantile functions, ranks and signs in dimension d: A measure transportation approach. *Ann. Statist.*, 49(2):1139 – 1165, 2021.
47. F. Han, Z. Miao, and Y. Shen. Nonparametric mixture MLEs under Gaussian-smoothed optimal transport distance. *IEEE Trans. Inform. Theory*, 2023.
48. S. Hundrieser, M. Klatt, T. Staudt, and A. Munk. A unifying approach to distributional limits for empirical optimal transport. *arXiv preprint: arXiv:2202.12790*, 2022.
49. L. V. Kantorovich. On the translocation of masses. In *Dokl. Akad. Nauk*, volume 37, pages 199–201, 1942.
50. M. Klatt, C. Tameling, and A. Munk. Empirical regularized optimal transport: Statistical theory and applications. *SIAM J. Math. Data Sci.*, 2(2):419–443, 2020.
51. S. Kolouri, P. E. Pope, C. E. Martin, and G. K. Rohde. Sliced Wasserstein auto-encoders. In *Proceedings of the International Conference on Learning Representations*, 2019.
52. C. Léonard. A survey of the Schrödinger problem and some of its connections with optimal transport. *Discrete Cont. Dyn. Syst.*, 34(4):1533, 2014.
53. S. Levental. Uniform limit theorems for Harris recurrent markov chains. *Probab. Theory Related Fields*, 80 (1):101–118, 1988.

54. P. Li, Q. Wang, and L. Zhang. A novel earth mover's distance methodology for image matching with gaussian mixture models. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1689–1696, 2013.
55. T. Lin, Z. Zheng, E. Y. Chen, M. Cuturi, and M. I. Jordan. On projection robust optimal transport: Sample complexity and model misspecification. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, pages 262–270, 2021.
56. L. Lovász and S. Vempala. The geometry of logconcave functions and sampling algorithms. *Random Structures Algorithms*, 30(3):307–358, 2007.
57. T. Manole and J. Niles-Weed. Sharp convergence rates for empirical optimal transport with smooth costs. *Ann. Appl. Probab. (to appear)*, 2021.
58. T. Manole, S. Balakrishnan, J. Niles-Weed, and L. Wasserman. Plugin estimation of smooth optimal transport maps. *arXiv preprint: arXiv 2107.12364*, 2021.
59. T. Manole, S. Balakrishnan, and L. Wasserman. Minimax confidence intervals for the sliced wasserstein distance. *Electron. J. Stat.*, 16(1):2252–2345, 2022.
60. G. Mena and J. Niles-Weed. Statistical bounds for entropic optimal transport: sample complexity and the central limit theorem. *Procceedings of the International Conference on Neural Information Processing Systems*, 32, 2019.
61. T. Mikami. Regularity of Schrodinger's functional equation and mean field pdes for $h$-path processes. *Osaka J. Math.*, 56(4):831–842, 2019.
62. T. Mikami. Regularity of Schrödinger's functional equation in the weak topology and moment measures. *J. Math. Soc. Japan*, 73(1):99–123, 2021.
63. E. Milman. On the role of convexity in isoperimetry, spectral gap and concentration. *Invent. Math.*, 177(1): 1–43, 2009.
64. G. Monge. Mémoire sur la théorie des déblais et des remblais. *Histoire de l'Académie Royale des Sciences de Paris*, 1781.
65. B. Muzellec, A. Vacher, F. Bach, F.-X. Vialard, and A. Rudi. Near-optimal estimation of smooth transport maps with kernel sums-of-squares. *arXiv preprint: arXiv:2112.01907*, 2021.
66. K. Nadjahi, A. Durmus, U. Simsekli, and R. Badeau. Asymptotic guarantees for learning generative models with the sliced-Wasserstein distance. In *Proceedings of the International Conference on Neural Information Processing Systems*, volume 32, 2019.
67. K. Nadjahi, A. Durmus, L. Chizat, S. Kolouri, S. Shahrampour, and U. Simsekli. Statistical and topological properties of sliced probability divergences. In *Proceedings of the International Conference on Neural Information Processing Systems*, volume 33, pages 20802–20812, 2020.
68. S. Nietert, Z. Goldfeld, and K. Kato. Smooth $p$-Wasserstein distance: structure, empirical approximation, and statistical applications. In *Proceedings of the International Conference on Machine Learning*, pages 8172–8183. PMLR, 2021.
69. J. Niles-Weed and P. Rigollet. Estimation of wasserstein distances in the spiked transport model. *Bernoulli*, 28(4):2663–2688, 2022.
70. Y. Nishiyama. Weak convergence of some classes of martingales with jumps. *Ann. Probab.*, 28(2):685–712, 2000.
71. M. Nutz and J. Wiesel. Entropic optimal transport: convergence of potentials. *Probab. Theory Related Fields*, pages 1–24, 2021.
72. M. Nutz and J. Wiesel. Stability of Schrödinger potentials and convergence of Sinkhorn's algorithm. *Ann. Probab.*, 51(2):699–722, 2023.
73. M. Okamoto. Distinctness of the eigenvalues of a quadratic form in a multivariate sample. *Ann. Statist.*, 1 (4):763 – 765, 1973.
74. D. N. Politis and J. P. Romano. Large sample confidence regions based on subsamples under minimal assumptions. *Ann. Statist.*, pages 2031–2050, 1994.
75. 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.

76. W. Römisch. Delta method, infinite dimensional. In *Encyclopedia of Statistical Sciences*. Wiley, 2004.
77. Y. Rubner, C. Tomasi, and L. J. Guibas. The earth mover's distance as a metric for image retrieval. *Int. J. Comput. Vis.*, 40(2):99–121, 2000.
78. R. Sadhu, Z. Goldfeld, and K. Kato. Limit distribution theory for the smooth 1-Wasserstein distance with applications. *arXiv preprint arXiv:2107.13494*, 2021.
79. R. Sandler and M. Lindenbaum. Nonnegative matrix factorization with earth mover's distance metric for image analysis. *IEEE Trans. Pattern Anal. Mach. Intell.*, 33(8):1590–1602, 2011.
80. F. Santambrogio. *Optimal transport for applied mathematicians*. Birkhäuser, 2015.
81. A. Saumard and J. A. Wellner. Log-concavity and strong log-concavity: a review. *Stat. Surv.*, 8:45, 2014.
82. E. Schrödinger. Über die umkehrung der naturgesetze. *Akad. Wiss. Berlin. Phys. Math.*, 144:144–153, 1931.
83. A. Shapiro. On concepts of directional differentiability. *J. Optim. Theory Appl.*, 66:477–487, 1990.
84. A. Shapiro. Asymptotic analysis of stochastic programs. *Ann. Oper. Res.*, 30:169–186, 1991.
85. J. Solomon, F. De Goes, G. Peyré, M. Cuturi, A. Butscher, A. Nguyen, T. Du, and L. Guibas. Convolutional Wasserstein distances: Efficient optimal transportation on geometric domains. *ACM Trans. Graph.*, 34(4): 66, 2015.
86. M. Sommerfeld and A. Munk. Inference for empirical Wasserstein distances on finite spaces. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 80:219–238, 2018.
87. C. Tameling, M. Sommerfeld, and A. Munk. Empirical optimal transport on countable metric spaces: Distributional limits and statistical applications. *Ann. Appl. Probab.*, 29:2744–2781, 2019.
88. I. Tolstikhin, O. Bousquet, S. Gelly, and B. Schölkopf. Wasserstein auto-encoders. In *Proceedings of the International Conference on Learning Representations*, 2018.
89. W. Torous, F. Gunsilius, and P. Rigollet. An optimal transport approach to causal inference. *arXiv preprint arXiv:2108.05858*, 2021.
90. A. Vacher, B. Muzellec, A. Rudi, F. Bach, and F.-X. Vialard. A dimension-free computational upper-bound for smooth optimal transport estimation. In *Proceedings of the Conference on Learning Theory*, pages 4143–4173, 2021.
91. A. W. van der Vaart. Efficiency and Hadamard differentiability. *Scand. J. Stat.*, 18(1):63–75, 1991.
92. A. W. van der Vaart. New Donsker classes. *Ann. Probab.*, 24(4):2128–2124, 1996.
93. A. W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, Cambridge, UK, 1998.
94. A. W. van der Vaart and J. A. Wellner. Weak convergence. In *Weak convergence and empirical processes*, pages 16–28. Springer, 1996.
95. C. Villani. *Topics in optimal transportation*. Number 58. American Mathematical Society, 2003.
96. C. Villani. *Optimal transport: old and new*. Springer, 2008.
97. F.-Y. Wang and J. Wang. Functional inequalities for convolution probability measures. *Ann. Inst. Henri Poincaré Probab. Stat.*, 52(2):898–914, 2016.
98. E. Wong, F. Schmidt, and Z. Kolter. Wasserstein adversarial examples via projected Sinkhorn iterations. In *Proceedings of the International Conference on Machine Learning*, pages 6808–6817, 2019.
99. J. Xi and J. Niles-Weed. Distributional convergence of the sliced Wasserstein process. *arXiv preprint arXiv:2206.00156*, 2022.
100. X. Xu and Z. Huang. Central limit theorem for the sliced 1-Wasserstein distance and the max-sliced 1-Wasserstein distance. *arXiv preprint arXiv:2205.14624*, 2022.
101. Y. Zhang, X. Cheng, and G. Reeves. Convergence of Gaussian-smoothed optimal transport distance with sub-gamma distributions and dependent samples. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, pages 2422–2430, 2021.