Minimax Confidence Intervals for the Sliced Wasserstein Distance

Tudor Manole, Sivaraman Balakrishnan, Larry Wasserman
Department of Statistics and Data Science, Carnegie Mellon University
tmanole@andrew.cmu.edu, {siva,larry}@stat.cmu.edu
February 23, 2021

Abstract

Motivated by the growing popularity of variants of the Wasserstein distance in statistics and machine learning, we study statistical inference for the Sliced Wasserstein distance—an easily computable variant of the Wasserstein distance. Specifically, we construct confidence intervals for the Sliced Wasserstein distance which have finite-sample validity under no assumptions or under mild moment assumptions. These intervals are adaptive in length to the regularity of the underlying distributions. We also bound the minimax risk of estimating the Sliced Wasserstein distance, and as a consequence establish that the lengths of our proposed confidence intervals are minimax optimal over appropriate distribution classes. To motivate the choice of these classes, we also study minimax rates of estimating a distribution under the Sliced Wasserstein distance. These theoretical findings are complemented with a simulation study demonstrating the deficiencies of the classical bootstrap, and the advantages of our proposed methods. We also show strong correspondences between our theoretical predictions and the adaptivity of our confidence interval lengths in simulations. We conclude by demonstrating the use of our confidence intervals in the setting of simulator-based likelihood-free inference. In this setting, contrasting popular approximate Bayesian computation methods, we develop uncertainty quantification methods with rigorous frequentist coverage guarantees.

1 Introduction

The Wasserstein distance is a metric between probability distributions which has received a surge of interest in statistics and machine learning (Panaretos & Zemel 2019a, Kolouri et al. 2017). This distance arises from the optimal transport problem (Villani 2003), and measures the work required to couple one distribution with another. Specifically, given probability distributions $P$ and $Q$ admitting at least $r \geq 1$ moments, with support in $\mathbb{R}^d$, $d \geq 1$, the $r$-th order Wasserstein distance between $P$ and $Q$ is defined by

$$W_r(P,Q) = \left( \inf_{\gamma \in \Pi(P,Q)} \int \|x-y\|^r d\gamma(x,y) \right)^{1/r},$$

where $\Pi(P,Q)$ denotes the set of joint probability distributions with marginals $P$ and $Q$, known as couplings. Any minimizer $\gamma$ is called an optimal coupling between $P$ and $Q$. The norm $\|\cdot\|$ is taken to be Euclidean in this paper, but may more generally be replaced by any metric on $\mathbb{R}^d$.

Keywords: Sliced Wasserstein Distance, Optimal Transport, Confidence Interval, Minimax Lower Bound, Adaptation, Likelihood-Free Inference
The Wasserstein distance has broadly served two uses in the statistics literature (see the review article (Panaretos & Zemel 2019b) and references therein). On the one hand, it has been used as a theoretical tool for asymptotic theory (see for instance Shorack & Wellner (2009), Shao & Tu (2012)), since convergence in r-Wasserstein distance is equivalent to weak convergence of probability measures and their r-th moments (Villani 2008). Wasserstein distances also play a prominent role in the analysis of mixture models (Nguyen 2013, Ho et al. 2019). On the other hand, increasingly many statistical applications employ the Wasserstein distance as a methodological tool in its own right. Unlike many common metrics between probability distributions, the Wasserstein distance does not presume distributions which are absolutely continuous with respect to a common dominating measure, and is sensitive to the underlying geometry of their support, due to the $\ell_2$-norm embedded in its definition. These considerations make it a natural and powerful data analytic tool—see for instance del Barrio et al. (1999, 2005), Courty et al. (2016), Ramdas et al. (2017), Arjovsky et al. (2017), Ho et al. (2017), Bernton et al. (2019a, b), Verdinelli & Wasserman (2019).

Despite the popularity of the Wasserstein distance, its high computational complexity often limits its applicability to large-scale problems. Developing efficient numerical approximations of the distance remains an active research area—see Peyré & Cuturi (2019) for a recent review. A key exception to the high computational cost is the univariate case, in which the Wasserstein distance admits a closed form as the $L^r$ norm between the quantile functions of $P$ and $Q$, which can be easily computed. This fact has led to the study of an alternate metric, known as the Sliced Wasserstein distance (Rabin et al. 2011, Bonneel et al. 2015), obtained by averaging the Wasserstein distance between random one-dimensional projections of the distributions $P$ and $Q$. The Sliced Wasserstein distance is generally a weaker metric than the Wasserstein distance (Bonnotte 2013), but nevertheless preserves many qualitatively similar properties which make it an attractive and easily computable alternative in many applications.

Motivated by the fact that the Wasserstein distance and its sliced analogue are sensitive to outliers and heavy tails, we introduce a trimmed version of the Sliced Wasserstein distance, denoted by $SW_{r,\delta}(P,Q)$ for some trimming constant $\delta \in [0,1/2]$ and defined formally in equation (12). This robustification of the Sliced Wasserstein distance compares distributions up to a $2\delta$ fraction of their probability mass, thereby generalizing the one-dimensional trimmed Wasserstein distance introduced by Munk & Czado (1998) (see also Álvarez-Esteban et al. (2008, 2012)). One of the aims of our paper is to derive confidence intervals for the trimmed Sliced Wasserstein distance which make either no assumptions or mild moment assumptions on the unknown distributions $P$ and $Q$. Specifically, given a level $\alpha \in (0,1)$ and i.i.d. samples $X_1,\ldots,X_n \sim P$ and $Y_1,\ldots,Y_m \sim Q$, we derive confidence sets $C_{nm} \subseteq \mathbb{R}$ such that

$$\inf_{P\sim Q} \mathbb{P}(SW_{r,\delta}(P,Q) \in C_{nm}) \geq 1 - \alpha,$$

where the infimum is over a suitable family of distributions $P,Q$.

One of the main reasons that the Wasserstein distance has found many applications is the fact that it is a useful notion of distance under weak assumptions. Unlike the Total Variation, Hellinger, Kullback-Leibler and other divergences, the Wasserstein distance between a pair of distributions can be estimated from samples (optimally) under mild assumptions without requiring any smoothing. However, existing results on inference for the Wasserstein distance (Munk & Czado 1998, Freitag et al. 2003, 2007, Freitag & Munk 2005), typically require strong smoothness assumptions and suggest different inferential procedures when $P=Q$ as compared to when $P \neq Q$. In contrast we construct various assumption-light confidence intervals $C_{nm}$ which have finite-sample validity under weak moment assumptions.

The confidence intervals we construct are adaptive to the regularity of the distributions $P$ and $Q$, as
measured by a functional $SJ_{r,\delta}(P)$ introduced formally in Section 3.1 (equation (17)). The magnitude of $SJ_{r,\delta}(P)$ is largely controlled by the tails of $P$ and by whether its one-dimensional projections have connected support. The one-dimensional counterpart of this functional was identified in the work of Bobkov & Ledoux (2019) who showed that when this functional is finite the empirical measure of $n$ samples converges to its underlying distribution in the Wasserstein distance at the fast rate of $O(1/\sqrt{n})$. On the other hand, when this functional is infinite Bobkov & Ledoux (2019) showed that this convergence happens at a slower rate of $O((1/n)^{1/2r})$. Our work shows that the role of the $SJ_{r,\delta}$ functional in inference is more nuanced. We show that when the $SJ_{r,\delta}$ functionals of $P$ and $Q$ are finite then our confidence intervals have length scaling at the fast rate of $O(1/\sqrt{n\wedge m})$, mirroring the rates of convergence in the work of Bobkov & Ledoux (2019). On the other hand, when the $SJ_{r,\delta}$ functionals are infinite, a dichotomy arises: in full generality, when $SW_{r,\delta}(P,Q)$ is allowed to take arbitrary (small) values uncertainty quantification is difficult and our intervals can have lengths scaling as $O((1/n\wedge m)^{1/2r})$ in the worst case. However, we find, somewhat surprisingly, even when the $SJ_{r,\delta}$ functional is infinite, accurate $O(1/\sqrt{n\wedge m})$-inference is possible so long as $SW_{r,\delta}(P,Q)$ is bounded away from 0 (i.e. the distributions are separated in the Sliced Wasserstein distance). To summarize, we find that accurate inference is possible when either $SW_{r,\delta}(P,Q)$ is bounded away from 0, or when the $SJ_{r,\delta}$ functionals are finite. We emphasize that the intervals we construct are adaptive, i.e. they have small lengths under appropriate conditions on the $SJ_{r,\delta}$ functional and $SW_{r,\delta}(P,Q)$, without needing the statistician to specify or have knowledge of these quantities. We also show that our confidence intervals have minimax optimal length over classes of distributions with varying magnitudes of $SJ_{r,\delta}(P)$.

To complement our results on confidence intervals for the Sliced Wasserstein distance we also consider the problem of estimating the Sliced Wasserstein distance between two distributions, given samples from each of them. We provide sharp minimax upper and lower bounds for this problem as well. Indeed, our minimax lower bounds for confidence interval length are derived directly from minimax lower bounds for estimating the Sliced Wasserstein distance by noting that the minimax length of a confidence interval is bounded from below by the corresponding minimax estimation rate.

We illustrate the practical significance of our methodology via an application to likelihood-free inference (Sisson et al. 2018), in which a parametrized stochastic simulator for the data-generating process is available, but its underlying distribution is intractable. Here, our goal is to construct confidence intervals for unknown parameters of the simulator, on the basis of minimizing its Sliced Wasserstein distance from an observed sample. Distributional assumptions such as those made in past work on inference for the one-dimensional Wasserstein distance (Munk & Czado 1998, Freitag et al. 2003, 2007, Freitag & Munk 2005) are typically unverifiable in such applications.

**Our Contributions.** We summarize the contributions of this paper as follows.

- We define the $\delta$-trimmed Sliced Wasserstein distance $SW_{r,\delta}$, and the functional $SJ_{r,\delta}$, generalizing the functional $J_r$ of Bobkov & Ledoux (2019). Hinging upon these results, we show that the finiteness of $SJ_{r,\delta}(P)$ is a sufficient condition for the empirical measure to estimate $P$ at the parametric rate under the Sliced Wasserstein distance, and we prove corresponding minimax lower bounds. We also derive the minimax rates of estimating the Sliced Wasserstein distance between two distributions. These minimax rates are sensitive to the magnitude of the $SJ_{r,\delta}$ functional.

- We propose two-sample confidence intervals for $SW_{r,\delta}(P,Q)$ which have finite-sample coverage under either no assumptions or under minimal moment assumptions. We bound the length of our confidence intervals, showing that they are adaptive both to the magnitude of $SJ_{r,\delta}(P),SJ_{r,\delta}(Q)$.
and to whether or not $P = Q$. The lengths of the intervals we construct match the minimax rates of estimating the Sliced Wasserstein distance, up to polylogarithmic factors.

- We show that the bootstrap is consistent in estimating the distribution of the empirical Sliced Wasserstein distance, away from the null $P = Q$, for all $r > 1$, when certain one-dimensional projections of $P$ and $Q$ are absolutely continuous. We then propose an assumption-lean confidence interval which combines the strengths of our finite-sample intervals and the bootstrap.

- We illustrate our theoretical findings with a simulation study and an application to likelihood-free inference.

**Paper Outline.** The rest of this paper is organized as follows. Section 2 contains background on Wasserstein distances, and defines the trimmed Sliced Wasserstein distance. In Section 3, we define the One-Dimensional Wasserstein Distance. In Section 4, we derive confidence intervals for the one-dimensional and Sliced Wasserstein distances which have finite-sample coverage under minimal assumptions. We discuss asymptotically-valid approaches in Section 5. We illustrate the performance of our confidence intervals via a simulation study in Section 6, and in Section 7, we describe applications of our methodology to likelihood-free inference. We close with discussions in Section 8.

**Notation.** In what follows, given a vector $x = (x_1, ..., x_d) \in \mathbb{R}^d$, $\|x\| = \left(\sum_{i=1}^d x_i^2\right)^{1/2}$ denotes the $\ell_2$ norm of $x$. For any $a, b \in \mathbb{R}$, $a \lor b$ denotes the maximum of $a$ and $b$, and $a \land b$ denotes the minimum of $a$ and $b$. For any function $f$ mapping a set $A$ to $\mathbb{R}$, its supremum norm is denoted by $\|f\|_\infty = \sup_{x \in A} |f(x)|$. For any sequences of real numbers $\{a_n\}_{n=1}^\infty$ and $\{b_n\}_{n=1}^\infty$, we write $a_n \lesssim b_n$ if there exists a constant $C > 0$ such that $a_n \leq C b_n$, and we write $a_n \asymp b_n$ if $a_n \lesssim b_n \lesssim a_n$. For any $x \in \mathbb{R}^d$, $\delta_x$ denotes the Dirac delta measure with mass at $x$. The $d$-dimensional unit ball is denoted by $S^{d-1} = \{u \in \mathbb{R}^d : \|u\| = 1\}$, and $\mu$ denotes the uniform measure on $S^{d-1}$. The Lebesgue measure on $\mathbb{R}^k$, for an integer $k \geq 1$ to be understood from context, is denoted $\lambda$. Given a map $T : \mathbb{R}^d \to \mathbb{R}$ and a Borel probability measure $P$ supported in $\mathbb{R}^d$, $T_# P$ denotes the pushforward of $P$ under $T$, defined by $T_# P(B) = P(T^{-1}(B))$ for all Borel sets $B \subseteq \mathbb{R}^d$. We also denote by $P^\otimes n$ the $n$-fold product measure of $P$. For any set $A \subseteq \mathbb{R}^d$, its diameter is denoted $\text{diam}(A) = \sup\{|x - y| : x, y \in A\}$. For any real numbers $a, b \in \mathbb{R}$, $I(a \leq b)$ is the indicator function equal to 1 if $a \leq b$ and 0 otherwise.

### 2 Background and Related Work

In this section we first provide some background on the Wasserstein distance and its sliced counterpart before turning our attention to a detailed discussion of related work.

#### 2.1 The Wasserstein Distance

Let $\mathcal{P}(\mathcal{X})$ denote the set of Borel probability measures whose support is contained in a set $\mathcal{X} \subseteq \mathbb{R}^d$. For all $r \geq 1$, let $\mathcal{P}_r(\mathcal{X})$ denote the subset of measures in $\mathcal{P}(\mathcal{X})$ admitting a finite $r$-th moment.

**The One-Dimensional Wasserstein Distance.** The infimum in the definition of the Wasserstein distance (1) is always achieved under the present setting of Borel probability measures over Euclidean spaces (cf. Theorem 4.1, Villani (2008)). Closed form expressions for the minimizer are, however, unavailable in general. The one-dimensional case is a key exception.
Let $\mathcal{X} \subseteq \mathbb{R}$, and consider two distributions $P,Q \in \mathcal{P}(\mathcal{X})$. Let $F,G$ denote the cumulative distribution functions (CDFs) of $P$ and $Q$, and denote their respective quantile functions by $F^{-1}$ and $G^{-1}$, where $F^{-1}(u) = \inf\{x \in \mathbb{R} : F(x) \geq u\}$ for all $u \in [0,1]$. We extend $F^{-1}$ to be defined over the entire real line under the convention $F^{-1}(u) = \inf(\mathcal{X})$ for all $u < 0$ and $F^{-1}(u) = \sup(\mathcal{X})$ for all $u > 1$, and similarly for $G^{-1}$. The one-dimensional Wasserstein distance admits the closed form (Bobkov & Ledoux 2019),

$$W_r(P,Q) = \left( \int_0^1 |F^{-1}(u) - G^{-1}(u)|^r \, du \right)^{1/r}. \tag{3}$$

**The $\infty$-Wasserstein Distance.** Let $\mathcal{X} \subseteq \mathbb{R}^d$ be a bounded set, and $P,Q \in \mathcal{P}(\mathcal{X})$. In this case, the limit,

$$W_\infty(P,Q) := \lim_{r \to \infty} W_r(P,Q) = \sup_{r \geq 1} W_r(P,Q) \tag{4}$$

exists, and defines a new metric $W_\infty$ on $\mathcal{P}(\mathcal{X})$. In the special case $\mathcal{X} \subseteq \mathbb{R}$, we have,

$$W_\infty(P,Q) = \sup_{0 \leq u \leq 1} |F^{-1}(u) - G^{-1}(u)|.$$

The relationship $W_r(P,Q) \leq W_\infty(P,Q)$ immediately shows that $W_\infty$ is a stronger metric than $W_r$ for any $r \geq 1$. In fact, it is strictly stronger: The empirical measure $P_n$ based on $n$ i.i.d. observations from the Bernoulli distribution $P$ satisfies $\sup_{r \geq 1} \lim_{n \to \infty} E W_r(P_n,P) = 0$ but $\liminf_{n \to \infty} E W_\infty(P_n,P) > 0$ (Bobkov & Ledoux 2019). In contrast, the metrics $W_r$ induce the same (weak) topology for all $r \geq 1$, when $\text{diam}(\mathcal{X}) < \infty$, as shown by the following interpolation inequalities (Villani 2003):

$$W_r(P,Q) \leq W_s(P,Q) \leq W_r^s(P,Q) \text{diam}(\mathcal{X})^{1 - \frac{r}{s}}, \quad \forall r \leq s. \tag{5}$$

One notable exception where $W_\infty$ also admits such a relationship with respect to $W_r$ metrics was established by Bouchitté et al. (2007): Under certain regularity conditions on $\mathcal{X}$, if $P$ is absolutely continuous with respect to the Lebesgue measure, with strictly positive density $p$ over $\mathcal{X}$, then for all $r > 1$,

$$W_\infty^{r+d}(P,Q) \leq C_{r,d}(\mathcal{X}) W_r(P,Q) \sup_{x \in \mathcal{X}} \left| \frac{1}{p(x)} \right|, \tag{6}$$

where $C_{r,d}(\mathcal{X}) > 0$ is a constant depending only on $p,d,\mathcal{X}$. Moreover, the choice $r=1$ is valid in equation (6) in the case $d=1$.

**The One-Dimensional Trimmed Wasserstein Distance.** Given distributions $P,Q \in \mathcal{P}(\mathbb{R})$ and a trimming constant $\delta \in [0,1/2)$, Munk & Czado (1998) define the $\delta$-trimmed Wasserstein distance (up to rescaling) by

$$W_{r,\delta}(P,Q) = \left( \frac{1}{1 - 2\delta} \int_0^{1-\delta} |F^{-1}(u) - G^{-1}(u)|^r \, du \right)^{1/r}. \tag{7}$$

When $\delta = 0$, $W_{r,\delta}$ reduces to the original Wasserstein distance $W_r$, and when $\delta > 0$, $W_{r,\delta}$ compares the distributions $P$ and $Q$ up to a $2\delta$ fraction of their tail mass, thereby providing a robustification of the Wasserstein distance. Specifically, let $P^\delta$ denote the distribution with CDF

$$F^\delta(x) = \frac{F(x) - \delta}{1 - 2\delta} I(F^{-1}(\delta) \leq x \leq F^{-1}(1 - \delta))$$
and similarly for \( Q^{\delta} \). Then, Álvarez-Esteban et al. (2008) show that

\[
W_{r,\delta}(P,Q) = W_r(P^{\delta},Q^{\delta}).
\]  

(8)

In addition, we define the trimmed \( \infty \)-Wasserstein distance by

\[
W_{\infty,\delta}(P,Q) = \sup_{\delta \leq u \leq 1-\delta} |F^{-1}(u) - G^{-1}(u)|.
\]

2.2 The Sliced Wasserstein Distance

The Sliced Wasserstein distance (Rabin et al. 2011, Bonneel et al. 2015) is defined as the average of Wasserstein distances between one-dimensional projections of the distributions \( P \) and \( Q \). Specifically, let \( S^{d-1} = \{x \in \mathbb{R}^d : \|x\| = 1\} \), and let \( \pi_\theta : x \in \mathbb{R}^d \rightarrow x^\top \theta \), for all \( \theta \in S^{d-1} \). Let \( P_\theta = \pi_\theta \# P \) and \( Q_\theta = \pi_\theta \# Q \), that is, \( P_\theta \) and \( Q_\theta \) are the respective probability distributions of \( X^\top \theta \) and \( Y^\top \theta \), for \( X \sim P \) and \( Y \sim Q \). Let \( \mu \) denote the uniform probability measure on \( S^{d-1} \). The \( r \)-th order Sliced Wasserstein distance between two distributions \( P,Q \in \mathcal{P}_r(\mathbb{R}^d) \) is given by

\[
SW_r(P,Q) = \left( \int_{S^{d-1}} W_r^r(P_\theta,Q_\theta) d\mu(\theta) \right)^{\frac{1}{r}}.
\]

(9)

Since \( P_\theta \) and \( Q_\theta \) are one-dimensional distributions, equation (9) admits the following closed form

\[
SW_r(P,Q) = \left( \int_{S^{d-1}} \int_0^1 |F_\theta^{-1}(u) - G_\theta^{-1}(u)|^r dud\mu(\theta) \right)^{\frac{1}{r}},
\]

(10)

where \( F_\theta^{-1} \) and \( G_\theta^{-1} \) are the respective quantile functions of \( P_\theta \) and \( Q_\theta \). Both integrals of the above expression can be approximated via Monte Carlo sampling from \( S^{d-1} \) and from the unit interval \([0,1]\). This fact makes the computation of the Sliced Wasserstein distance significantly simpler than that of the Wasserstein distance. Moreover, the Sliced Wasserstein distance retains some of the qualitative behaviour of the Wasserstein distance, at least over an appropriate class of distributions. Indeed, Bonnotte (2013) showed that for any distributions \( P,Q \in \mathcal{P}(\{x \in \mathbb{R}^d : \|x\| \leq M\}) \), where \( M > 0 \), we have

\[
SW_r^r(P,Q) \leq c_{d,r} W_r^r(P,Q) \leq C_{d,r} M^{r-1/(d+1)} SW_r^{1/(d+1)}(P,Q),
\]

(11)

where \( C_{d,r} > 0 \) is a constant depending on \( d \) and \( r \), but not depending on \( M \), and \( c_{d,r} = \frac{1}{d} \int_{S^{d-1}} \|\theta\|^r d\mu(\theta) \), which is bounded above by \( 1/d \) whenever \( r \geq 2 \). In particular, it follows that the metrics \( W_r \) and \( SW_r \) are topologically equivalent over \( \mathcal{P}(\mathcal{X}) \) when \( \text{diam}(\mathcal{X}) < \infty \). As we shall see, however, the statistical behaviour of the Wasserstein and Sliced Wasserstein distances can differ dramatically for large dimensions \( d \).

The Trimmed Sliced Wasserstein Distance. In analogy to the trimmed Wasserstein distance in equation (7), we further define the trimmed Sliced Wasserstein distance by

\[
SW_{r,\delta}(P,Q) = \left( \frac{1}{1-2\delta} \int_{S^{d-1}} \int_{\delta}^{1-\delta} |F_\theta^{-1}(u) - G_\theta^{-1}(u)|^r dud\mu(\theta) \right)^{\frac{1}{r}},
\]

(12)

6
for some $\delta \in [0, 1/2)$. We also define the trimmed $\infty$-Sliced Wasserstein distance by

$$SW_{\infty, \delta}(P, Q) := \sup_{r \geq 1} SW_{r, \delta}(P, Q) = \int_{S^{d-1}} W_{\infty, \delta}(P_{\theta}, Q_{\theta}) d\mu(\theta).$$

### 2.3 Related Work

We are unaware of any other work regarding statistical inference for the Sliced Wasserstein distance, except in the special case $d = 1$ when it coincides with the one-dimensional Wasserstein distance. In this case, Munk & Czado (1998) study limiting distributions of the empirical (plug-in) Wasserstein distance estimator, and Freitag et al. (2003, 2007), Freitag & Munk (2005) establish sufficient conditions for the validity of the bootstrap in estimating the distribution of the empirical second-order Wasserstein distance. While these results are very useful, they assume that (i) $P$ and $Q$ are absolutely continuous, (ii) with densities supported on connected sets, and (iii) require different inferential procedures at the classical null ($P = Q$) and away from the null ($P \neq Q$). In contrast, the confidence intervals derived in the present paper are valid under either no assumptions or mild moment assumptions on $P$ and $Q$, and are applied more generally to the Sliced Wasserstein distance in arbitrary dimension. Though our methodology is assumption-light, our confidence intervals are adaptive to (iii), and assumptions (i) and (ii) are closely related to the finiteness of $SJ_{r, \delta}(P), SJ_{r, \delta}(Q)$, to which our confidence intervals are also adaptive.

The Sliced Wasserstein distance is one of many modifications of the Wasserstein distance based on low-dimensional projections. We mention here the Generalized Sliced (Kolouri et al. 2019), Tree-Sliced (Le et al. 2019), max-Sliced (Deshpande et al. 2019), Subspace Robust (Paty & Cuturi 2019, Niles-Weed & Rigollet 2019), and Distributional Sliced (Nguyen et al. 2020) Wasserstein distances. It is also possible to define various other interesting distances by slicing (averaging along univariate projections, Kim et al. (2020)).

Beyond the aforementioned inferential results for the one-dimensional Wasserstein distance, statistical inference for Wasserstein distances over finite or countable spaces has been studied by Sommerfeld & Munk (2018), Tameling et al. (2019), Klatt, Tameling & Munk (2020), Klatt, Munk & Zemel (2020). For distributions with multidimensional support, Rippl et al. (2016) consider the situation where $P$ and $Q$ only differ by a location-scale transformation. Imaizumi et al. (2019) study the validity of the multiplier bootstrap for estimating the distribution of a plug-in estimator of an approximation of the Wasserstein distance. Central Limit Theorems for the Wasserstein distance between empirical measures in general dimension have been established by del Barrio et al. (2019), but with unknown centering constants which are a barrier to using these results for statistical inference.

Rates of convergence for the problem of estimating a distribution under the Wasserstein distance (Dudley 1969, Bolley et al. 2006, Boissard & Le Gouic 2014, Fournier & Guillin 2015, Bobkov & Ledoux 2019, Weed & Bach 2019, Singh & Póczos 2019, Lei 2020) have received significantly more attention than the problem of estimating the Wasserstein distance, the latter being more closely related to our work. Minimax rates of estimating the Wasserstein distance between two distributions have been established by Niles-Weed & Rigollet (2019), as well as by Liang (2019) when $r = 1$. In the special case $d = 1$, where the Sliced Wasserstein distance coincides with the Wasserstein distance, our results refine those of Niles-Weed & Rigollet (2019) by showing that faster rates can be achieved depending on the finiteness of the $SJ_{r, \delta}$ functional, and on the magnitude of $SW_{r, \delta}(P, Q)$.

Likelihood-free inference methodology with respect to the Wasserstein and Sliced Wasserstein distances has recently been developed by Bernton et al. (2019b) and Nadjahi et al. (2020), respectively. In contrast
to these methods, both of which employ approximate Bayesian computation, our work provides frequentist coverage guarantees under minimal assumptions.

3 Minimax Rates for Estimating the Sliced Wasserstein Distance

The goal of this section is to bound the minimax risk of estimating the Sliced Wasserstein distance between two distributions, that is

$$ R_{nm} \equiv R_{nm}(O; r) = \inf_{\hat{S}_{nm}(P,Q) \in O} \mathbb{E}_{P \otimes n \otimes Q \otimes m} \left| \hat{S}_{nm} - SW_{r,\delta}(P,Q) \right|, $$

where the infimum is over all estimators $\hat{S}_{nm}$ of the Sliced Wasserstein distance based on a sample of size $n$ from $P$ and a sample of size $m$ from $Q$, and $O \subseteq \mathcal{P}(\mathbb{R}^d) \times \mathcal{P}(\mathbb{R}^d)$ is an appropriately chosen collection of pairs of distributions. Our motivation for studying this quantity is the observation that $R_{nm}$ lower bounds the minimax length of a confidence interval for the Sliced Wasserstein distance. We construct confidence intervals with matching length in Section 4.

The estimation problem in equation (13) is related to, but distinct from, the problem of estimating a distribution under the Sliced Wasserstein distance. The minimax risk associated with this problem is given by

$$ M_n \equiv M_n(J; r) = \inf_{\hat{P}_n \in J} \mathbb{E}_{P \otimes n} \left\{ SW_{r,\delta}(\hat{P}_n, P) \right\}, $$

where the infimum is over all estimators $\hat{P}_n$ of Borel probability distributions $P$, based on a sample of size $n$ from $P$, and $J \subseteq \mathcal{P}(\mathbb{R}^d)$. Problems (13) and (14) are related as follows: Given estimators $\hat{P}_n$ and $\hat{Q}_m$ for two distributions $P,Q \in \mathcal{P}(\mathbb{R}^d)$, which are minimax-optimal in the sense of equation (14), we have, by the triangle inequality,

$$ R_{nm}(O; r) \leq \mathbb{E} \left| SW_{r,\delta}(\hat{P}_n, \hat{Q}_m) - SW_{r,\delta}(P,Q) \right| $$

$$ \leq \mathbb{E} SW_{r,\delta}(\hat{P}_n, P) + \mathbb{E} SW_{r,\delta}(\hat{Q}_m, Q) $$

$$ \lesssim M_{n\wedge m}(J; r), $$

for suitable families $J$ and $O$ (typically $O := J \times J$). Inequality (16) implies that estimating a distribution under $SW_{r,\delta}$ is a more challenging problem, statistically, than that of estimating the Sliced Wasserstein distance between two distributions. It is unclear, however, whether the rate $M_{n\wedge m}$ is a tight upper bound on $R_{nm}$, or whether the latter can be further reduced. For the Wasserstein distance $W_r$ in general dimension Liang (2019) and Niles-Weed & Rigollet (2019) showed that there is no gap between these minimax risks (ignoring poly-logarithmic factors) for compactly supported distributions.

Let us now briefly summarize the main results of this section. We bound the minimax risks $M_n$ and $R_{nm}$, and we show that there can be a large gap between these minimax risks when the pairs of distributions in $O$ are appropriately separated. In the special case when $d = 1$, $SW_{r,\delta}$ reduces to the one-dimensional (trimmed) Wasserstein distance, and our results imply faster rates than those of Liang (2019) and Niles-Weed & Rigollet (2019), for estimating the Wasserstein distance between distributions which are bounded away from each other. We also emphasize that in stark contrast to the minimax risk for estimating the Wasserstein distance and estimating in the Wasserstein distance, the minimax risks we obtain for the Sliced Wasserstein distance when $d > 1$ are dimension-free.
Though our primary interest is in the quantity $R_{nm}$ (due to its direct connection to confidence intervals) we begin by studying the rate $M_n$ to motivate our choices of families $O$. Inspired by Bobkov & Ledoux (2019), in Section 3.1 we define a functional $SJ_{r,\delta}$, whose magnitude is related to the regularity of the supports of $P$ and $Q$, and whose finiteness implies improved rates of decay for $M_n$. Based on these considerations, we then study the minimax risk $R_{nm}$ over various families $O$ in Section 3.2.

### 3.1 Minimax Rates for Distribution Estimation under $SW_{r,\delta}$

Let $\delta \in [0,1/2)$, $P \in \mathcal{P}(\mathbb{R}^d)$, and let $X_1,\ldots,X_n \sim P$ be an i.i.d. sample. Let $P_n = \frac{1}{n} \sum_{i=1}^n \delta X_i$ denote the corresponding empirical measure. The goal of this section is to characterize the rates of convergence of $P_n$ to the distribution $P$ under the (trimmed) Sliced Wasserstein distance. We focus on upper bounds on the expectation $\mathbb{E}[SW_{r,\delta}(P_n,P)]$, extending the comprehensive treatment by Bobkov & Ledoux (2019) of this quantity with $\delta=0$ in dimension one. We then provide corresponding minimax lower bounds on $M_n$.

For any $\theta \in \mathbb{S}^{d-1}$, let $p_\theta$ denote the density of the absolutely continuous component in the Lebesgue decomposition of the measure $P_\theta$. Recall that $P_\theta$ denotes the probability distribution of $X_1^\top \theta$, with CDF $F_\theta$. Define the functional

$$SJ_{r,\delta}(P) = \frac{1}{1-2\delta} \int_{\mathbb{S}^{d-1}} \int_{F_\theta^{-1}(\delta)} \left[ F_\theta(x)(1-F_\theta(x)) \right]^{r/2} \frac{1}{p_\theta(x)^{r-1}} \, dx \, d\mu(\theta),$$

(17)

with the convention that $0/0=0$. When $d=1$, we write $J_{r,\delta}$ instead of $SJ_{r,\delta}$, and in the untrimmed case $\delta=0$, we omit the subscript $\delta$ and write $SJ_r$ or $J_r$. When $d=1$ and $\delta=0$, Bobkov & Ledoux (2019) prove that the finiteness of $J_r(P)$ is a necessary and sufficient condition for $\mathbb{E}[W_r(P_n,P)]$ to decay at the parametric rate $n^{-1/2}$. The magnitude of $J_r$ is thus closely related to the convergence behaviour of empirical measures under one-dimensional Wasserstein distances, and we show below that the same is true for the $SJ_{r,\delta}$ functional with respect to trimmed Sliced Wasserstein distances.

It can be seen that a necessary condition for the finiteness of $SJ_{r,\delta}(P)$ is that for $\mu$-almost every $\theta \in \mathbb{S}^{d-1}$, the density $p_\theta$ is supported on a (possibly infinite) interval. When $\delta=0$, the value of $SJ_{r,\delta}(P)$ depends in part on the tail behaviour of $P$ and the value of $r$. For example, if $P = N(0,I_d)$ is the standard Gaussian distribution, it can be shown that $SJ_{r,\delta}(P) < \infty$ whenever $\delta > 0$, whereas for $\delta = 0$, $SJ_r(P) < \infty$ if and only if $1 \leq r < 2$ by a similar argument as Bobkov & Ledoux (2019), p. 46. On the other hand, if $P = \frac{1}{2} U(0,\Delta_1) + \frac{1}{2} U(\Delta_2,1)$, for some $0 < \Delta_1 \leq \Delta_2 < 1$, where $U(a,b)$ denotes the uniform distribution on the interval $(a,b) \subseteq \mathbb{R}$, then a direct calculation reveals that $SJ_{r,\delta}(P) < \infty$ if and only if $\Delta_1 = \Delta_2$, for every $\delta \in [0,1/2)$.

We now provide two upper bounds on $\mathbb{E}[SW_{r,\delta}(P_n,P)]$, which are effective when $SJ_{r,\delta}(P) < \infty$ and $SJ_{r,\delta}(P) = \infty$ respectively.

**Proposition 1.** Let $\delta \in [0,1/2)$, and $r \geq 1$.

(i) Suppose $SJ_{r,\delta}(P) < \infty$. Then, there exists a universal constant $C>0$ depending only on $r$ such that

$$\mathbb{E}[SW_{r,\delta}(P_n,P)] \leq \left( \mathbb{E}[SW_{r,\delta}(P_n,P)] \right)^{\frac{1}{2}} \leq C \frac{SJ_{r,\delta}(P)}{\sqrt{n}}.$$

(18)
(ii) Under no conditions on SJ_{r,\delta} we have,
\[ \mathbb{E}[SW_{r,\delta}(P_n,P)] \leq \left( \mathbb{E}[SW_{r,\delta}(P_n,P)] \right)^{\frac{1}{r}} \leq C_{r,\delta}(P)n^{-1/2r}, \]
where \[ C_{r,\delta}(P) = \left( \frac{r^{2r-1}}{1-2\delta} \int_{S^{d-1}} \int_{F_{\theta}^{-1}(\delta)} |x|^{r-1} \sqrt{F_{\theta}(x) - F_{\theta}(x) \mu(\theta)} dx d\mu(\theta) \right)^{\frac{1}{r}}. \]

Remarks.

- Proposition 1(i) implies that the parametric rate of convergence \( n^{-1/2} \) is achievable by the empirical measure provided that \( SJ_{r,\delta}(P) < \infty \).

- Even when the latter condition is violated, Proposition 1(ii) implies that the rate of estimating a distribution under \( SW_{r,\delta} \) is no worse than \( n^{-1/2r} \), assuming that the constant \( C_{r,\delta}(P) \) is finite. This assumption is mild: For example, it follows from Lemma 3 in Appendix A that \( C_{r,\delta}(P) \) is uniformly bounded over the class of distributions \( K_r(b) = \{ P \in \mathcal{P}(\mathbb{R}^d) : \mathbb{E}_{X \sim P}[(X^\top \theta)^2]^{\frac{1}{2}} d\mu(\theta) \leq b \} \), \( b > 0 \), (19)

which contains the set of distributions \( P \) whose second moment \( \mathbb{E}_P[\|X\|^2] \) is bounded above by \( b^{2/r} \). It follows that
\[ \sup_{P \in K_r(b)} \mathbb{E}_{P^\otimes n}[SW_{r,\delta}(P_n,P)] \lesssim n^{-1/2r}. \] (20)

- The rates in equations (18) and (20) do not depend on the dimension \( d \), contrasting generic rates of convergence of the empirical measure under the Wasserstein distance. For instance, if \( P \) is supported on a bounded set in \( \mathbb{R}^d \), Lei (2020) (see also Fournier & Guillin (2015) and Weed & Bach (2019)) shows that
\[ \mathbb{E}[W_r(P_n,P)] \preceq \begin{cases} n^{-1/2r}, & d < 2r \\ n^{-1/2r}(\log n)^{1/r}, & d = 2r \\ n^{-1/d}, & d > 2r. \end{cases} \] (21)

The convergence of the empirical measure under \( W_r \) thus exhibits a poor dependence on the dimension \( d \) when \( d > 2r \). In fact, the rate achieved in equation (21) when \( d \leq 2r \) is the rate achieved under \( SW_{r,\delta} \), in equation (20), for any \( d \). These considerations show that estimating a distribution in the Sliced Wasserstein distance does not suffer from the curse of dimensionality despite metrizing the same topology on \( \mathcal{P}(\mathbb{R}^d) \)—see equation (11).

- Finally, we note that the proof of Proposition 1 follows from a straightforward generalization of Theorem 5.3 and Theorem 7.16 of Bobkov & Ledoux (2019), and is therefore omitted.

We close this subsection by bounding the minimax risk \( M_n \) in equation (14). In view of Proposition 1 and equation (20), it is natural to carry out our analysis over the following class of distributions,
\[ \mathcal{J}(s) = \{ P \in K_r(b) : SJ_{r,\delta}(P) \leq s \}, \quad s \in [0, \infty]. \]
Proposition 2. Let \( r \geq 1 \). Then, for any \( s \in [0, \infty) \), there exist constants \( C_1, C_2 > 0 \) depending on \( b, r, \delta, s \) such that for all \( s_1 \leq s \),

\[
\mathcal{M}_n(\mathcal{J}(s_1); r) \geq C_1 s_1^{1/2} n^{-1/2}, \quad \text{and} \quad \mathcal{M}_n(\mathcal{J}(\infty); r) \geq C_2 n^{-1/2r}.
\]

Proposition 2 implies that the rates of convergence achieved by the empirical measure in Proposition 1 are minimax optimal. The proof of this result will follow as a special case of our minimax lower bounds for estimating the Sliced Wasserstein distance, to which we turn our attention next.

3.2 Minimax Estimation of the Sliced Wasserstein Distance

In this section, we bound the minimax risk \( R_{nm} \) of estimating the Sliced Wasserstein distance between two distributions, as defined in equation (13). We begin by providing upper bounds on the estimation error of the empirical Sliced Wasserstein distance, \( \text{SW}_{r, \delta}(P_n, Q_m) \). Recall that, \( P_n \) and \( Q_m \) denote the empirical measures of i.i.d. samples \( X_1, \ldots, X_n \sim P \) and \( Y_1, \ldots, Y_m \sim Q \) respectively.

Proposition 3. Suppose \( P, Q \in \mathcal{K}_r(b) \). Let \( \delta \in (0, 1/2) \) be arbitrary but fixed.

\( \quad (i) \) There exists a universal constant \( c_1 > 0 \), possibly depending on \( r, b, \delta \), such that

\[
E|\text{SW}_{r, \delta}(P_n, Q_m) - \text{SW}_{r, \delta}(P, Q)| \leq c_1 \left( n^{-1/2} + m^{-1/2} \right) \wedge \left( \frac{\text{SJ}_{r, \delta}^1(P)}{\sqrt{n}} + \frac{\text{SJ}_{r, \delta}^1(Q)}{\sqrt{m}} \right).
\]

\( \quad (ii) \) Suppose \( \text{SW}_{r, \delta}(P, Q) \geq \Gamma \), for some real number \( \Gamma > 0 \). Then, there exists a universal constant \( c_2 > 0 \), possibly depending on \( \Gamma, b, r, \delta \), such that

\[
E|\text{SW}_{r, \delta}(P_n, Q_m) - \text{SW}_{r, \delta}(P, Q)| \leq c_2 \left( \frac{1}{\sqrt{n}} + \frac{1}{\sqrt{m}} \right).
\]

Remarks.

- Proposition 3(i) is an immediate consequence of inequality (16), which implies that the rate of estimating the Sliced Wasserstein distance with the plug-in estimator \( \text{SW}_{r, \delta}(P_n, Q_m) \) is no worse than the rate of convergence of the empirical measure under \( \text{SW}_{r, \delta} \), which was established in Proposition 1. In particular, these results show that the parametric rate \( n^{-1/2} + m^{-1/2} \) for estimating \( \text{SW}_{r, \delta} \) is achievable for distributions satisfying \( \text{SJ}_{r, \delta}^1(P), \text{SJ}_{r, \delta}^1(Q) < \infty \).

- On the other hand, Proposition 3(ii) implies that the parametric rate of estimating \( \text{SW}_{r, \delta}(P, Q) \) is always achievable when \( P \) and \( Q \) are bounded away from each other by a positive constant, in the distance \( \text{SW}_{r, \delta} \). This fast rate of convergence is obtained irrespective of the values of \( \text{SJ}_{r, \delta}^1(P) \) and \( \text{SJ}_{r, \delta}^1(Q) \). Discrepancies between rates of convergence at the null \( (P = Q) \) and away from the null \( (P \neq Q) \) have previously been noted by Sommerfeld & Munk (2018) for Wasserstein distances over finite spaces—indeed, their rates match those of Proposition 3 when \( \text{SJ}_{r, \delta}^1(P), \text{SJ}_{r, \delta}^1(Q) = \infty \).
Finally, we note that the natural estimator \( \text{SW}_{r,\delta}(P_n, Q_n) \) is \textit{adaptive} to the typically unknown \( \text{SJ}_{r,\delta} \) functionals of \( P \) and \( Q \), and does not require the statistician to specify if \( P = Q \) or \( P \neq Q \). Instead, the estimator adapts and yields favorable rates in favorable situations—when either the \( \text{SJ}_{r,\delta} \) functionals are finite, or when \( P \) and \( Q \) are sufficiently well-separated.

A natural question in this context is to study the optimality of the plug-in estimator described above. We now provide corresponding lower bounds on the minimax risk \( R_{nm} \). Inspired by Proposition 3, we define the following collection of pairs of distributions,

\[
\mathcal{O}(\Gamma; s_1, s_2) = \{(P, Q) \in \mathcal{K}_r^2(b) : \text{SJ}_{r,\delta}(P) \leq s_1, \text{SJ}_{r,\delta}(Q) \leq s_2, \text{SW}_{r,\delta}(P, Q) \geq \Gamma\}, \quad \Gamma \geq 0,
\]

where \( s_1, s_2 \in [0, \infty] \). In constructing our lower bounds, one minor concern that we have to address is that if \( \Gamma \) is chosen sufficiently large relative to \( b \) then the class \( \mathcal{O}(\Gamma; s_1, s_2) \) might be empty. Consequently, we always assume that,

\[
\Gamma' \leq c_r b, \tag{22}
\]

for some sufficiently small constant \( c_r > 0 \) depending on \( r \). With these definitions in place we now state our minimax lower bounds on the risk \( R_{nm} \).

**Theorem 1.** Let \( r \geq 1 \) and \( \delta \in (0, 1/2) \). Fix an arbitrary real number \( s > 0 \), and assume \( b \geq (2s)^{1/r} \).

(i) For any real numbers \( s_1, s_2 \geq 0 \), there exists a constant \( C_1 > 0 \) depending on \( \delta, r, b, s \) such that

\[
R_{nm}(\mathcal{O}(0; s_1, s_2); r) \geq C_1 \begin{cases} n^{-\frac{1}{2r}} + m^{-\frac{1}{2r}}, & s_1 \vee s_2 = \infty \\ \frac{s_1^{\frac{1}{r}}}{\sqrt{n}} + \frac{s_2^{\frac{1}{r}}}{\sqrt{m}}, & s_1 \vee s_2 \leq s \end{cases}.
\]

(ii) Let \( s_1 = s_2 = \infty \). For any \( \Gamma > 0 \) satisfying equation (22), there exists a constant \( C_2 > 0 \) depending on \( \delta, r, b, \Gamma \) such that

\[
R_{nm}(\mathcal{O}(\Gamma; s_1, s_2); r) \geq C_2 \left( \frac{1}{\sqrt{n}} + \frac{1}{\sqrt{m}} \right).
\]

Remarks.

- Theorem 1 implies that the rates achieved by the empirical Sliced Wasserstein distance \( \text{SW}_r(P_n, Q_m) \) in Proposition 3, as well as their dependence on the \( \text{SJ}_{r,\delta} \) functional, are minimax optimal.

- We defer the proof of this result to Appendix C. This result is proved by a standard information-theoretic technique of constructing several pairs of distributions (each satisfying the various hypotheses). These distributions are carefully chosen to have small Kullback-Leibler divergence, but have very different Sliced Wasserstein distances. We then obtain lower bounds via an application of Le Cam’s inequality (see, for instance, Theorem 2.2 of Tsybakov (2008)). Beyond this careful choice of pairs of distributions, the bulk of our technical effort is in computing or tightly bounding the various Sliced Wasserstein distances (see Lemma 5 in the Appendix).

In what follows, we construct finite-sample confidence intervals for \( \text{SW}_{r,\delta}(P, Q) \) whose lengths achieve these same rates of convergence (up to polylogarithmic factors), and which are adaptive to the magnitude of \( \text{SJ}_{r,\delta} \).
4 Confidence Intervals for the Sliced Wasserstein Distance

In this section, we introduce and study our assumption-light confidence intervals for the one-dimensional and Sliced Wasserstein distances.

4.1 Finite-Sample Confidence Intervals in One-Dimension

Throughout this subsection, let \( r \geq 1 \) be given, let \( P, Q \in \mathcal{P}(\mathbb{R}) \) be probability distributions with respective CDFs \( F, G \), and let \( X_1,...,X_n \sim P \) and \( Y_1,...,Y_m \sim Q \) be i.i.d. samples. Let \( F_n(x) = \frac{1}{n} \sum_{i=1}^{n} I(X_i \leq x) \) and \( G_m(x) = \frac{1}{m} \sum_{j=1}^{m} I(Y_j \leq x) \) denote their corresponding empirical CDFs, for all \( x \in \mathbb{R} \). We derive confidence intervals \( C_{nm} \subseteq \mathbb{R} \) for the \( \delta \)-trimmed Wasserstein distance, with the following non-asymptotic coverage guarantee

\[
\inf_{P,Q \in \mathcal{P}(\mathbb{R})} \mathbb{P}(W_{r,\delta}(P,Q) \in C_{nm}) \geq 1 - \alpha, \tag{23}
\]

for some pre-specified level \( \alpha \in (0,1) \). Our approach hinges on the fact that the one-dimensional Wasserstein distance may be expressed as the \( L^1 \) norm of the quantile functions of \( P \) and \( Q \), suggesting that a confidence interval may be derived via uniform control of the empirical quantile process. Specifically, the starting point for our confidence intervals is a confidence band of the form

\[
\inf_{P \in \mathcal{P}(\mathbb{R})} \mathbb{P}\left( F_n^{-1}(\gamma_{\alpha,n}(u)) \leq F^{-1}(u) \leq F_n^{-1}(\eta_{\alpha,n}(u)), \forall u \in [\delta,1-\delta] \right) \geq 1 - \alpha/2, \tag{24}
\]

for some sequences of functions \( \gamma_{\alpha,n}, \eta_{\alpha,n} : [\delta,1-\delta] \to \mathbb{R} \). The study of uniform quantile bounds of the form (24) is a classical topic (see for instance, the book of Shorack & Wellner (2009)). We discuss two prominent examples that will form the basis of our development.

Example 1. By the Dvoretzky-Kiefer-Wolfowitz (DKW) inequality (Dvoretzky et al. 1956, Massart 1990), we have

\[
\mathbb{P}\left( |F_n(x) - F(x)| \leq \beta_n, \forall x \in \mathbb{R} \right) \geq 1 - \frac{\alpha}{2}, \quad \beta_n = \sqrt{\frac{1}{2n} \log(4/\alpha)}. \tag{25}
\]

Inverting this inequality leads to the choice

\[
\gamma_{\alpha,n}(u) = u - \beta_n, \quad \eta_{\alpha,n}(u) = u + \beta_n, \quad u \in [\delta,1-\delta]. \tag{26}
\]

Example 2. Scale-dependent choices of \( \gamma_{\alpha,n} \) and \( \eta_{\alpha,n} \) may be obtained via the relative Vapnik-Chervonenkis (VC) inequality (Vapnik 2013). The latter implies the inequality

\[
\mathbb{P}\left( |F_n(x) - F(x)| \leq \nu_{\alpha,n} \sqrt{F_n(x)(1-F_n(x))}, \forall x \in \mathbb{R} \right) \geq 1 - \frac{\alpha}{2}, \tag{27}
\]

where \( \nu_{\alpha,n} := \sqrt{\frac{16}{n} \log(16/\alpha) + \log(2n+1)} \). As shown in Appendix G.5, inverting inequality (27) leads to the choice

\[
\gamma_{\alpha,n}(u) = \frac{2u + \nu_{\alpha,n}^2 - \nu_{\alpha,n} \sqrt{\nu_{\alpha,n}^2 + 4u(1-u)}}{2(1+\nu_{\alpha,n}^2)}, \quad \eta_{\alpha,n}(u) = \frac{2u + \nu_{\alpha,n}^2 + \nu_{\alpha,n} \sqrt{\nu_{\alpha,n}^2 + 4u(1-u)}}{2(1-\nu_{\alpha,n}^2)}, \tag{28}
\]

for all \( u \in [\delta,1-\delta] \).
Given sequences of functions \( \gamma_{\alpha,n}, \eta_{\alpha,n} \) satisfying equation (24), it is easy to see that with probability at least \( 1-\alpha \),
\[
    A_{nm}(u) \leq |F^{-1}(u) - G^{-1}(u)| \leq B_{nm}(u), \quad \forall u \in [\delta, 1-\delta],
\]
where,
\[
    A_{nm}(u) = \left[ F_n^{-1}(\gamma_{\alpha,n}(u)) - G_m^{-1}(\eta_{\alpha,m}(u)) \right] \vee \left[ G_m^{-1}(\gamma_{\alpha,m}(u)) - F_n^{-1}(\eta_{\alpha,n}(u)) \right] \lor 0,
\]
\[
    B_{nm}(u) = \left[ F_n^{-1}(\eta_{\alpha,n}(u)) - G_m^{-1}(\gamma_{\alpha,m}(u)) \right] \vee \left[ G_m^{-1}(\eta_{\alpha,m}(u)) - F_n^{-1}(\gamma_{\alpha,n}(u)) \right].
\]
This observation readily leads to the following Proposition.

**Proposition 4.** Let \( \delta \in (0, 1/2) \) and \( r \geq 1 \) be fixed. The interval
\[
    C_{nm} = \left[ \frac{1}{1-2\delta} \int_{\delta}^{1-\delta} A_{nm}(u) \, du \right]^{\frac{1}{r}}, \left[ \frac{1}{1-2\delta} \int_{\delta}^{1-\delta} B_{nm}(u) \, du \right]^{\frac{1}{r}}, \tag{29}
\]
satisfies
\[
    \inf_{P,Q \in \mathcal{P}(\mathbb{R})} \mathbb{P}(W_{r,\delta}(P,Q) \in C_{nm}) \geq 1-\alpha.
\]

Proposition 4 establishes the finite-sample coverage of the confidence interval \( C_{nm} \), under no assumptions on the distributions \( P,Q \). We emphasize, however, that for distributions \( P,Q \) with unbounded support, the interval \( C_{nm} \) only has finite length under the following condition,

**A1(\( \delta; \alpha \))** The trimming constant \( \delta \) satisfies
\[
    \gamma_{\alpha,n \land m}(\delta) > 0 \quad \text{and} \quad \eta_{\alpha,n \land m}(1-\delta) < 1.
\]

For instance, if \( \gamma_{\alpha,n}, \eta_{\alpha,n} \) are chosen via the DKW inequality (26), these inequalities imply the choice \( \delta \gtrsim (n \land m)^{-1/2} \). This excludes the untrimmed case \( \delta = 0 \), for which statistical inference for the Wasserstein distance is not possible without any assumptions on the tail behaviour of \( P \) and \( Q \). Indeed, the untrimmed \( r \)-Wasserstein distance itself fails to be well-defined for distributions \( P,Q \) which do not admit \( r \) moments. Under explicit tail assumptions on the distributions \( P \) and \( Q \), the interval \( C_{nm} \) may be modified to cover the untrimmed Wasserstein distance. We do not pursue this avenue here, and assume that \( \delta \) is chosen to satisfy **A1(\( \delta; \alpha \))**. In what follows, we build upon the interval \( C_{nm} \) and consider confidence intervals for the Sliced Wasserstein distance in general dimension \( d \). We then study the statistical properties of these confidence intervals, with a particular focus on their minimax-optimality and adaptivity.

### 4.2 Finite-Sample Confidence Intervals in General Dimension

We now use Proposition 4 to derive a confidence interval for \( \text{SW}_{r,\delta}(P,Q) \), where \( P,Q \in \mathcal{P}(\mathbb{R}^d) \). In analogy to the previous subsection, an immediate first approach is to choose functions \( \overline{\gamma}_{\alpha,n} \) and \( \overline{\eta}_{\alpha,n} \) such that
\[
    \inf_{P \in \mathcal{P}(\mathbb{R}^d)} \mathbb{P}\left( F_{\theta,n}^{-1}(\overline{\gamma}_{\alpha,n}(u)) \leq F_{\theta}^{-1}(u) \leq F_{\theta,n}^{-1}(\overline{\eta}_{\alpha,n}(u)), \quad \forall u \in [\delta, 1-\delta], \quad \forall \theta \in \mathbb{S}^{d-1} \right) \geq 1-\alpha/2, \tag{30}
\]
where \( F_{\theta,n} : x \in \mathbb{R} \mapsto \frac{1}{n} \sum_{i=1}^{n} I(X_i \leq x) \) for all \( \theta \in \mathbb{S}^{d-1} \), and \( F_{\theta}^{-1} \) denotes the quantile function of \( P_{\theta} = \pi_{\theta \#} P \). Such a bound can be obtained, for instance, by an application of the VC inequality (Vapnik
over the set of half-spaces in $\mathbb{R}^d$. An assumption-light confidence interval for $SW_{r,\delta}(P,Q)$ with finite-sample coverage may then be constructed by following the same lines as in the previous section. Due to the uniformity of equation (30) over the unit sphere, however, it can be seen that the length of such an interval is necessarily dimension-dependent. In what follows, we instead show that it is possible to obtain a confidence interval with dimension-independent length by exploiting the fact that the Sliced Wasserstein distance is a mean with respect to the distribution $\mu$.

Let $\theta_1, \ldots, \theta_N$ be an i.i.d. sample from the distribution $\mu$, for some integer $N \geq 1$, and let $\mu_N = \frac{1}{N} \sum_{i=1}^{N} \delta_{\theta_i}$ denote the corresponding empirical measure. Consider the following Monte Carlo approximation of the Sliced Wasserstein distance between the distributions $P$ and $Q$,

$$SW_{r,\delta}^{(N)}(P,Q) = \left( \int_{S_{d-1}} W_{r,\delta}(P_{\theta},Q_{\theta}) d\mu_N(\theta) \right)^{\frac{1}{r}} = \left( \frac{1}{N} \sum_{j=1}^{N} W_{r,\delta}^{T}(P_{\theta_j},Q_{\theta_j}) \right)^{\frac{1}{r}}. $$

For any $\theta \in S_{d-1}$, let $[L_{N,nm}(\theta), U_{N,nm}(\theta)]$ be the confidence interval in equation (29) for $W_{r,\delta}(P_{\theta},Q_{\theta})$, at level $1 - \alpha/N$. Let

$$C_{nm}^{(N)} = \left[ \left( \int_{S_{d-1}} L_{N,nm}(\theta) d\mu_N(\theta) \right)^{\frac{1}{r}}, \left( \int_{S_{d-1}} U_{N,nm}(\theta) d\mu_N(\theta) \right)^{\frac{1}{r}} \right].$$

By a simple Bonferroni correction, we can obtain conditional coverage of the (random) parameter $SW_{r,\delta}^{(N)}(P,Q)$, i.e.

$$\inf_{P,Q \in \mathcal{P}(\mathbb{R}^d)} \mathbb{P}\left( SW_{r,\delta}^{(N)}(P,Q) \in C_{nm}^{(N)} \mid \theta_1, \ldots, \theta_N \right) \geq 1 - \alpha.$$ 

The following result, ensuring coverage of the (trimmed) Sliced Wasserstein distance, is now a straightforward consequence of the Central Limit Theorem, when the dimension $d$ is fixed.

**Proposition 5.** Let $\delta \in (0,1/2)$, $b > 0$, and $r,d \geq 1$ be fixed. Then,

$$\liminf_{N \to \infty} \inf_{P,Q \in \mathcal{K}_{r,d}(b)} \mathbb{P}\left( SW_{r,\delta}(P,Q) \in C_{nm}^{(N)} \right) \geq 1 - \alpha.$$ 

Although the coverage of $C_{nm}^{(N)}$ requires almost no assumptions on $P$ and $Q$, we now show that its length achieves the minimax rates established in Theorem 1, and is adaptive to the magnitude of $SJ_{r,\delta}(P), SJ_{r,\delta}(Q)$.

### 4.2.1 Bounds on the Length of $C_{nm}^{(N)}$

In this section, we provide upper bounds on the length of the interval $C_{nm}^{(N)}$. We begin with a general result (Theorem 2) which provides upper bounds with respect to the functions $\gamma_{\alpha,n}$ and $\eta_{\alpha,n}$. We subsequently specialize this general result through Corollaries 1, 2 (and Corollary 3 in Appendix H) to illustrate the different rates of convergence which can be obtained under various choices of these functions, and under various conditions on the underlying distributions.

In what follows, we assume $\gamma_{\alpha,n}$ and $\eta_{\alpha,n}$ are both differentiable, invertible with differentiable inverses over $[\delta,1-\delta]$, and are respectively increasing and decreasing as functions of $\alpha$. Given $\epsilon \in (0,1)$, for
The proof is a straightforward consequence of Examples 1, 2, and their derivations in Appendix G.5. For any fixed \( n \), quantities which follow. For any fixed \( n \) decays to zero as \( n \to \infty \). It can be seen that \( U \) are satisfied by the choices of assumptions. Here, recall \( B1 \) and \( B3 \) regarding the functions \( \gamma, \eta, \kappa, \epsilon, \sigma \) and \( \theta \), \( \kappa, \epsilon, \sigma \) are chosen as in equation (26), then

\[
\kappa_{\epsilon,n} \leq \frac{2}{\sqrt{2n}} \log(4/\epsilon) \quad \text{and} \quad V_{\epsilon,n}(P) \leq \kappa_{\epsilon,n}^{-1} \int_{S^{d-1}} \int_{\delta}^{1-\delta} \left( p_\theta(F^{-1}_\theta(u)) \right)^r \, d\mu(\theta).
\]

2. If \( \gamma, \eta, \kappa, \epsilon, \sigma \) are chosen as in equation (28), then for a universal constant \( k > 0 \),

\[
\kappa_{\epsilon,n} \leq k \kappa_{\epsilon,n}, \quad \text{and} \quad V_{\epsilon,n}(P) \leq \kappa_{\epsilon,n} S J_{r,\delta}(P).
\]

The proof is a straightforward consequence of Examples 1, 2, and their derivations in Appendix G.5, and is therefore omitted. We now define the functional

\[
U_{\epsilon,n}(P) = \frac{1}{1-2\delta} \int_{S^{d-1}} \left( \sup_{\delta \leq u \leq 1-\delta} |F^{-1}_\theta(u+h)-F^{-1}_\theta(u)|^{r-1} \right) d\mu(\theta).
\]

It can be seen that \( U_{\epsilon,n}(P) \) is an upper bound on the magnitude of the largest jump discontinuity of the quantile function \( F^{-1}_\theta \), averaged over directions \( \theta \in S^{d-1} \). When \( S J_{r,\delta}(P) < \infty \), the quantile function \( F^{-1}_\theta \) is absolutely continuous for almost all \( \theta \in S^{d-1} \) (see Lemma 2 in Appendix A), implying that \( U_{\epsilon,n}(P) \) decays to zero as \( n \to \infty \). The lengths of our confidence intervals will now ultimately depend on the quantities which follow. For any fixed \( s > 0 \), define

\[
\psi_{\epsilon,n,m} = \begin{cases} 
(\text{SW}_{r,\delta}(P,Q)+V_{\epsilon,n}(P)+V_{\epsilon,m}(Q))^{r-1}[V_{\epsilon,n}(P)]^{\frac{1}{r}}, & S J_{r,\delta}(P) \vee S J_{r,\delta}(Q) \leq s \\
(\text{SW}_{\infty,\delta}(P,Q)+U_{\epsilon,n}(P)+U_{\epsilon,m}(Q)) \kappa_{\epsilon,n}, & S J_{r,\delta}(P) \vee S J_{r,\delta}(Q) > s
\end{cases},
\]

and

\[
\varphi_{\epsilon,n,m} = \begin{cases} 
(\text{SW}_{r,\delta}(P,Q)+V_{\epsilon,n}(P)+V_{\epsilon,m}(Q))^{r-1}[V_{\epsilon,m}(Q)]^{\frac{1}{r}}, & S J_{r,\delta}(P) \vee S J_{r,\delta}(Q) \leq s \\
(\text{SW}_{\infty,\delta}(P,Q)+U_{\epsilon,n}(P)+U_{\epsilon,m}(Q)) \kappa_{\epsilon,m}, & S J_{r,\delta}(P) \vee S J_{r,\delta}(Q) > s
\end{cases}.
\]
With this notation in place, we arrive at the following upper bound on the length of \( C_{nm}^{(N)} \).

**Theorem 2** (Length of \( C_{nm}^{(N)} \)). Let \( r \geq 1, \delta \in (0,1/2), \alpha \in (0,1), \) and \( P,Q \in K_{2\varphi} \) for some \( b > 0 \). Assume conditions **A1**(\( \delta/2; \alpha/N \)), **B1-B3** hold. Then, there exist constants \( c,c_1 > 0 \) depending only on \( s,b,r,\delta \), and a random variable \( \kappa_{N} \) depending on \( \mu_{N} \), such that \( \mathbb{E}_{\mu^{\otimes N}} [\kappa_{N}] \leq c_1 N^{-1/2r} I(d \geq 2) \), and such that for all \( \varepsilon \in (0,1) \), with probability at least \( 1 - \varepsilon \),

\[
\lambda(C_{nm}^{(N)}) \leq \left\{ SW_{r,\delta}(P,Q) + c(\psi_{\varepsilon, nm} + \varphi_{\varepsilon, nm} + \kappa_{N}) \right\}^{1/r} - SW_{r,\delta}(P,Q).
\]

At a high-level the presence of Sliced Wasserstein distances \( SW_{\infty,\delta}(P,Q) \) or \( SW_{r,\delta}(P,Q) \) in equations (32), (33) implies distinct rates of decay for \( \lambda(C_{nm}^{(N)}) \), depending on whether \( P,Q \) are near the classical null \( P = Q \) or far from it. The fact that \( SW_{\infty,\delta} \) is a stronger metric than \( SW_{r,\delta} \), and the presence of the functional \( U_{\varepsilon,n} \), will imply a second dichotomy in the rate of \( \lambda(C_{nm}^{(N)}) \), based on whether or not \( SJ_{r,\delta}(P) \lor SJ_{r,\delta}(Q) \leq s \), as we shall explore in the sequel. We emphasize that \( s \) is an arbitrary positive constant, which the constants \( c,c_1 \) in Theorem 2 depend upon, thus the condition \( SJ_{r,\delta}(P) \lor SJ_{r,\delta}(Q) \leq s \) may be replaced by \( SJ_{r,\delta}(P) \lor SJ_{r,\delta}(Q) \leq s \) without changing the rates in Theorem 2, at the expense of introducing constants depending (possibly unboundedly) on \( SJ_{r,\delta}(P), SJ_{r,\delta}(Q) \).

We consider several special cases of Theorem 2 to illustrate its implications. The following result specializes Theorem 2 to Examples 1,2.

**Corollary 1.** Let \( P,Q \in K_{2\varphi} \) be arbitrary, and assume the conditions of Theorem 2. Let \( \varepsilon \in (0,1) \).

(i) Suppose \( \gamma_{\varepsilon,n}, \eta_{\varepsilon,n} \) are chosen as in Example 1. Then, with probability at least \( 1 - \varepsilon \),

\[
\lambda(C_{nm}^{(N)}) \lesssim \frac{1}{N} \left\{ \log(N/\varepsilon) \left( n^{-\frac{1}{2r}} + m^{-\frac{1}{2r}} \right), \quad SJ_{r,\delta}(P) \lor SJ_{r,\delta}(Q) > s \right\}.
\]

(ii) Suppose \( \gamma_{\varepsilon,n}, \eta_{\varepsilon,n} \) are chosen as in Example 2. Then, with probability at least \( 1 - \varepsilon \),

\[
\lambda(C_{nm}^{(N)}) \lesssim \frac{1}{N} \left\{ \log(n \land m) + \log(N/\varepsilon) \left( \frac{1}{2} \left( \frac{1}{\sqrt{m}} + \frac{1}{\sqrt{n}} \right), \quad SJ_{r,\delta}(P) \lor SJ_{r,\delta}(Q) > s \right\}
\]

Corollary 1(i) shows that the length of the DKW-based interval achieves the minimax optimal rate, implied by Theorem 1(i), up to a polylogarithmic factor in \( N \), and up to the approximation error \( \kappa_{N} \). It does not, however, depend on the magnitude of \( SJ_{r,\delta}(P) \) and \( SJ_{r,\delta}(Q) \), when the latter are less than \( s \), as could have been expected from Proposition 3(i) and Theorem 1(i). This is a consequence of the DKW inequality not adapting to the variance of the distributions therein. On the other hand, Corollary 1(ii) shows that the length of the relative VC-based interval does depend on the magnitude of the \( SJ_{r,\delta} \) functional, at the expense of a polylogarithmic term in \( n,m \).

When the distributions \( P \) and \( Q \) are assumed to be bounded away from each other in \( SW_{r,\delta} \), Theorem 1(ii) suggests that the nonparametric rate \( n^{-\frac{1}{2r}} + m^{-\frac{1}{2r}} \) in Corollary 1 is improvable. This is indeed the case, as shown below.
Corollary 2. Suppose two distributions $P,Q \in \mathcal{K}_2$ (b) satisfy $SW_{r,\delta}(P,Q) \geq \Gamma$ for some constant $\Gamma > 0$. Then, under the assumptions of Theorem 2, for all $\epsilon \in (0,1)$, we have with probability at least $1-\epsilon$,

$$\lambda(C^{(N)}_{nm}) \lesssim \epsilon N + \begin{cases} \kappa_{\epsilon,n} + \kappa_{\epsilon,m}, & SJ_{r,\delta}(P) \vee SJ_{r,\delta}(Q) > s \\ V_{\epsilon,m}(P) + V_{\epsilon,m}(Q), & SJ_{r,\delta}(P) \vee SJ_{r,\delta}(Q) \leq s. \end{cases}$$

For example, when $\gamma_{\epsilon,n,\eta_{\epsilon,m}}$ are based on the DKW inequality (Example 1), Corollary 2 implies that the length of $C^{(N)}_{nm}$ achieves the parametric rate $n^{-1/2} + m^{-1/2}$ with high probability (again, ignoring factors depending only on $N$), under the mere condition that $P$ and $Q$ are bounded away from each other. Theorem 1(ii) implies that this rate is minimax-optimal. As before, adaptivity to the magnitudes of $SJ_{r,\delta}(P), SJ_{r,\delta}(Q)$ is available using the relative VC-based interval in Example 2.

We close this section by briefly commenting on the situation where $SJ_{r,\delta}(P) \leq s$ but $SJ_{r,\delta}(Q) = \infty$. In this case, Corollary 2 continues to imply that the parametric rate is achievable for distributions $P$ and $Q$ which are bounded away from each other. When $P$ and $Q$ are permitted to approach each other, the conditions $SJ_{r,\delta}(P) \leq s, SJ_{r,\delta}(Q) = \infty$ need not be violated so long as $P \neq Q$. In this specialized situation, Corollary 1 merely implies the nonparametric rate $n^{-\frac{1}{2}} + m^{-\frac{1}{2}}$, though one may expect a better dependence on $n$ as a result of the condition $SJ_{r,\delta}(P) \leq s$, as well as a better dependence on $m$ by virtue of $Q$ being in proximity to $P$. By appealing to equation (6), it turns out that such a result indeed holds, and is further discussed in Appendix II.

5 Asymptotic Confidence Intervals and a Hybrid Bootstrap Approach

We now discuss several existing asymptotic confidence intervals for the one-dimensional Wasserstein distance, their extensions to the Sliced Wasserstein distance, and we compare them to our finite-sample confidence intervals in Section 4. We then discuss how the strengths of these methods can be combined.

In the context of goodness-of-fit testing, Munk & Czado (1998) (see also del Barrio et al. (2018)) prove central limit theorems of the form

$$\sqrt{\frac{nm}{n+m}} \left\{ W_{2,\delta}(P_n, Q_m) - W_{2,\delta}(P,Q) \right\} \overset{\text{d}}{\sim} N(0, \sigma^2), \quad (34)$$

as $n,m \to \infty$ such that $n/(n+m) \to a \in (0,1)$, for distributions $P,Q \in \mathbb{R}$ and for some $\sigma > 0$. They also construct a consistent estimator $\hat{\sigma}$ of $\sigma$. This result assumes, in particular, that $P \neq Q$, and that each of the distributions $P$ and $Q$ satisfy the following smoothness condition,

(C) $F$ is twice continuously differentiable, with density $p$, which is strictly positive over the interval $[F^{-1}(\delta), F^{-1}(1-\delta)]$. Moreover,

$$\sup_{x \in [F^{-1}(\delta), F^{-1}(1-\delta)]} \frac{F'(x)(1-F(x))}{p'(x)p(x)} < \infty.$$

Assumption (C) originates from strong approximation theorems for the empirical quantile process (Csorgo & Revesz 1978), and entails that $P$ and $Q$ have differentiable densities, whose supports are intervals.
Under the weaker assumption that $P$ and $Q$ merely admit densities with respect to the Lebesgue measure supported on an interval, and still retaining the assumption that $P \neq Q$, Freitag et al. (2003, 2007), Freitag & Munk (2005) prove the consistency of the bootstrap in estimating the distribution of $W_2^2(P, Q_m)$. The Wasserstein distance is well-defined between any pairs of (possibly mutually singular) distributions with sufficient moments, unlike other classical metrics between probability distributions such as the Hellinger and $L^1$ metrics. Indeed, this feature of the Wasserstein distance is a primary motivation for its use in statistical applications. Smoothness assumptions such as (C) are therefore prohibitive in inferential problems for the Wasserstein distance, and motivated our development of assumption-light confidence intervals in the previous section. Nevertheless, when a smoothness assumption such as (C) happens to hold, asymptotic confidence intervals based on limit laws such as (34), or those based on the bootstrap, may have shorter length than those described in Section 4.

In what follows, our goal is to combine the strengths of the bootstrap with those of our finite-sample intervals. We start by proving in Section 5.1 that, when $P \neq Q$ admit absolutely continuous projections, the bootstrap is valid in estimating the distribution of $SW_{r, \delta}(P_n, Q_m)$ for all $r > 1$, thereby generalizing the results of Freitag et al. (2003, 2007), Freitag & Munk (2005) from the case $d = 1$ and $r = 2$. In particular, we weaken their assumptions by showing that the connectedness of the supports of $P, Q$ with respect to the Lebesgue measure is not necessary for the validity of the bootstrap. In Section 5.2, we then define an assumption-light hybrid confidence interval based on the bootstrap and on the intervals of Section 4.

5.1 Bootstrapping the Sliced Wasserstein Distance

Let $P, Q \in \mathcal{P}([\mathbb{R}]^d)$, and let $X_1, \ldots, X_n \sim P$, $Y_1, \ldots, Y_m \sim Q$ be i.i.d. samples. Furthermore, let $P_n$ and $Q_m$ denote the corresponding empirical measures, and let $P_n^*$ and $Q_m^*$ denote their bootstrap counterparts (that is, $P_n^*$ is the sampling distribution of a sample of size $n$ drawn from $P_n$). Lemma 8 in Appendix E establishes the Hadamard differentiability of the Sliced Wasserstein distance over pairs of distributions $(P, Q)$ admitting absolutely continuous projections. The consistency of the bootstrap in estimating the distribution of $SW_{r, \delta}(P_n, Q_m)$ then follows from the functional delta method (van der Vaart & Wellner 1996), as outlined in the following Theorem. In what follows $BL_1$ denotes the set of 1-Lipschitz functions $f : \mathbb{R} \to \mathbb{R}$, such that $\|f\|_\infty \leq 1$, and $\mathcal{K}(b) = \{P \in \mathcal{P}([\mathbb{R}]^d) : \|\|x\|^2 dP(x) \leq b\}$ for $b > 0$.

**Theorem 3.** Let $\delta \in (0, 1/2)$, and $P, Q \in \mathcal{K}(b)$. Assume $P \neq Q$, and that $P_\theta, Q_\theta$ are absolutely continuous for $\mu$-almost every $\theta \in \mathbb{S}^{d-1}$. Then, the bootstrap is consistent, namely,

$$
\sup_{h \in BL_1} \mathbb{E} \left[ h \left( \sqrt{\frac{nm}{n+m}} \left( SW_{r, \delta}(P_n^*, Q_m^*) - SW_{r, \delta}(P_n, Q_m) \right) \right) \right] \rightarrow 0,
$$

in outer probability, where the limit is taken as $n, m \to \infty$ such that $\frac{n}{n+m} \to a \in (0, 1)$.

Letting $F_{nm}^*$ denote the CDF of $SW_{r, \delta}(P_n^*, Q_m^*)$, and setting $q_{\alpha/2} = F_{nm}^*(\alpha/2)$, $q_{1-\alpha/2} = F_{nm}^*(1-\alpha/2)$, for some $\alpha \in (0, 1)$, it follows that an asymptotic $(1 - \alpha)$ confidence interval for $SW_{r, \delta}(P, Q)$ is given by

$$
C_{nm}^* = \left[ \left( SW_{r, \delta}(P_n, Q_m) - q_{\alpha/2} \sqrt{\frac{n+m}{nm}} \right)^{\frac{1}{\alpha}}, \left( SW_{r, \delta}(P_n, Q_m) + q_{\alpha/2} \sqrt{\frac{n+m}{nm}} \right)^{\frac{1}{\alpha}} \right]. \quad (35)
$$
In practice, the CDF $F_{nm}^*$ is estimated via Monte Carlo (Efron & Tibshirani 1994).

By Theorem 3, the primary assumptions for the validity of $C_{nm}^*$ are the absolute continuity of the projections of $P, Q$, and the condition $P \neq Q$. The former assumption can be anticipated from the lack of Hadamard differentiability of the Wasserstein distance over finite spaces (Sommerfeld & Munk 2018), where all probability measures are purely atomic. Regarding the latter assumption, failure of the bootstrap at the null is due to the Sliced Wasserstein distance being a functional with first-order degeneracy (Munk & Czado 1998), for which corrections such as those of Chen & Fang (2019), or the $m$-out-of-$n$ bootstrap (Sommerfeld & Munk 2018), yield a consistent procedure, but are practically less attractive as they introduce further tuning parameters. In what follows, we show that a simple test for the absolute continuity of $P$ and $Q$, together with a two-sample test, can be used to form a confidence interval with minimal assumptions which either equals $C_{nm}^*$ or a finite-sample confidence interval from the previous section.

5.2 A Hybrid Bootstrap Approach

Let $\alpha \in (0,1)$, and define

$$D_n = \min_{1 \leq i, j \leq n} \|X_i - X_j\|, \quad D'_m = \min_{1 \leq i, j \leq m} \|Y_i - Y_j\|,$$

The quantities $D_n$ and $D'_m$ denote the smallest sample spacing, and small values of these quantities indicate that $P$ and $Q$ may admit atoms. We let $C_{nm}^*$ denote the bootstrap confidence interval in equation (35) at level $1 - \alpha$, and let $C_{nm}^\dagger$ denote the assumption-light confidence interval for $SW_{r,\delta}(P,Q)$ in equation (31) at level $1 - \alpha$. We define the $(1-\alpha)$-hybrid confidence interval as:

$$C_{nm} = \begin{cases} C_{nm}^\dagger & \text{if } \{0 \in C_{nm}^\dagger\} \cup \{D_n = 0\} \cup \{D'_m = 0\} \\ C_{nm}^* & \text{otherwise} \end{cases}.$$  \hspace{1cm} (36)

Roughly, we use the bootstrap interval if we are reasonably certain that $P \neq Q$ and that the two distributions do not admit atoms, and fall back on the assumption-light finite-sample interval otherwise. The following result characterizes the coverage guarantees we provide for the hybrid interval.

**Theorem 4.** Let $\alpha \in (0,1), r \geq 1, b > 0$ and $\delta \in (0,1/2)$, and assume $A1(\delta; \alpha)$ holds. Assume further that for almost all $\theta \in S^{d-1}$, the measures $P_\theta$ and $Q_\theta$ admit no singular components. Then,

$$\inf_{P,Q \in K(b)n,m} \liminf_{n,m \to \infty} \mathbb{P}(SW_{r,\delta}(P,Q) \in C_{nm}) \geq 1 - \alpha,$$  \hspace{1cm} (37)

where the limit inferior is taken such that $\frac{n}{n+m} \to a \in (0,1)$.

**Remarks:**

- Theorem 4 establishes the asymptotic coverage of $C_{nm}$ under a mild moment assumption on $P,Q$, and the assumption that $P$ and $Q$ admit no singular components along almost all projections. It is worth noting that this last assumption is rather mild, and does not prevent $P$ and $Q$ themselves from admitting singular components.
• Equation (37) is strictly weaker than the uniform coverage guarantee

$$\liminf_{n,m \to \infty} \inf_{P,Q \in \mathcal{K}(b)} \mathbb{P}(\text{SW}_{r,\delta}(P,Q) \in C_{nm}) \geq 1 - \alpha.$$  \hspace{1cm} (38)

Indeed, absent other assumptions, it is not possible to construct a non-trivial test for the absolute continuity of the projections of $P$ and $Q$, which controls the Type I error uniformly over the class of distributions in $\mathcal{K}(b)$. Following our proof with minor modifications, a coverage statement of the form (38) can be shown to hold under a uniform upper bound on the densities of the absolutely continuous component of $P_\theta$ and $Q_\theta$, for $\theta \in S_d^{-1}$.

6 Simulation Study

We perform a simulation study to illustrate the coverage and length of the confidence intervals described in Sections 4 and 5. All simulations were performed in Python 3.5 on a typical Linux machine with eight cores. Implementations for all confidence intervals described in this paper, along with code for reproducing the following simulations, are publicly available\(^1\).

Comparison of Asymptotic and Finite Sample Confidence Intervals. We compare the following confidence intervals: (i) The finite sample interval in equation (31) (or (29) in the one-dimensional case), based on the DKW inequality from Example 1, (ii) The standard bootstrap confidence interval in equation (35), and (iii) The hybrid interval in equation (36). We also implemented the finite-sample interval (31) with respect to the relative VC inequality (Example 2), however we rarely noticed an improvement over the DKW finite sample interval in practice. This is likely due to the sub-optimal constants in the relative VC inequality (27), unlike those in the DKW inequality of Massart (1990), and consequently we do not consider this method in the present simulation study.

| Model | $P$ | $Q$ |
|-------|-----|-----|
| 1     | $\frac{1}{2}N\left((-1,1)^\top J_2\right) + \frac{1}{2}N\left((1,1)^\top J_2\right)$ | $N(0,I_2)$ |
| 2     | $\frac{1+\sqrt{n}}{2}\delta_2 + \delta_1$ | $\frac{1}{2}\delta_2 + \delta_5$ |
| 3     | $\mathbb{T}(1,1)$ | $\mathbb{T}(\frac{1}{2},5)$ |
| 4     | $.95N(0,1) + .05N(0,1)$ | $N(0,1)$ |
| 5     | $.55N\left((-5,-5)^\top J_2\right) + .45N\left((5,5)^\top J_2\right)$ | $\frac{1}{2}N\left((-5,-5)^\top J_2\right) + \frac{1}{2}N\left((5,5)^\top J_2\right)$ |

Table 1: Parameter settings for Models 1-5. For any $R>r>0$, $\mathbb{T}(r,R)$ denotes the uniform distribution over the torus $\{(r+\cos\theta)\cos\psi,(r+\cos\theta)\sin\psi,rsin\theta\)^\top \leq \theta,\psi \leq 2\pi \subseteq \mathbb{R}^3$. Sampling from $\mathbb{T}(r,R)$ is performed using the rejection sampling algorithm of Diaconis et al. (2013).

We generate 100 samples of size $n=m=600, 900, 1200$ and 1500, from each of the pairs of distributions $(P,Q)$ described in Table 1. We choose the level $\alpha=.05$, the trimming constant $\delta=.1$ and the Monte Carlo sample size $N=500$, for which Assumption A1$(\delta;\alpha/N)$ is met for the sample sizes under consideration. The number of bootstrap replications is set to $B=1,000$. The order of the Wasserstein distance $r$ is

\(^1\)https://github.com/tmanole/SW-inference.
chosen to be 2. The empirical coverage and average lengths of the three confidence intervals are reported in Figure 1 for Models 1-3, and in Figure 2 for Models 4-5.

Model 1 satisfies the regularity conditions required for the validity of the bootstrap, and we indeed observe its valid coverage for all sample sizes considered. When $n = 600$, the finite sample confidence interval does not distinguish $SW_{r,\delta}(P,Q)$ from zero on every replication, hence the hybrid confidence interval exhibits length and coverage between those of the finite sample and bootstrap intervals. For larger sample sizes, the hybrid interval coincides with the bootstrap interval. Model 2 is not absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}$, and the bootstrap is seen to markedly undercover the true Wasserstein distance. Model 3 is also not absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^3$, since the supports of $P$ and $Q$ are two-dimensional manifolds. Nevertheless, the linear projections of $P$ and $Q$ are absolutely continuous and well-separated, ensuring the validity of the bootstrap. Models 4 and 5 consist of pairs of measures admitting Sliced Wasserstein distance near zero, causing the bootstrap to undercover. We also report the average runtime under these Models in Figures 2(c) and 2(f), showing a clear computational advantage of our finite sample intervals over the bootstrap.
Adaptivity and Asymptotic Confidence Interval Length. We now illustrate the behaviour predicted by Theorem 2, and Corollaries thereafter, regarding the length of our finite-sample intervals. Consider the following two pairs of distributions,

Model 6(i). \[ P_1 = \frac{1}{2} \delta_{-5} + \frac{1}{2} \delta_5, \quad Q_{1, \Delta} = \left( \frac{1}{2} + \Delta \right) \delta_{-5} + \left( \frac{1}{2} - \Delta \right) \delta_5, \]

Model 6(ii). \[ P_2 = U(-5,5), \quad Q_{2, \Delta} = \left( \frac{1}{2} + \Delta \right) U(-5,0) + \left( \frac{1}{2} - \Delta \right) U(0,5), \]

for any \( \Delta \in [0,1/2] \). Notice that \( J_{r, \delta}(P_1) = J_{r, \delta}(Q_{1, \Delta}) = \infty \), while \( J_{r, \delta}(P_2), J_{r, \delta}(Q_{2, \Delta}) < \infty \). We report the average length of the finite sample confidence interval (29), for each of Models 6.(i) and 6.(ii), based on 100 samples of sizes \( n = m \in [250, 50,000] \). In Figures 3(a) and 3(b), we do so for \( \Delta = 0 \) with respect to Wasserstein distances of varying orders \( r \in [1,16] \), while in Figures 3(c) and 3(d), we do so for a range of values \( \Delta \in [0,0.4] \) with respect to the Wasserstein distance of order \( r = 2 \). In the former case, the average confidence interval length for the pair \( (P_1,Q_{1,0}) \) decays at an increasingly slow rate as the order \( r \) increases, while that of the pair \( (P_2,Q_{2,0}) \) remains nearly unchanged. This behaviour was predicted by Corollary
1, which indicates that the length of the finite sample interval scales at the \( n^{-1/2r} \) rate in the worst case, but does so at the faster rate \( n^{-1/2} \) for distributions admitting finite SJ\(_{r,\delta} \) values. When \( r=2 \) is fixed, Figures 3(c) and 3(d) similarly exhibit increasing interval lengths for the pair \((P_1,Q_{1,\Delta})\) as \( \Delta \) decreases to zero, yet nearly constant lengths for the pair \((P_2,Q_{2,\Delta})\). This behaviour is in line with Corollary 2.

7 Application to Likelihood-Free Inference

In a wide range of statistical applications, the likelihood function for a parametric model of interest may be intractable, though samples from the model can be easily generated. Examples include proton-proton collisions in particle physics (Stoye et al. 2018, Brehmer et al. 2020a,b), predator-prey dynamics in ecology (Lotka 1920a,b), inference for cosmological parameters in astronomy (Dalmasso et al. 2020a,b), and network dynamics in queuing theory (Ebert et al. 2019). In such applications, the practitioner typically has access to a parametrized stochastic simulator for the data generating process, which produces samples from a distribution \( P_{\theta} \in \mathcal{P}(\mathbb{R}^d) \) with unknown closed form, and which depends on some physically meaningful parameters \( \theta \in \Theta \subseteq \mathbb{R}^D \). The goal of likelihood-free inference is to characterize the values of \( \theta \) for which an observed sample \( X_1,\ldots,X_n \) is likely to have been generated by the simulator.

Approximate Bayesian computation (ABC; Sisson et al. (2018)) is arguably the most popular family of methodologies for likelihood-free inference. ABC methods repeatedly simulate parameter values in \( \Theta \), and accept those for which the simulator \( P_{\theta} \) produces a similar synthetic sample to the observed sample. The similarity of two samples is typically measured on the basis of summary statistics of the datasets. These summary statistics are often application-specific, and can be difficult to specify. Furthermore, due to the intractability of the likelihood, summary statistics can rarely be chosen as sufficient statistics for \( \theta \), making information loss inevitable. These considerations have motivated the development of methods which replace tailored summary statistics by distances between empirical measures of the synthetic and observed samples (Park et al. 2016, Gutmann et al. 2018, Jiang 2018). In particular, Bernton et al. (2019b) and Nadjahi et al. (2020) suggest the use of the Wasserstein and Sliced Wasserstein distances for this purpose.

In what follows, we propose a simple alternative to such ABC methods, which provides frequentist guarantees for likelihood-free inference. Using the method developed in Section 4, we build confidence
sets for the simulator’s parameters on the basis of minimizing the (Sliced) Wasserstein distance between
the empirical measures of an observed sample and synthetic samples from the simulator. We focus on
the case where $X_1,\ldots,X_n$ is an i.i.d. sample from a distribution $P \in \mathcal{P}(\mathbb{R}^d)$. Fix
\[
\theta_0 \in \arg\min_{\theta \in \Theta} SW_{r,\delta}(P,P_{\theta}),
\]
and let $\epsilon_0 = SW_{r,\delta}(P_{\theta_0},P)$. Here $\theta_0$ denotes an $SW_{r,\delta}$-projection of the distribution $P$ onto the family
$\{P_\theta\}_{\theta \in \Theta}$. If the simulator is correctly-specified, we have $P = P_{\theta_0}$ and $\epsilon_0 = 0$, whereas if the simulator is
misspecified, the set $\{\theta \in \Theta : SW_{r,\delta}(P,P_\theta) \leq \epsilon\}$, is empty for sufficiently small values of $\epsilon \geq 0$.

We propose to construct confidence sets for $\theta_0$. For any $\theta \in \Theta$, and for any synthetic sample $Y_{1}^\theta,...,Y_{n}^\theta \sim P_{\theta}$, let $[\ell_{nm}(\theta),u_{nm}(\theta)]$ be a $(1-\alpha)$-confidence interval for $SW_{r,\delta}(P,P_{\theta})$, obtained via equation (31) on the basis of $Y_{1}^\theta,...,Y_{n}^\theta$ and the observed sample $X_{1},...,X_{n}$. A confidence set for $\theta_0$ is then easily given by the following Proposition.

**Proposition 6.** Let $\alpha \in (0,1)$, $r \geq 1$, and assume the same conditions as Proposition 5. Given any fixed
real number $\epsilon \geq \epsilon_0$, define
\[
C_{nm} = \{\theta \in \Theta : \ell_{nm}(\theta) \leq \epsilon\}. \tag{41}
\]
Then,
\[
\liminf_{N \to \infty} \inf_{P \in \mathcal{K}_{2r}(b)} \mathbb{P}\left(\theta_0 \in C_{nm}\right) \geq 1 - \alpha.
\]

Proposition 6 provides a $(1-\alpha)$-confidence set for the projection $\theta_0$. In the well-specified setting $\epsilon_0 = 0$, $C_{nm}$ is simply a confidence set for the parameter $\theta_0$, which corresponds to the ground truth $P$. We emphasize that no assumptions were made in the statement of Proposition 6 beyond the mild moment assumption $P,P_\theta \in \mathcal{K}_{2r}(b)$ (which can be removed when $d = 1$). The intractibility of the likelihood function makes such assumption-lean inference particularly attractive.

In practice, computation of the lower confidence bounds $\ell_{nm}$ in equation (41) may be carried out over a finite grid $\{\theta_1,\ldots,\theta_M\} \subseteq \Theta$ of candidate parameter values. While such a search may be computationally expensive, particularly for parameter spaces of high dimension $D$, it is akin to repeated sampling of parameters in ABC, or similar operations in other likelihood-free methods. Nevertheless, efficient computation of the individual intervals $[\ell_{nm}(\theta),u_{nm}(\theta)]$ can dramatically reduce the computational burden of $C_{nm}$. The simulation study in Section 6 suggests that the runtime of our finite-sample confidence intervals is considerably lower than that of bootstrap-based methods (cf. Figures 2(c) and 2(f)).

**Example: The Toggle Switch Model.** We illustrate our methodology in a systems biology model used by Bonassi et al. (2011), Bonassi & West (2015). This model was analyzed by Bernton et al. (2019b) using an ABC method based on the Wasserstein distance, and serves as a realistic example of likelihood-free inference with independent data. The toggle switch model describes the expression level of two genes across $n$ cells over $T \in \mathbb{N}$ time points. Specifically, we let $(U_{i,t},V_{i,t})$ denote their expression level in cell $i \in \{1,\ldots,n\}$, and at time $t \in \{1,\ldots,T\}$. Given a starting value $(U_{i,0},V_{i,0})$ for every $i = 1,\ldots,n$, the model is given for $t = 1,\ldots,T$ by
\[
\begin{align*}
U_{i,t+1} &= U_{i,t} + \frac{\alpha_1}{1+V_{i,t}^\alpha} - (1+0.03U_{i,t}) + \frac{1}{2} \xi_{i,t} \\
V_{i,t+1} &= V_{i,t} + \frac{\alpha_2}{1+U_{i,t}^\alpha} - (1+0.03V_{i,t}) + \frac{1}{2} \xi_{i,t}
\end{align*}
\tag{42}
\]
where $\alpha_1, \alpha_2 \in \mathbb{R}$ and $\beta_1, \beta_2 \geq 0$ are parameters, and $\xi_{i,t}, \zeta_{i,t}$ are independent standard Gaussian random variables. Following Bernton et al. (2019b), $\xi_{i,t}$ and $\zeta_{i,t}$ are truncated such that $U_{i,t}, V_{i,t}$ remain nonnegative for all $i,t$. In applications, the full evolution (42) is not observed, except for the following noisy measurement at time $T$,

$$X_i = U_{i,T} + \epsilon_i, \quad i = 1, \ldots, n,$$

where $\epsilon_i \sim N(\mu, \mu \sigma / U_{i,T}^2)$, and $\mu \in \mathbb{R}$ and $\sigma, \gamma \geq 0$ are parameters. $X_1, \ldots, X_n$ thus forms an i.i.d. sample from a distribution $P_\theta$ on $\mathbb{R}$ with respect to the parameter $\theta = (\alpha_1, \alpha_2, \beta_1, \beta_2, \mu, \sigma, \gamma) \in \mathbb{R}^7$. A closed form for $P_\theta$ is unclear, but the evolution (42) makes simulation from $P_\theta$ simple, making this model a good candidate for likelihood-free inference.

We illustrate our methodology on simulated observations from this model in both a well-specified and a misspecified case. We treat the exponent parameters $\beta_1, \beta_2, \gamma$ as known, but possibly misspecified, and perform inference on $(\alpha_1, \alpha_2, \mu, \sigma)$. In what follows, we set $U_{i,0} = V_{i,0} = 10$, $i = 1, \ldots, n$, and we generate $n = 2,000$ observations from $P_{\theta_0}$ with $\theta_0 = (22, 12, 4, 4.5, 325, 25, 15)$, matching the parameter setting of Bernton et al. (2019b).

- **Well-specified Setting.** Treat $\beta_1 = 4, \beta_2 = 4.5, \gamma = 0.15$ as known and correctly specified. We compute the confidence set (41) with $r = 1$ and $\epsilon = 0$, by repeatedly simulating $m$ observations from a grid of candidate values of $(\alpha_1, \alpha_2, \mu, \sigma) \in \mathbb{R}^4$, for $m \in \{5 \cdot 10^3, 10^4, 2 \cdot 10^4\}$. The resulting two-dimensional confidence sets for the parameters $(\alpha_1, \alpha_2)$, which are of primary interest, are reported in Figure 4. These confidence sets can be seen to cover the true parameter value, and naturally have decreasing area as $m$ increases.

- **Misspecified Setting.** Using the same observed sample as before, we now misspecify the simulator with the values $\beta_1 = 2$ and $\beta_2 = 2$. The resulting confidence set $C_{nm}$ is shown in Figure 5 for several choices of $\epsilon$, and can be seen to cover the projection parameter $\theta_0$. The latter was approximated by $\tilde{\theta}_0 = \arg\min_\theta W_{1,\delta}(P_M, P_\theta)$ over a grid of parameters $\theta$, where $P_M$ denotes a simulated empirical measure based on $M = 10,000$ observations.

### 8 Conclusion and Discussion

Our aim in this paper has been to develop assumption-light finite sample confidence intervals for the Sliced Wasserstein distance. After deriving the minimax rates for estimating the Sliced Wasserstein distance, which are of independent interest, we bounded the length of our confidence intervals, showing that they achieve minimax optimal length up to polylogarithmic factors. Their length is also shown to be adaptive to whether or not the underlying distributions are near the classical null, as well as to their regularity, as measured by the magnitude of the functional $\text{SI}_{\lambda, \delta}$. These findings contrast asymptotic methods such as the bootstrap, whose validity we show is subject to certain prohibitive assumptions on the underlying distributions, and whose asymptotic length does not enjoy the same adaptivity as that of our finite sample intervals.

Our work leaves open the problem of statistical inference for Wasserstein distances in dimension greater than one, for which new techniques would have to be developed. Indeed, our work has hinged upon the representation of the one-dimensional Wasserstein distance as the $L^p$ distance between quantile functions, which is unavailable in general dimension. For the same reason, our work does not shed light on statistical
inference for other modifications of the Wasserstein distance based on projections of distributions to low-dimensions greater than one, such as those summarized in Section 2.3. We have shown that the Sliced Wasserstein distance can be estimated at dimension-independent rates, and it is of interest to understand how this finding changes for other low-dimensional modifications of the Wasserstein distance.

Acknowledgments

Tudor Manole would like to thank Niccolò Dalmasso for conversations which inspired the likelihood-free inference application in Section 7. This work was supported in part by the NSF grants DMS-1713003 and CIF-1763734, and by the Natural Sciences and Engineering Research Council of Canada.
References

Álvarez-Esteban, P. C., Del Barrio, E., Cuesta-Albertos, J. A. & Matrán, C. (2008), ‘Trimmed comparison of distributions’, *Journal of the American Statistical Association* **103**, 697–704. 2, 6

Álvarez-Esteban, P. C., del Barrio, E., Cuesta-Albertos, J. A. & Matrán, C. (2012), ‘Similarity of samples and trimming’, *Bernoulli* **18**, 606–634. 2

Arjovsky, M., Chintala, S. & Bottou, L. (2017), Wasserstein generative adversarial networks, *in* ‘Proceedings of the 34th International Conference on Machine Learning’, Vol. 70, pp. 214–223. 2

Bernton, E., Jacob, P. E., Gerber, M. & Robert, C. P. (2019a), ‘Approximate Bayesian computation with the Wasserstein distance’, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **81**, 235–269. 2

Bernton, E., Jacob, P. E., Gerber, M. & Robert, C. P. (2019b), ‘On parameter estimation with the Wasserstein distance’, *Information and Inference: A Journal of the IMA*. 2, 7, 24, 25, 26

Bobkov, S. & Ledoux, M. (2019), ‘One-dimensional empirical measures, order statistics, and Kantorovich transport distances’, *Memoirs of the American Mathematical Society* **261**. 3, 5, 7, 9, 10, 33

Boissard, E. & Le Gouic, T. (2014), On the mean speed of convergence of empirical and occupation measures in Wasserstein distance, *in* ‘Annales de l’Institut Henri Poincare Probabilités et Statistiques’, Vol. 50, pp. 539–563. 7

Bolley, F., Guillin, A. & Villani, C. (2006), ‘Quantitative Concentration Inequalities for Empirical Measures on Non-compact Spaces’, *Probability Theory and Related Fields* **137**, 541–593. 7

Bonassi, F. V. & West, M. (2015), ‘Sequential Monte Carlo with Adaptive Weights for Approximate Bayesian Computation’, *Bayesian Analysis* **10**, 171–187. 25

Bonassi, F. V., You, L. & West, M. (2011), ‘Bayesian learning from marginal data in bionetwork models’, *Statistical Applications in Genetics and Molecular Biology* **10**. 25

Bonneel, N., Rabin, J., Peyré, G. & Pfister, H. (2015), ‘Sliced and Radon Wasserstein Barycenters of Measures’, *Journal of Mathematical Imaging and Vision* **51**, 22–45. 2, 6

Bonnote, N. (2013), Unidimensional and Evolution Methods for Optimal Transportation, PhD thesis, Paris 11. 2, 6

Bouchitté, G., Jimenez, C. & Rajesh, M. (2007), ‘A new $L^\infty$ estimate in optimal mass transport’, *Proceedings of the American Mathematical Society* **135**, 3525–3535. 5

Brehmer, J., Kling, F., Espejo, I. & Cranmer, K. (2020a), ‘MadMiner: Machine learning-based inference for particle physics’, *Computing and Software for Big Science* **4**, 1–25. 24

Brehmer, J., Louppe, G., Pavez, J. & Cranmer, K. (2020b), ‘Mining gold from implicit models to improve likelihood-free inference’, *Proceedings of the National Academy of Sciences* **117**, 5242–5249. 24

Chen, Q. & Fang, Z. (2019), ‘Inference on Functionals under First Order Degeneracy’, *arXiv preprint arXiv:1901.04861*. 20
Courty, N., Flamary, R., Tuia, D. & Rakotomamonjy, A. (2016), ‘Optimal transport for domain adaptation’, *IEEE transactions on pattern analysis and machine intelligence* 39, 1853–1865. 2

Csorgo, M. & Révész, P. (1978), ‘Strong Approximations of the Quantile Process’, *The Annals of Statistics* 6, 882–894. 18

Dalmasso, N., Izbicki, R. & Lee, A. B. (2020a), ‘Confidence Sets and Hypothesis Testing in a Likelihood-Free Inference Setting’, *arXiv preprint arXiv:2002.10399*. 24

Dalmasso, N., Lee, A., Izbicki, R., Pospisil, T., Kim, I. & Lin, C.-A. (2020b), Validation of approximate likelihood and emulator models for computationally intensive simulations, in ‘Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics’, pp. 3349–3361. 24

del Barrio, E., Cuesta-Albertos, J. A., Matrán, C. & Rodríguez-Rodríguez, J. M. (1999), ‘Tests of goodness of fit based on the $L_2$-Wasserstein distance’, *The Annals of Statistics* 27, 1230–1239. 2

del Barrio, E., Giné, E. & Utzet, F. (2005), ‘Asymptotics for $L^2$ functionals of the empirical quantile process, with applications to tests of fit based on weighted Wasserstein distances’, *Bernoulli* 11, 131–189. 2

del Barrio, E., Gordaliza, P., Lescornel, H. & Loubes, J.-M. (2019), ‘Central limit theorem and bootstrap procedure for Wasserstein’s variations with an application to structural relationships between distributions’, *Journal of Multivariate Analysis* 169, 341–362. 7

del Barrio, E., Gordaliza, P. & Loubes, J.-M. (2018), ‘A Central Limit Theorem for $L^p$ transportation cost with applications to Fairness Assessment in Machine Learning’, *arXiv:1807.06796*. 18

Deshpande, I., Hu, Y.-T., Sun, R., Pyrros, A., Siddiqui, N., Koyejo, S., Zhao, Z., Forsyth, D. & Schwing, A. G. (2019), Max-Sliced Wasserstein distance and its use for GANs, in ‘Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition’, pp. 10648–10656. 7

Diaconis, P., Holmes, S. & Shahshahani, M. (2013), Sampling from a manifold, in ‘Advances in Modern Statistical Theory and Applications: A Festschrift in Honor of Morris L. Eaton’, Institute of Mathematical Statistics, pp. 102–125. 21

Dudley, R. M. (1969), ‘The speed of mean Glivenko-Cantelli convergence’, *The Annals of Mathematical Statistics* 40, 40–50. 7

Dvoretzky, A., Kiefer, J. & Wolfowitz, J. (1956), ‘Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator’, *The Annals of Mathematical Statistics* 27, 642–669. 13

Ebert, A., Dutta, R., Mengersen, K., Mira, A., Ruggeri, F. & Wu, P. (2019), ‘Likelihood-free parameter estimation for dynamic queueing networks: Case study of passenger flow in an international airport terminal’, *arXiv preprint arXiv:1804.02526*. 24

Efron, B. & Tibshirani, R. J. (1994), *An Introduction to the Bootstrap*, CRC press. 20

Fournier, N. & Guillin, A. (2015), ‘On the rate of convergence in Wasserstein distance of the empirical measure’, *Probability Theory and Related Fields* 162, 707–738. 7, 10
Freitag, G., Czado, C. & Munk, A. (2007), ‘A nonparametric test for similarity of marginals—With applications to the assessment of population bioequivalence’, *Journal of Statistical Planning and Inference* **137**, 697–711. 2, 3, 7, 19

Freitag, G. & Munk, A. (2005), ‘On Hadamard differentiability in k-sample semiparametric models—with applications to the assessment of structural relationships’, *Journal of Multivariate Analysis* **94**, 123–158. 2, 3, 7, 19

Freitag, G., Munk, A. & Vogt, M. (2003), Assessing structural relationships between distributions-a quantile process approach based on Mallows distance, in ‘Recent Advances and Trends in Nonparametric Statistics’, Elsevier, pp. 123–137. 2, 3, 7, 19

Gutmann, M. U., Dutta, R., Kaski, S. & Corander, J. (2018), ‘Likelihood-free inference via classification’, *Statistics and Computing* **28**, 411–425. 24

Ho, N., Nguyen, X., Yurochkin, M., Bui, H. H., Huynh, V. & Phung, D. (2017), Multilevel Clustering via Wasserstein Means, in ‘Proceedings of the 34th International Conference on Machine Learning’, pp. 1501–1509. 2

Ho, N., Yang, C.-Y. & Jordan, M. I. (2019), ‘Convergence Rates for Gaussian Mixtures of Experts’, *arXiv preprint arXiv:1907.04377*. 2

Imaizumi, M., Ota, H. & Hamaguchi, T. (2019), ‘Hypothesis Test and Confidence Analysis with Wasserstein Distance on General Dimension’, *arXiv preprint arXiv:1910.07773*. 7

Jiang, B. (2018), Approximate Bayesian Computation with Kullback-Leibler Divergence as Data Discrepancy, in ‘Proceedings of the 21st International Conference on Artificial Intelligence and Statistics’, pp. 1711–1721. 24

Kim, I., Balakrishnan, S. & Wasserman, L. (2020), ‘Robust multivariate nonparametric tests via projection-averaging’, *arXiv preprint arXiv:1803.00715, to appear in the Annals of Statistics*. 7

Klatt, M., Munk, A. & Zemel, Y. (2020), ‘Limit Laws for Empirical Optimal Solutions in Stochastic Linear Programs’, *arXiv preprint arXiv:2007.13473*. 7

Klatt, M., Tameling, C. & Munk, A. (2020), ‘Empirical regularized optimal transport: Statistical theory and applications’, *SIAM Journal on Mathematics of Data Science* **2**, 419–443. 7

Kolouri, S., Nadjahi, K., Simsekli, U., Badeau, R. & Rohde, G. (2019), Generalized sliced wasserstein distances, in ‘Advances in Neural Information Processing Systems 32’, pp. 261–272. 7

Kolouri, S., Park, S. R., Thorpe, M., Slepecev, D. & Rohde, G. K. (2017), ‘Optimal mass transport: Signal processing and machine-learning applications’, *IEEE signal processing magazine* **34**, 43–59. 1

Le, T., Yamada, M., Fukumizu, K. & Cuturi, M. (2019), Tree-sliced variants of Wasserstein distances, in ‘Advances in Neural Information Processing Systems 32’, pp. 12283–12294. 7

Lei, J. (2020), ‘Convergence and concentration of empirical measures under Wasserstein distance in unbounded functional spaces’, *Bernoulli* **26**, 767–798. 7, 10
Liang, T. (2019), ‘On the Minimax Optimality of Estimating the Wasserstein Metric’, arXiv preprint arXiv:1908.10324 7, 8

Lotka, A. J. (1920a), ‘Undamped oscillations derived from the law of mass action.’, Journal of the American Chemical Society 42, 1595–1599. 24

Lotka, A. J. (1920b), ‘Analytical note on certain rhythmic relations in organic systems’, Proceedings of the National Academy of Sciences 6, 410–415. 24

Massart, P. (1990), ‘The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality’, The Annals of Probability pp. 1269–1283. 13, 21

Munk, A. & Czado, C. (1998), ‘Nonparametric Validation of Similar Distributions and Assessment of Goodness of Fit’, Journal of the Royal Statistical Society: Series B (Statistical Methodology) 60, 223–241. 2, 3, 5, 7, 18, 20

Nadjahi, K., De Bortoli, V., Durmus, A., Badeau, R. & Şimşekli, U. (2020), Approximate Bayesian computation with the sliced-Wasserstein distance, in ‘IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)’, pp. 5470–5474. 7, 24

Nguyen, K., Ho, N., Pham, T. & Bui, H. (2020), ‘Distributional Sliced-Wasserstein and Applications to Generative Modeling’, arXiv preprint arXiv:2002.07367 7

Nguyen, X. (2013), ‘Convergence of latent mixing measures in finite and infinite mixture models’, The Annals of Statistics 41, 370–400. 2

Niles-Weed, J. & Rigollet, P. (2019), ‘Estimation of Wasserstein distances in the Spiked Transport Model’, arXiv preprint arXiv:1909.07513 7, 8

Panaretos, V. M. & Zemel, Y. (2019a), An Invitation to Statistics in Wasserstein Space, Springer Nature. 1

Panaretos, V. M. & Zemel, Y. (2019b), ‘Statistical Aspects of Wasserstein Distances’, Annual Review of Statistics and Its Application 6, 405–431. 2

Park, M., Jitkrittum, W. & Sejdinovic, D. (2016), K2-ABC: Approximate Bayesian Computation with Kernel Embeddings, in ‘Proceedings of the 19th International Conference on Artificial Intelligence and Statistics’, pp. 398–407. 24

Paty, F.-P. & Cuturi, M. (2019), Subspace Robust Wasserstein Distances, in ‘Proceedings of the 36th International Conference on Machine Learning’, pp. 5072–5081. 7

Peyré, G. & Cuturi, M. (2019), ‘Computational optimal transport’, Foundations and Trends® in Machine Learning 11, 355–607. 2

Rabin, J., Peyré, G., Delon, J. & Bernot, M. (2011), Wasserstein barycenter and its application to texture mixing, in ‘International Conference on Scale Space and Variational Methods in Computer Vision’, Springer, pp. 435–446. 2, 6

Ramdas, A., Trillos, N. & Cuturi, M. (2017), ‘On wasserstein two-sample testing and related families of nonparametric tests’, Entropy 19, 47. 2
Rippl, T., Munk, A. & Sturm, A. (2016), ‘Limit laws of the empirical Wasserstein distance: Gaussian distributions’, *Journal of Multivariate Analysis* **151**, 90–109. 7

Shao, J. & Tu, D. (2012), *The Jackknife and Bootstrap*, Springer Science & Business Media. 2

Shorack, G. R. & Wellner, J. A. (2009), *Empirical Processes with Applications to Statistics*, SIAM. 2, 13

Singh, S. & Póczos, B. (2019), ‘Minimax Distribution Estimation in Wasserstein Distance’, *arXiv preprint arXiv:1802.08855*. 7

Sisson, S. A., Fan, Y. & Beaumont, M. (2018), *Handbook of Approximate Bayesian Computation*, Chapman and Hall/CRC. 3, 24

Sommerfeld, M. & Munk, A. (2018), ‘Inference for empirical Wasserstein distances on finite spaces’, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **80**, 219–238. 7, 11, 20

Stoye, M., Brehmer, J., Louppe, G., Pavez, J. & Cranmer, K. (2018), ‘Likelihood-free inference with an improved cross-entropy estimator’, *arXiv preprint arXiv:1808.00973*. 24

Tameling, C., Sommerfeld, M. & Munk, A. (2019), ‘Empirical optimal transport on countable metric spaces: Distributional limits and statistical applications’, *The Annals of Applied Probability* **29**, 2744–2781. 7

Tsybakov, A. B. (2008), *Introduction to Nonparametric Estimation*, Springer Science & Business Media. 12, 36

van der Vaart, A. W. & Wellner, J. A. (1996), *Weak Convergence and Empirical Processes*, Springer. 19, 51, 52, 54

Vapnik, V. (2013), *The Nature of Statistical Learning Theory*, Springer Science & Business Media. 13, 14, 56

Verdinelli, I. & Wasserman, L. (2019), ‘Hybrid Wasserstein distance and fast distribution clustering’, *Electronic Journal of Statistics* **13**, 5088–5119. 2

Villani, C. (2003), *Topics in Optimal Transportation*, American Mathematical Soc. 1, 5, 41

Villani, C. (2008), *Optimal Transport: Old and New*, Springer Science & Business Media. 2, 4

Weed, J. & Bach, F. (2019), ‘Sharp asymptotic and finite-sample rates of convergence of empirical measures in Wasserstein distance’, *Bernoulli* **25**, 2620–2648. 7, 10
Appendix

This Appendix is organized as follows. Appendix A collects several preliminary results which will be frequently used in the sequel. Appendices B, C, D, E, and F, respectively contain the proofs of our main results: Proposition 3, Theorem 1, Theorem 2, Theorem 3, and Theorem 4. All remaining results are proven in Appendix G. Further discussion of Theorem 2 appears in Appendix H.

A Preliminary Results

In this section, we collect several preliminary results which will frequently be used in the sequel. We begin with the following straightforward Lemma, which follows from Proposition A.17 of Bobkov & Ledoux (2019).

Lemma 2. Let $P \in \mathcal{P}(\mathbb{R}^d)$, $r \geq 1$, and $\delta \in (0,1/2)$. Let $F_{\theta}^{-1}$ denote the quantile function of $P_{\theta} = \pi_{\theta \#} P$ for all $\theta \in S^{d-1}$. If $S_{r,\delta}(P) < \infty$, then $F_{\theta}^{-1}$ is absolutely continuous for $\mu$-almost all $\theta \in S^{d-1}$.

Furthermore, we describe the following characterization of distributions falling in the collection $K_r(b)$.

Lemma 3. Let $\delta \in (0,1/2)$, $r \geq 1$ and $b > 0$, be fixed real numbers. Then, there exists a real number $A_r > 0$, depending only on $b,r,\delta$, such that for all distributions $P \in K_r(b)$, we have

$$\int_{S^{d-1}} |F_{\theta}^{-1}(a)|^r d\mu(\theta) \leq A_r, \quad a \in \{\delta,1-\delta\}.$$  

Furthermore, there exists a real number $A'_r > 0$, depending only on $b,r,\delta$, such that for all distributions $P \in K_r(b)$, we have

$$\sup_{\theta \in S^{d-1}} |F_{\theta}^{-1}(a)| \leq A'_r, \quad a \in \{\delta,1-\delta\}.$$  

Proof of Lemma 3. The proof is a straightforward consequence of Markov’s inequality. We have

$$1-\delta \leq \mathbb{P}\left( X^\top \theta \geq F_{\theta}^{-1}(\delta) \right) \leq \frac{\mathbb{E}_X[(X^\top \theta)^2]}{[F_{\theta}^{-1}(\delta)]^2}.$$  

Thus, since $P \in K_r(b)$,

$$\int_{S^{d-1}} |F_{\theta}^{-1}(\delta)|^r d\mu(\theta) \leq \int_{S^{d-1}} \left( \frac{\mathbb{E}_X[(X^\top \theta)^2]}{1-\delta} \right)^{\frac{r}{2}} d\mu(\theta) \leq \frac{b}{(1-\delta)^2} =: A_r.$$  

The claim follows by a similar argument for $a = 1-\delta$. For the second part, we similarly have

$$\sup_{\theta \in S^{d-1}} |F_{\theta}^{-1}(\delta)| \leq \sqrt{\frac{1}{1-\delta} \mathbb{E}_X \left[ \sup_{\theta \in S^{d-1}} (X^\top \theta)^2 \right]} = \sqrt{\frac{1}{1-\delta} \mathbb{E}_X \left[ \|X\|^2 \right]} \leq \sqrt{\frac{b}{1-\delta}} =: A'_r.$$  

The claim follows. \qed
**B Proof of Proposition 3**

Part (i) of the claim is an immediate consequence of Proposition 1 and equation (16). Furthermore, part (ii) is immediate when \( r = 1 \), thus we only prove part (ii) for \( r > 1 \).

Let \( D_{nm} = SW^r_{\epsilon,\delta}(P_n, Q_m) - SW^r_{\epsilon,\delta}(P, Q) \). Then,

\[
\mathbb{E}\left| SW^r_{\epsilon,\delta}(P_n, Q_m) - SW^r_{\epsilon,\delta}(P, Q) \right| = SW^r_{\epsilon,\delta}(P, Q) \mathbb{E} \left( \frac{D_{nm}}{SW^r_{\epsilon,\delta}(P, Q)} + 1 \right)^{\frac{1}{r}} - 1 \\
\leq \frac{1}{r} \cdot SW^r_{\epsilon,\delta}(P, Q) \left( \mathbb{E}|D_{nm}| \right) \left( 1 + o(1) \right) \\
\leq \mathbb{E}|D_{nm}|,
\]

by Taylor expansion of the map \( x \mapsto (x+1)^{1/r} - 1 \) about \( x = 0 \) for \( r > 1 \), where we have used the assumption that \( SW_{\epsilon,\delta}(P, Q) \) is bounded away from zero. It will therefore suffice to prove \( \mathbb{E}|D_{nm}| \lesssim (n \wedge m)^{-1/2} \), where the symbol “\( \lesssim \)” hides constants possibly depending on \( b, r, \delta \).

Recall that \( F_{\theta}^{-1} \) (resp. \( G_{\theta}^{-1}, F_{\theta,n}^{-1}, G_{\theta,n}^{-1} \)) denotes the quantile function of the distribution \( P_{\theta} = \pi_{\theta \#} P \) (resp. \( Q_{\theta} = \pi_{\theta \#} Q, P_{\theta,n} = \pi_{\theta \#} P_n, Q_{\theta,m} = \pi_{\theta \#} Q_m \)), for any \( \theta \in \mathbb{S}^{d-1} \). We have,

\[
SW^r_{\epsilon}(P_n, Q_m) \\
= \int_{\mathbb{S}^{d-1}} \int_{\delta}^{1-\delta} |F_{\theta,n}^{-1}(u) - G_{\theta,m}^{-1}(u)|^r du \\
= \int_{\mathbb{S}^{d-1}} \int_0^1 \left\{ |F_{\theta}^{-1}(u) - G_{\theta}^{-1}(u)|^r \\
+ \text{sgn}(F_{\theta,n}^{-1}(u) - G_{\theta,m}^{-1}(u)) |\tilde{F}_{\theta,n}^{-1}(u) - \tilde{G}_{\theta,m}^{-1}(u)| \\
- 1 \left\{(F_{\theta,n}^{-1}(u) - F_{\theta}^{-1}(u)) - (G_{\theta,m}^{-1}(u) - G_{\theta}^{-1}(u)) \right\} \right\} du,
\]

by a Taylor expansion of the map \( (x,y) \mapsto |x-y|^r \) about \( (x,y) = (F_{\theta}^{-1}(u), G_{\theta}^{-1}(u)) \), where \( \tilde{F}_{\theta,n}^{-1}(u) \) (resp. \( \tilde{G}_{\theta,m}^{-1}(u) \)) is a real number on the line joining \( F_{\theta}^{-1}(u) \) and \( F_{\theta,n}^{-1}(u) \) (resp. \( G_{\theta}^{-1}(u) \) and \( G_{\theta,m}^{-1}(u) \)). Setting

\[
a_{\theta} = \min \left\{ F_{\theta}^{-1}(\delta), F_{\theta,n}^{-1}(\delta), G_{\theta}^{-1}(\delta), G_{\theta,m}^{-1}(\delta) \right\}, \\
b_{\theta} = \min \left\{ F_{\theta}^{-1}(1-\delta), F_{\theta,n}^{-1}(1-\delta), G_{\theta}^{-1}(1-\delta), G_{\theta,m}^{-1}(1-\delta) \right\},
\]

34
we then have
\[
|D_{nm}| \leq r \int_{S^{d-1}} \int_{-\delta}^{1-\delta} |\bar{\phi}_{\theta}^{-1}(u) - \bar{G}_{\theta,m}^{-1}(u)| \, du \, d\mu(\theta)
\]
\[
\leq r \int_{S^{d-1}} (b_\theta - a_\theta)^{-1} \int_{-\delta}^{1-\delta} \left[ |F_{\theta,n}^{-1}(u) - F_{\theta}^{-1}(u)| + |G_{\theta,m}^{-1}(u) - G_{\theta}^{-1}(u)| \right] \, du \, d\mu(\theta)
\]
\[
\leq r \int_{S^{d-1}} (b_\theta - a_\theta)^{-1} \left\{ \int_{a_\theta}^{b_\theta} |F_{\theta,n}(x) - F_{\theta}(x)| \, dx + \int_{a_\theta}^{b_\theta} |G_{\theta,m}(x) - G_{\theta}(x)| \, dx \right\} d\mu(\theta)
\]
\[
\leq r \int_{S^{d-1}} (b_\theta - a_\theta)^{r} \left\{ \sup_{x \in \mathbb{R}} |F_{\theta,n}(x) - F_{\theta}(x)| + \sup_{x \in \mathbb{R}} |G_{\theta,m}(x) - G_{\theta}(x)| \right\} d\mu(\theta), \tag{43}
\]
\[
=: r \int_{S^{d-1}} Z_{nm}(\theta) d\mu(\theta). \tag{44}
\]
Fix $\theta \in S^{d-1}$. The Dvoretzky-Kiefer-Wolfowitz inequality (Example 1) implies that, for all $t > 0$, with probability at least $1 - 4 \exp(-2(n \wedge m)t^2)$,
\[
a_\theta \geq F_{\theta}^{-1}(\delta - t) \wedge G_{\theta}^{-1}(\delta - t), \quad b_\theta \leq F_{\theta}^{-1}(1 - \delta + t) \vee G_{\theta}^{-1}(1 - \delta + t),
\]
and
\[
\sup_{x \in \mathbb{R}} |F_{\theta,n}(x) - F_{\theta}(x)| \leq t, \quad \sup_{x \in \mathbb{R}} |G_{\theta,m}(x) - G_{\theta}(x)| \leq t
\]
Since $\delta$ is held fixed, there exists a constant $c_1 > 0$ such that $\delta \geq c_1(n \wedge m)^{-1/2}$. We deduce that when $t \geq t_{nm} := c_1(n \wedge m)^{-1/2}/2$, with probability at least $1 - 4 \exp(-2(n \wedge m)t^2)$, $Z_{nm}(\theta) \lesssim tM_{nm}(\theta)$, where
\[
M_{nm}(\theta) = \max \left\{ |F_{\theta}^{-1}(\delta/2)|, |G_{\theta}^{-1}(\delta/2)|, |F_{\theta}^{-1}(1 - \delta/2)|, |G_{\theta}^{-1}(1 - \delta/2)| \right\}^r
\]
We then have
\[
\mathbb{E}_{P \otimes_{n \otimes m} Q \otimes m}[Z_{nm}(\theta)] = \int_0^\infty \mathbb{P}(Z_{nm}(\theta) \geq t) \, dt
\]
\[
= \int_0^{t_{nm}M_{nm}(\theta)} \mathbb{P}(Z_{nm}(\theta) \geq t) \, dt + \int_{t_{nm}M_{nm}(\theta)}^\infty \mathbb{P}(Z_{nm}(\theta) \geq t) \, dt
\]
\[
\leq t_{nm}M_{nm}(\theta) + \int_{t_{nm}M_{nm}(\theta)}^\infty 4 \exp\left(-2nt^2/M_{nm}^2(\theta)\right) \, dt
\]
\[
\lesssim \frac{M_{nm}(\theta)}{\sqrt{n \wedge m}}.
\]
Returning to equation (44), applying Fubini’s Theorem and Lemma 3, we obtain
\[
\mathbb{E}[D_{nm}] \leq r \int_{S^{d-1}} \mathbb{E}[Z_{nm}(\theta)] \, d\mu(\theta) \lesssim (n \wedge m)^{-1/2} \int_{S^{d-1}} M_{nm}(\theta) \, d\mu(\theta) \lesssim (n \wedge m)^{-1/2}.
\]
The claim follows.
C Proof of Theorem 1

Throughout the proof, KL denotes the Kullback-Leibler divergence, and $\chi^2$ denotes the $\chi^2$-divergence. In view of equation (8) and its natural analogue for the Sliced Wasserstein distance, together with the fact that all distributions considered below are compactly supported, there will be no loss of generality in assuming $\delta = 0$ in what follows.

At a high-level, our general approach is to carefully construct two pairs of distributions $(P_0,Q_0)$ and $(P_1,Q_1) \in \mathcal{O}$ such that the corresponding product measures $(P_0^\otimes n \otimes Q_0^\otimes m)$ and $(P_1^\otimes n \otimes Q_1^\otimes m)$ are close in the KL distance, but such that $SW_r(P_0,Q_0)$ and $SW_r(P_1,Q_1)$ are sufficiently different. In particular, if we can show that $KL(P_0^\otimes n \otimes Q_0^\otimes m, P_1^\otimes n \otimes Q_1^\otimes m) \leq \zeta < \infty$, then via an application of Le Cam’s inequality (see for instance, Theorem 2.2 of Tsybakov (2008)), we obtain the minimax lower bound that,

$$\mathcal{R}_{nm}(\mathcal{O},r) \gtrsim c_\zeta |SW_r(P_0,Q_0) - SW_r(P_1,Q_1)|,$$

(45)

where $c_\zeta > 0$ is a constant depending only on $\zeta$. We will use four separate constructions to handle various cases of the Theorem.

Let $\epsilon_n = k_r n^{-1/2}$, for a constant $k_r \in (0,1)$, possibly depending on $r$, to be determined below. We use the following pairs of distributions.

- **Construction 1.** For a vector $A = (a,0,...,0) \in \mathbb{R}^d$, and for $g > 0$, we define:

  $$P_{01} = \frac{1}{2} \delta_0 + \frac{1}{2} \delta_A, \quad Q_{01} = \frac{1}{2} \delta_gA + \frac{1}{2} \delta_{(1+g)A}$$

  $$P_{11} = \left(\frac{1}{2} + \epsilon_n\right) \delta_0 + \left(\frac{1}{2} - \epsilon_n\right) \delta_A, \quad Q_{11} = \frac{1}{2} \delta_gA + \frac{1}{2} \delta_{(1+g)A}.$$

- **Construction 2.** For $\gamma_2, \Delta > 0$ to be chosen in the sequel we let $P_{02}, P_{12}, Q_{02}, Q_{12} \in \mathcal{P}(\mathbb{R}^d)$ be the probability distributions of random vectors of the form $(X,0,...,0) \in \mathbb{R}^d$, with $X$ respectively distributed according to the distributions

  $$P_{02}^{(1)} = U\left(0,\gamma_2^{1/r}\right), \quad Q_{02}^{(1)} = U\left(\Delta \gamma_{2}^{1/r}, (1+\Delta) \gamma_{2}^{1/r}\right)$$

  $$P_{12}^{(1)} = \frac{1 + \epsilon_n}{2} U\left(0, \gamma_2^{1/r}/2\right) + \frac{1 - \epsilon_n}{2} U\left(\gamma_2^{1/r}/2, \gamma_2^{1/r}\right), \quad Q_{12}^{(1)} = U\left(\Delta \gamma_{2}^{1/r}, (1+\Delta) \gamma_{2}^{1/r}\right).$$

- **Construction 3.** For $0 < s_1 \leq s_2$ we let $P_{03}, P_{13}, Q_{03}, Q_{13} \in \mathcal{P}(\mathbb{R}^d)$ be the probability distributions of random vectors of the form $(X,0,...,0) \in \mathbb{R}^d$, with $X$ respectively distributed according to the distributions

  $$P_{03}^{(1)} = U\left(0,s_1^{1/r}\right), \quad Q_{03}^{(1)} = U\left(0,s_2^{1/r}\right),$$

  $$P_{13}^{(1)} = U\left(0,s_1^{1/r}\right), \quad Q_{13}^{(1)} = (1 - \epsilon_m) U\left(0,s_2^{1/r}\right) + \epsilon_m \delta_{s_2^{1/r}}.$$

- **Construction 4.** For $0 < s_2 \leq s_1$ we let $P_{04}, P_{14}, Q_{04}, Q_{14} \in \mathcal{P}(\mathbb{R}^d)$ be the probability distributions of random vectors of the form $(X,0,...,0) \in \mathbb{R}^d$, with $X$ respectively distributed according to the
Constructions 1 uses pairs of distributions with infinite $S J_r$, while Constructions 2-4 use pairs of distributions with finite $S J_r$. To compactly state our next result we define several terms,

$$t_r := \left( \int_{S^{d-1}} |\theta_1|^r d\mu(\theta) \right)^{\frac{1}{r}},$$

$$\beta := (s_2/s_1)^{1/r}, \quad \bar{\beta} := 1/\beta,$$

$$\Delta_\beta := \beta - 1, \quad \Delta_{\bar{\beta}} := \bar{\beta} - 1.$$

With these definitions in place the following technical lemma describes the main features of our constructions.

**Lemma 4.** There exists a choice of constant $k_r \in (0, 1)$ for which the following statements hold.

- **Construction 1.** Let $g := \Gamma / \left( \int_{S^{d-1}} |A^\top \theta|^r d\mu(\theta) \right)^{1/r}$. Then, there exists a constant $c_1 > 0$, possibly depending on $r$, such that

  $$SW_r^r(P_{01}, Q_{01}) = \Gamma^r,$$

  $$SW_r^r(P_{11}, Q_{11}) \geq \Gamma^r + c_1 \epsilon_n.$$

  Furthermore, there exists a choice of the vector $A$ for which $P_{01}, Q_{01}, P_{11}, Q_{11} \in \mathcal{O}(\Gamma; \infty, \infty)$.

- **Construction 2.** There exists a constant $c_2 > 0$, possibly depending on $r$, such that

  $$SW_r^r(P_{02}, Q_{02}) = \gamma_2 (t_r \Delta)^r,$$

  $$SW_r^r(P_{12}, Q_{12}) \geq \gamma_2 t_r^r \left\{ \Delta^r + c_2 \Delta^{r-1} \epsilon_n \right\}.$$

  Furthermore, $P_{02}, Q_{02}, P_{12}, Q_{12} \in \mathcal{O}(0; \gamma_2, \gamma_2)$.

- **Construction 3.** Assume that $\bar{\beta} \in (0, 1]$. Then,

  $$SW_r^r(P_{03}, Q_{03}) = \frac{s_2 t_r^r |\Delta_{\bar{\beta}}|^r}{r+1},$$

  $$SW_r^r(P_{13}, Q_{13}) \geq \frac{t_r^r s_2}{r+1} \left\{ |\Delta_{\bar{\beta}}|^r + r |\Delta_{\bar{\beta}}|^{r-1} \epsilon_m \right\}.$$

  Furthermore, $P_{03}, Q_{03}, P_{13}, Q_{13} \in \mathcal{O}(0; s_1, s_2)$.  

37
Construction 4. Assume that $\beta \in (0,1]$. Then,

$$SW_r(P_{04}, Q_{04}) = \frac{s_1 t_r^r |\Delta|}{r+1},$$

$$SW_r(P_{14}, Q_{14}) \geq \frac{t_r^r s_1}{r+1} \left\{ |\Delta|^r + r|\Delta|^{r-1} \epsilon_n \right\}.$$

Furthermore, $P_{04}, Q_{04}, P_{14}, Q_{14} \in O(0; s_1, s_2)$.

In each case, for some fixed universal constant $\zeta > 0$ we have that,

$$KL(P_{0i}^{\otimes m} \otimes Q_{0i}^{\otimes m}, P_{1i}^{\otimes m} \otimes Q_{1i}^{\otimes m}) \leq \zeta < \infty, \quad i = 1, 2, 3, 4.$$

Taking this result as given, we can now complete the proof of the Theorem. Using Construction 1 with $\Gamma = 0$, we obtain from equation (45) that

$$R_{nm}(O(0; \infty, \infty); r) \geq c_\zeta |SW_r(P_{01}, Q_{01}) - SW_r(P_{11}, Q_{11})| \geq \epsilon_n^{1/r} \asymp n^{-1/2r}.$$

Reversing the roles of $n$ and $m$ we obtain the first claim of part (i) of the Theorem. Choosing $\Gamma$ to be a strictly positive constant, we instead obtain

$$R_{nm}(O(\Gamma, \infty, \infty); r) \geq \Gamma \left\{ 1 - \left(1 + \frac{c_1 \epsilon_n}{\Gamma r} \right)^{\frac{1}{2}} \right\} = \frac{c_1 k_r n^{-1/2}}{\Gamma_{r-1} (1+o(1))},$$

by a first-order Taylor expansion of the map $x \mapsto (1+x)^{\frac{1}{2}}$. The fact that $\Gamma$ is bounded away from zero then implies $R_{nm}(O(\Gamma, \infty, \infty); r) \geq n^{-1/2}$ which proves part (ii) of the theorem, again upon reversing the roles of $n$ and $m$. It thus only remains to prove the second claim of part (i).

Without loss of generality we assume that $n \leq m$ in the remainder of the proof, noting that as above we may always reverse the roles of $n$ and $m$ and repeat our constructions. We consider four cases.

Case 1: $-1 \leq \Delta \leq -\epsilon_n$.

In this case, the condition $\Delta \leq 0$ implies $s_1 \geq s_2$. Since $n \leq m$, it therefore suffices to prove $R_{nm}(O(0, s_1, s_2); r) \geq s_1^{1/r} \epsilon_n$. Furthermore, since $\beta \leq 1$, we may invoke Construction 4 to obtain

$$|SW_r(P_{04}, Q_{04}) - SW_r(P_{14}, Q_{14})| \geq \frac{s_1^{1/r} t_r^r |\Delta|}{(r+1)^{1/r}} \left( \left(1 + \frac{r \epsilon_n}{|\Delta|} \right)^{\frac{1}{2}} - 1 \right) \asymp s_1^{1/r} \epsilon_n,$$

where we have used the assumption $|\Delta| \geq \epsilon_n$ in the last order assessment of the above display. This fact together with equation (45) yields the desired lower bound for Case 1.

Case 2: $-\epsilon_n < \Delta \leq 0$. The inequality $s_1 \geq s_2$ continues to hold, thus it suffices to prove $R_{nm}(O(0; s_1, s_2); r) \geq s_1^{1/r} \epsilon_n$. Notice further that

$$s_2^{1/r} \epsilon_n = s_1^{1/r} \beta \epsilon_n > s_1^{1/r} (1-\epsilon_n) \epsilon_n = s_1^{1/r} \epsilon_n (1+o(1)).$$

It will therefore suffice to prove $R_{nm}(O(0; s_1, s_2); r) \geq s_2^{1/r} \epsilon_n$. We use Construction 2, and choose $\gamma_2 = s_2$, and $\Delta \in (0,1]$ to be a constant larger than $\epsilon_n$. We observe that all distributions have $SJ_r$ functional at
most $s_2 = \min\{s_1, s_2\}$. Furthermore,

$$|SW_r(P_{02}, Q_{02}) - SW_r(P_{12}, Q_{12})| \geq s_2^{1/r} t_r \Delta \left( \frac{1 + c_2 \epsilon_n}{\Delta} \right)^{\frac{1}{2}} - 1.$$ 

Since $\Delta \geq \epsilon_n$, it is a straightforward observation that the factor in braces of the above display is of order $\epsilon_n$, thus we have shown

$$|SW_r(P_{02}, Q_{02}) - SW_r(P_{12}, Q_{12})| \geq s_2^{1/r} \epsilon_n,$$

and this together with equation (45) yields the desired lower bound for Case 2.

**Case 3:** $-1 \leq \Delta_3 \leq -\epsilon_m$ and $s_1^{1/r} \epsilon_m \leq s_2^{1/r} \epsilon_m$. In this case, it suffices to prove $\mathcal{R}_{nm}(O(0; s_1, s_2); r) \geq s_2^{1/r} \epsilon_m$. Notice that $\Delta_3 \leq 1$, hence we may use Construction 3 to obtain

$$|SW_r(P_{03}, Q_{03}) - SW_r(P_{13}, Q_{13})| \geq s_2^{1/r} t_r |\Delta_3| \left( \frac{1 + \epsilon_m}{|\Delta_3|} \right)^{\frac{1}{2}} - 1 \geq s_2^{1/r} \epsilon_m,$$

where we have used the assumption $|\Delta_3| \geq \epsilon_m$ in the last order assessment of the above display. This fact together with equation (45) yields the desired lower bound for Case 1.

**Case 4:** $-\epsilon_m < \Delta_3 < 0$ or $s_1^{1/r} \epsilon_m > s_2^{1/r} \epsilon_m$. Notice that if the condition $\Delta_3 > -\epsilon_m$ is satisfied, it implies

$$s_1^{1/r} \epsilon_m = s_2^{1/r} \beta \epsilon_m > (1 - \epsilon_m) \epsilon_m s_2^{1/r} \geq (1 - \epsilon_m) \epsilon_m s_2^{1/r} \geq s_1^{1/r} \epsilon_m.$$

For this case, it will thus suffice to prove $\mathcal{R}_{nm}(O(0; s_1, s_2); r) \geq s_1^{1/r} \epsilon_m$. Since $\Delta_3 \leq 0$, we observe that all distributions have SJ$_r$ functional at most $s_1 = \min\{s_1, s_2\}$. Invoking Construction 2 with $\gamma_2 = s_1$, the remainder of the argument follows similarly as in Case 2.

It remains to establish Lemma 4 and we turn our attention to this now.

**C.1 Proof of Lemma 4**

Bounding the KL divergence in each case is straightforward. We observe that for each $1 \leq i \leq 4$,

$$\text{KL}(P_{0i}^{\otimes n} \otimes Q_{0i}^{\otimes m}, (P_{1i}^{\otimes n} \otimes Q_{1i}^{\otimes m})) = n \text{KL}(P_{0i}, P_{1i}) + m \text{KL}(Q_{0i}, Q_{1i}) \leq n \chi^2(P_{0i}, P_{1i}) + m \chi^2(Q_{0i}, Q_{1i}).$$

The $\chi^2$ divergences in each construction can be computed in closed form. Doing so yields the bounds:

$$\text{KL}(P_{0i}^{\otimes n} \otimes Q_{0i}^{\otimes m}, P_{1i}^{\otimes n} \otimes Q_{1i}^{\otimes m}) \leq n \epsilon_n^2, \quad i = 1, 2, 4,$$

$$\text{KL}(P_{03}^{\otimes n} \otimes Q_{03}^{\otimes m}, P_{13}^{\otimes n} \otimes Q_{13}^{\otimes m}) \leq m \epsilon_m^2.$$ 

Together with the definition of $\epsilon_n$, we obtain the desired bounds on the KL divergence.

As a second preliminary let us verify that for appropriate choice of various parameters the distributions we construct have appropriately bounded moments, and belong to the class $K_r(b)$ defined in equation (19).

Notice first that the distributions $P_{01}, Q_{01}, P_{11}, Q_{11}$ have support with diameter bounded above by $(1 + G) a.$
Choosing a (possibly depending on $G$ and hence $\Gamma$) such that this expression is bounded above by $b^{1/r}$ ensures $P_{01}, Q_{01}, P_{11}, Q_{11} \in K_r(b)$. We are guaranteed that such a choice exists by using the assumption in equation (22), which ensures that $\Gamma$ cannot be too large.

Furthermore, the distributions $P_{ij}, Q_{ij}$ for $i=2,3,4$ and $j=0,1$ have supports with diameter bounded above by $s(1+\Delta) \leq 2s$. The assumption $b \geq (2s)^{1/r}$ therefore guarantees $P_{ij}, Q_{ij} \in K_r(b)$ for $i=2,3,4$ and $j=0,1$.

We now consider each construction in turn, establishing the remaining claims. As a preliminary technical result, it will be useful to study the Wasserstein distance between several pairs of univariate distributions.

**Lemma 5.** 1. Let $\Delta \geq \epsilon > 0$, and define the distributions

\[ \nu = \frac{1+\epsilon}{2} U(0,1/2) + \frac{1-\epsilon}{2} U(1/2,1), \]

and $\rho = U(\Delta, 1+\Delta)$. Then,

\[ W_r^r(\nu, \rho) \geq \Delta^r + \frac{r}{4} \epsilon \Delta^{r-1}. \]

2. Given $\xi \in (0,1], \Delta_\xi = \xi - 1$, define for all $\epsilon \in (0,1]$,

\[ \nu = U(0, \xi), \quad \rho = (1-\epsilon) U(0,1) + \epsilon \delta_1. \]

Then,

\[ W_r(\nu, \rho) \geq \frac{1}{r+1} \left[ |\Delta_\xi|^r + (r+1) \epsilon |\Delta_\xi|^{r-1} \right] \]

We prove this result in Appendix C.1.1. Taking this result as given, we can now compute the various Sliced Wasserstein distances and $S_{1/r}$ functionals.

**Computing the Sliced Wasserstein distances.**

- **Construction 1.** For any $\theta \in S^{d-1}$, let $F_{01,\theta}^{-1}, F_{11,\theta}^{-1}, G_{01,\theta}^{-1}$ and $G_{11,\theta}^{-1}$ denote the respective quantile functions of $\pi_\theta \# P_{01}, \pi_\theta \# P_{11}, \pi_\theta \# Q_{01}, \pi_\theta \# Q_{11}$. We have

\[ F_{01,\theta}^{-1}(u) = \begin{cases} 0 \land A^\top \theta, & u \in (0, 1/2) \\ 0 \lor A^\top \theta, & u \in [1/2, 1) \end{cases}, \quad F_{11,\theta}^{-1}(u) = \begin{cases} 0 \land A^\top \theta, & u \in (0, 1/2 + \epsilon_n) \\ 0 \lor A^\top \theta, & u \in [1/2 + \epsilon_n, 1) \end{cases}, \]

\[ G_{01,\theta}^{-1}(u) = G_{11,\theta}^{-1} = \begin{cases} gA^\top \theta \land (1+g)A^\top \theta, & u \in (0, 1/2) \\ gA^\top \theta \lor (1+g)A^\top \theta, & u \in [1/2, 1) \end{cases}. \]
Therefore,
\[
\text{SW}_r(P_0, Q_0) = \int_{\mathbb{S}^{d-1}} \int_0^1 |F_{01, \theta}^{-1}(u) - G_{01, \theta}^{-1}(u)|^r \, d\mu(\theta)
\]
\[
= \frac{1}{2} \int_{\theta \in \mathbb{S}^{d-1}, A^\top \theta \geq 0} |gA^\top \theta|^r \, d\mu(\theta) + \frac{1}{2} \int_{\theta \in \mathbb{S}^{d-1}, A^\top \theta < 0} |A^\top \theta - (1+g)A^\top \theta|^r \, d\mu(\theta)
\]
\[
= g^r \int_{\mathbb{S}^{d-1}} |A^\top \theta|^r \, d\mu(\theta)
\]
\[
= \Gamma^r.
\]

Furthermore,
\[
\text{SW}_r(P_{11}, Q_{11}) = \int_{\mathbb{S}^{d-1}} \int_0^1 |F_{1, \theta}^{-1}(u) - G_{11, \theta}^{-1}(u)|^r \, d\mu(\theta)
\]
\[
= \int_{\mathbb{S}^{d-1}} \left( \int_0^{1/2} |0 \wedge A^\top \theta - gA^\top \theta \wedge (1+g)A^\top \theta|^r \, du \right.
\]
\[
+ \int_{1/2}^{1/2 + \epsilon_n} |0 \wedge A^\top \theta - gA^\top \theta \vee (1+g)A^\top \theta|^r \, du \right)
\]
\[
+ \int_{1/2 + \epsilon_n}^1 |0 \wedge A^\top \theta - gA^\top \theta \vee (1+g)A^\top \theta|^r \, d\mu(\theta)
\]
\[
= (1-\epsilon_n)g^r \int_{\mathbb{S}^{d-1}} |A^\top \theta|^r \, d\mu(\theta) + \epsilon_n \int_{\mathbb{S}^{d-1}} |0 \wedge A^\top \theta - gA^\top \theta \vee (1+g)A^\top \theta|^r \, d\mu(\theta)
\]
\[
= (1-\epsilon_n)g^r \Gamma^r + \epsilon_n \int_{\mathbb{S}^{d-1}} |0 \wedge A^\top \theta - gA^\top \theta \vee (1+g)A^\top \theta|^r \, d\mu(\theta)
\]
\[
= \Gamma^r + c_1 \epsilon_n,
\]
for a positive constant \(c_1 > 0\). It follows that \(\text{SW}_r(P_{11}, Q_{11}) \geq \text{SW}_r(P_0, Q_0) \geq \Gamma\), thus \((P_0, Q_0), (P_{11}, Q_{11}) \in \mathcal{O}(\Gamma; \infty, \infty)\), and
\[
|\text{SW}_r(P_0, Q_0) - \text{SW}_r(P_{11}, Q_{11})| = |\Gamma - (\Gamma^r + c_1 \epsilon_n)^{1/r}|.
\]

- **Construction 2.** We use the first part of Lemma 5, and let \(\nu = \frac{1+\epsilon_n}{2} U(0, 1/2) + \frac{1-\epsilon_n}{2} U(1/2, 1)\), and \(\rho = U(\Delta, 1+\Delta)\). Notice that if \(X \sim \nu\), then \(\gamma_2^{1/r} X \sim P_{12}^{(1)}\), and if \(Y \sim \rho\), then \(\gamma_2^{1/r} Y \sim Q_{02}^{(1)}\). Therefore, by Proposition 7.16 of Villani (2003), \(W_r(\pi_{\theta \# P_{12}, \pi_{\theta \# Q_{12}}}) = |\theta_1|^{1/r} W_r(\nu, \rho)\). Thus,
\[
\text{SW}_r(P_{12}, Q_{12}) = \int_{\mathbb{S}^{d-1}} W_r(\pi_{\theta \# P_{12}, \pi_{\theta \# Q_{12}}}) \, d\mu(\theta)
\]
\[
= \int_{\mathbb{S}^{d-1}} |\theta_1|^{r} \gamma_2 W_r(\nu, \rho) \, d\mu(\theta) \geq \gamma_2^r \left[ \Delta^r + \frac{r}{4} \Delta^{r-1} \epsilon_n \right].
\]
by Lemma 5. Furthermore, it is easy to show that
\[
SW_r^r(P_{02}, Q_{02}) = \gamma_2 \Delta^r \int_{S^d-1} |\theta_1|^r d\mu(\theta) = \gamma_2 (t_r \Delta)^r.
\]

**Construction 3.** We use the second part of Lemma 5. We set
\[
\nu = U(0, \tilde{\beta}), \quad \rho = (1 - \epsilon_m) U(0, 1) + \epsilon_m \delta_1.
\]
Then, for all \( \epsilon \in (0, 1) \),
\[
W_r(\nu, \rho) \geq \frac{1}{r+1} \left[ |\Delta_{\tilde{\beta}}|^r + r\epsilon_m |\Delta_{\tilde{\beta}}|^{r-1} \right]
\]
We then obtain
\[
SW_r^r(P_{13}, Q_{13}) = \int_{S^d-1} |\theta_1|^r s_2 W_r^r(\nu, \rho) d\mu(\theta) \geq \frac{t_r s_2}{r+1} \left[ |\Delta_{\tilde{\beta}}|^r + r\epsilon_m |\Delta_{\tilde{\beta}}|^{r-1} \right].
\]
On the other hand, it is easy to see that
\[
SW_r^r(P_{03}, Q_{03}) = \frac{s_2 t_r |\Delta_{\tilde{\beta}}|^r}{r+1}.
\]

**Construction 4.** We again use the second part of Lemma 5, setting
\[
\nu = U(0, \beta), \quad \rho = (1 - \epsilon) U(0, 1) + \epsilon_\alpha \delta_1.
\]
The rest follows by the same argument as for Construction 3.

**Computing the SJ functionals.** Our next step will be to compute the SJ\(_r\) functionals for the various distributions we have constructed. We note that for Construction 1 our distributions are allowed to have infinite SJ\(_r\) so we only need to consider Constructions 2-4. The calculations for Construction 3 and 4 follow along very similar lines to those of Construction 2, which we detail below.

**Construction 2.** We have
\[
SJ_r(Q_{02}) = SJ_r(Q_{12}) = SJ_r(P_{02}) \leq \int_{S^d-1} \int_0^1 \left( \frac{\sqrt{u(1-u)}}{1/|\theta_1|^{1/r_2}} \right)^r dud\mu(\theta) \leq \int |\theta_1|^{1/r_2} d\mu(\theta) \leq \gamma_2.
\]
Furthermore,
\[
SJ_r(Q_{13}) \leq \int_{S^d-1} \int_0^1 \left( \frac{\sqrt{u(1-u)}}{(1-\epsilon_n)/|\theta_1|^{1/r_2}} \right)^r dud\mu(\theta) \leq \frac{\gamma_2}{(1-\epsilon_n)^r} \int_0^1 [u(1-u)]^{\frac{r}{2}} du.
\]
Choosing the constant \( k_r > 0 \) to satisfy \( k_r < 1 - \left( \int_0^1 [u(1-u)]^{\frac{r}{2}} du \right)^\frac{1}{r} \) guarantees that the above display is bounded above by \( \gamma_2 \).

To complete the proof it remains to prove Lemma 5.
C.1.1 Proof of Lemma 5

We prove each of the two claims in turn.

**Proof of Claim (1).** Notice that the quantile functions of \( \nu \) and \( \rho \) are respectively given by

\[
F^{-1}(u) = \begin{cases} 
\frac{u}{1+\epsilon}, & 0 \leq u \leq (1+\epsilon)/2, \\
\frac{1}{2} + \frac{u-(1+\epsilon)/2}{1-\epsilon}, & (1+\epsilon)/2 \leq u \leq 1,
\end{cases} \\
G^{-1}(u) = (u+\Delta)I(0 \leq u \leq 1).
\]

Thus,

\[
W_r(\nu, \rho) = \int_0^1 |F^{-1}(u) - G^{-1}(u)|^r du
= \int_0^{(1+\epsilon)/2} \left| \Delta + u - \frac{u}{1+\epsilon} \right|^r du + \int_{(1+\epsilon)/2}^1 \left| \Delta + u - \frac{u-(1+\epsilon)/2}{1-\epsilon} \right|^r du
= (I) + (II),
\]

say. We have,

\[
(I) = \int_0^{(1+\epsilon)/2} \left[ \Delta + \frac{\epsilon}{1+\epsilon} u \right]^r du = \frac{1+\epsilon}{\epsilon(r+1)} \left\{ \left( \Delta + \frac{\epsilon}{2} \right)^{r+1} - \Delta^{r+1} \right\}.
\]

Also,

\[
(II) = \int_{(1+\epsilon)/2}^1 \left( \Delta + \frac{\epsilon}{1-\epsilon} - u \frac{\epsilon}{1-\epsilon} \right)^r du
= -\frac{1-\epsilon}{(r+1)\epsilon} \left\{ \left( \Delta + \frac{\epsilon}{1-\epsilon} - \frac{\epsilon}{1-\epsilon} \right)^{r+1} - \left( \Delta + \frac{\epsilon}{1-\epsilon} - \frac{(1+\epsilon)\epsilon}{2(1-\epsilon)} \right)^{r+1} \right\}
= -\frac{1-\epsilon}{\epsilon(r+1)} \left\{ \Delta^{r+1} - \left( \Delta + \frac{\epsilon}{2} \right)^{r+1} \right\}.
\]

Thus,

\[
(I) + (II) = \left\{ \left( \Delta + \frac{\epsilon}{2} \right)^{r+1} - \Delta^{r+1} \right\} \left\{ \frac{1+\epsilon}{\epsilon(r+1)} + \frac{1-\epsilon}{\epsilon(r+1)} \right\}
= \frac{2}{\epsilon(r+1)} \left\{ \left( \Delta + \frac{\epsilon}{2} \right)^{r+1} - \Delta^{r+1} \right\}
= \frac{2}{\epsilon(r+1)} \left\{ \Delta^{r+1} + \frac{r+1}{2} \Delta r + \frac{r(r+1)}{8} (\Delta + \epsilon)^{r-1} \epsilon^2 - \Delta^{r+1} \right\},
\]

for some \( \epsilon \in (0, \epsilon/2) \), by a first-order Taylor expansion. Therefore,

\[
W_r(\nu, \rho) = \Delta^{r+1} + \frac{r}{4} (\Delta + \epsilon)^{r-1} \epsilon \geq \Delta^{r+1} + \frac{r}{4} \Delta^{r-1} \epsilon,
\]

and the claim follows. □
Proof of Claim (2). The respective quantile functions of \( \nu \) and \( \rho \) are given by

\[
F^{-1}(u) = \begin{cases} 
\frac{u}{1-\epsilon}, & 0 \leq u \leq 1 - \epsilon, \\
1, & 1 - \epsilon < u \leq 1
\end{cases}, \quad G^{-1}(u) = \xi u I(0 \leq u \leq 1).
\]

Thus,

\[
SW_r(r, \nu, \rho) = \int |F^{-1}(u) - G^{-1}(u)| r \, du
\]

\[
= \int_{0}^{1-\epsilon} \left[ \frac{u}{1-\epsilon} - \xi u \right]^r \, du + \int_{1-\epsilon}^{1} |1 - \xi u|^r \, du \\
= \int_{0}^{1-\epsilon} \left[ \frac{u}{1-\epsilon} - \xi u \right]^r \, du + \int_{1-\epsilon}^{1} |1 - \xi u|^r \, du \\
= \frac{1-\epsilon}{r+1} - \Delta \xi + \epsilon \xi + \frac{1}{\xi(r+1)} \left[ (-\Delta \xi + \epsilon \xi)^{r+1} - (-\Delta \xi)^{r+1} \right] \\
= \frac{1}{r+1} \left[ (-\Delta \xi + \epsilon \xi)^r \left( 1 - \epsilon + \frac{-\Delta \xi + \epsilon \xi}{\xi} \right) - \frac{|\Delta \xi|^{r+1}}{\xi} \right] \\
= \frac{1}{\xi(r+1)} \left[ (|\Delta \xi| + \epsilon \xi)^r - |\Delta \xi|^{r+1} \right] \\
= \frac{1}{\xi(r+1)} \left[ (|\Delta \xi| + \epsilon \xi)^r - |\Delta \xi|^{r+1} \right] \\
\geq \frac{1}{\xi(r+1)} \left[ |\Delta \xi|^r + r \epsilon |\Delta \xi|^r - |\Delta \xi|^{r+1} \right] \\
= \frac{|\Delta \xi|^r}{r+1} + \frac{r \epsilon |\Delta \xi|^r}{r+1}.
\]

The claim follows. \( \square \)

D Proof of Theorem 2

We begin by formally stating assumptions B1-B3, made in the statement of Theorem 2.

B1 \( \gamma_{\epsilon,n,\eta_k,n} \), viewed as functions over \([\delta,1-\delta]\), are differentiable, invertible with differentiable inverses, and respectively increasing and decreasing functions of \( \epsilon \in (0,1) \). Furthermore, both are increasing functions of \( u \in [0,1] \).

B2 There exists a constant \( K_1 > 0 \) such that for each \( f \in \{\gamma_{r,N,n,\eta_k,n}^{-1} : \tau \in \{\alpha \wedge \epsilon, \epsilon\} \} \), we have

\[
\sup_{\delta \leq u \leq 1 - \delta} \left| \frac{\partial f(u)}{\partial u} - 1 \right| \leq K_1 \kappa_{\epsilon,n},
\]

where recall that \( \epsilon = (\epsilon \wedge \alpha)/N \).
B3 There exists a constant $K_2$ such that for each $f,g \in \{\gamma_{\tau,N,n}, \eta_{\tau,N,n} : \tau \in \{\epsilon, \epsilon \wedge \alpha\}\}$, we have
$$\left|g^{-1}(f_n(x))\right| \leq K_2 \kappa_{\epsilon,n}$$ for $x=0,1$, and
$$\sup_{\delta \leq u \leq 1-\delta} \left|\frac{\partial g^{-1}(f(u))}{\partial u} - 1\right| \leq K_2 \kappa_{\epsilon,n}.$$

The following Lemma will be used in the sequel.

**Lemma 6.** Let $s \in [0,\infty)$, and let $P \in K_2(b)$ be such that $S_{\tau \delta}(P) \leq s$. Then, there exists a constant $K > 0$ depending only on $s,r,\delta$ and a function $\tilde{z}_N$ of $\mu_N$ such that $\mathbb{E}_{\mu_N}[\tilde{z}_N] \leq KN^{-1/2}I(d \geq 2)$, such that for all $f,g \in \{\gamma_{\tau,n}, \eta_{\tau,n} : \tau \in \{\alpha, \epsilon, \epsilon \wedge \alpha\}\}$ and $h \in \{\gamma_{\tau,n}, \eta_{\tau,n} : \tau \in \{\alpha, \epsilon, \epsilon \wedge \alpha\}\}$ we have
$$\frac{1}{1-2\delta} \int_{\mathbb{S}^{d-1}} \int_{h(\delta)}^{h(1-\delta)} \left[F_{\theta}^{-1}(f(u)) - F_{\theta}^{-1}(g(u))\right]^r d\mu_N(\theta) \leq V_{\epsilon,n}(P)(1+o(1)) + \tilde{z}_N.$$

The proof of Lemma 6 appears in Appendix D.1. We now turn to the proof of the main result.

**Proof of Theorem 2.** When the symbol \( \lesssim \) is used during the proof, the underlying constants may be dependent on $r,s,b,\delta$, but not on $P$ and $Q$. We prove the claim in five steps.

**Step 0: Setup.** With probability at least $1-\epsilon$, uniformly in $j=1,...,N$, we have both
$$F_{\theta_j,n}^{-1}(\gamma_{\epsilon/N,n}(u)) \leq F_{\theta_j,n}^{-1}(u) \leq F_{\theta_j,n}^{-1}(\eta_{\epsilon/N,n}(u)),$$ and,
$$G_{\theta_j,m}^{-1}(\gamma_{\epsilon/m,n}(u)) \leq G_{\theta_j,m}^{-1}(u) \leq G_{\theta_j,m}^{-1}(\eta_{\epsilon/m,n}(u)).$$

All derivations which follow will be carried out on the event that the above two inequalities are satisfied, which has probability at least $1-\epsilon$. For notational simplicity, we will write $a = \alpha/N$, $\epsilon = \epsilon/N$, and we recall that $\epsilon = \epsilon \wedge \alpha$.

Recall that for all $\theta \in \{\theta_1,...,\theta_N\}$, $F_{\theta,n}$, $G_{\theta,m}$ denote the empirical CDFs of $F$ and $Q$ respectively, and $F_{\theta_j,n}^{-1}$, $G_{\theta_j,m}^{-1}$ their corresponding quantile functions. We may write
$$C_{nm} = \left[\left(\frac{1}{1-2\delta} \int_{\mathbb{S}^{d-1}} \int_{h(\delta)}^{h(1-\delta)} A_{\theta,n,m} d\mu_N(\theta)\right)^{\frac{1}{r}}, \left(\frac{1}{1-2\delta} \int_{\mathbb{S}^{d-1}} \int_{h(\delta)}^{h(1-\delta)} B_{\theta,n,m} d\mu_N(\theta)\right)^{\frac{1}{r}}\right],$$

where
$$A_{\theta,n,m}(u) = [F_{\theta,n}^{-1}(\gamma_{\alpha,n}(u)) - G_{\theta,m}^{-1}(\eta_{\alpha,m}(u))] \vee [G_{\theta,m}^{-1}(\gamma_{\alpha,m}(u)) - F_{\theta,n}^{-1}(\eta_{\alpha,n}(u))] \wedge 0,$$
$$B_{\theta,n,m}(u) = [F_{\theta,n}^{-1}(\gamma_{\alpha,n}(u)) - G_{\theta,m}^{-1}(\eta_{\alpha,m}(u))] \vee [G_{\theta,m}^{-1}(\gamma_{\alpha,m}(u)) - F_{\theta,n}^{-1}(\eta_{\alpha,n}(u))].$$

We will first show that for some $\tilde{z}_N$ depending on $\mu_N$,
$$\left|\frac{1}{1-2\delta} \int_{\mathbb{S}^{d-1}} \int_{h(\delta)}^{h(1-\delta)} A_{\theta,n,m}(u) d\mu_N(\theta) - [SW_{r \delta}^{(N)}(P,Q)]^r \right| \lesssim \psi_{\epsilon,n,m} + \varphi_{\epsilon,n,m} + \tilde{z}_N.$$

A similar argument can be used to bound this expression with $A_{\theta,n,m}$ replaced by $B_{\theta,n,m}$, and will lead to the claim.
We will assume without loss of generality that \( r > 1 \) in what follows. As will be clear from the proof, the arguments of Steps 2 and 4 alone may easily be used to prove the claim when \( r = 1 \).

**Step 1: Taylor Expansion Setup.** Let \( u \in [\delta, 1 - \delta] \) and \( \theta \in \{\theta_1, \ldots, \theta_N\} \). By Taylor expansion of the map \((x, y) \in \mathbb{R}^2 \to (x - y)^r\), there exists \( F_{\theta,n}^{-1}(u) \) (resp. \( G_{\theta,m}^{-1}(u) \)) on the segment joining \( F_{\theta,n}^{-1}(\gamma_{a,n}(u)) \) and \( F_{\theta,n}^{-1}(u) \) (resp. \( G_{\theta,m}^{-1}(\eta_{a,m}(u)) \)) such that

\[
[F_{\theta,n}^{-1}(\gamma_{a,n}(u)) - G_{\theta,m}^{-1}(\eta_{a,m}(u))]^r = [F_{\theta,n}^{-1}(u) - G_{\theta,m}^{-1}(u)]^r + \xi_{nm}(u),
\]

where

\[
\xi_{\theta,nm}(u) = r \bigg( \frac{\Gamma(n) - \gamma_{a,n}(u)}{F_{\theta,n}^{-1}(u) - G_{\theta,m}^{-1}(u)} \bigg)^{r-1} \bigg\{ \bigg( F_{\theta,n}^{-1}(\gamma_{a,n}(u)) - F_{\theta,n}^{-1}(u) \bigg) - \bigg( G_{\theta,m}^{-1}(\eta_{a,m}(u)) - G_{\theta,m}^{-1}(u) \bigg) \bigg\}.
\]

Likewise, there exists \( F_{\theta,n}^{-1}(u) \) (resp. \( G_{\theta,m}^{-1}(u) \)) on the segment joining \( F_{\theta,n}^{-1}(u) \) and \( F_{\theta,n}^{-1}(\eta_{a,n}(u)) \) (resp. \( G_{\theta,m}^{-1}(\gamma_{a,m}(u)) \)) and \( G_{\theta,m}^{-1}(u) \) such that

\[
[G_{\theta,m}^{-1}(\gamma_{a,m}(u)) - F_{\theta,n}^{-1}(\eta_{a,n}(u))]^r = [G_{\theta,n}^{-1}(u) - F_{\theta,n}^{-1}(u)]^r + \zeta_{\theta,nm}(u),
\]

where

\[
\zeta_{\theta,nm}(u) = r \bigg( \frac{\Gamma(n) - \eta_{a,n}(u)}{F_{\theta,n}^{-1}(u) - G_{\theta,m}^{-1}(u)} \bigg)^{r-1} \bigg\{ \bigg( F_{\theta,n}^{-1}(\eta_{a,n}(u)) - F_{\theta,n}^{-1}(u) \bigg) - \bigg( G_{\theta,m}^{-1}(\gamma_{a,m}(u)) - G_{\theta,m}^{-1}(u) \bigg) \bigg\}.
\]

Now, consider the numerical inequality \(|(a^r + b^r) + (a^r - b^r) \leq 3(|a| + |d|)|, for all \( a \in \mathbb{R}, b, d \in \mathbb{R}^+ \). Taking \( a = (F_{\theta}^{-1} - G_{\theta}^{-1}) \), \( b = \xi_{\theta,nm} \), and \( d = \zeta_{\theta,nm} \), this inequality implies

\[
\frac{1}{1 - 2\delta} \int_{\mathcal{S}^d - \delta}^{1 - \delta} |A_{\theta,nm}(u)d\mu_N(\theta) - \left[ \text{SW}_{r,\delta}^{(N)}(P,Q)^r \right] | \leq \frac{1}{1 - 2\delta} \int_{\mathcal{S}^d - \delta}^{1 - \delta} \left| (F_{\theta,n}^{-1}(u) - G_{\theta,n}^{-1}(u))^r + \xi_{\theta,nm}(u) \right| \bigg\} \bigg\} + 0
\]

\[
\leq \frac{1}{1 - 2\delta} \int_{\mathcal{S}^d - \delta}^{1 - \delta} \left| (F_{\theta,n}^{-1}(u) - G_{\theta,n}^{-1}(u))^r + \xi_{\theta,nm}(u) \right| \bigg\} \bigg\} + 0
\]

\[
\leq \frac{1}{1 - 2\delta} \int_{\mathcal{S}^d - \delta}^{1 - \delta} \left| \xi_{\theta,nm}(u) \right| d\mu_N(\theta) + \frac{1}{1 - 2\delta} \int_{\mathcal{S}^d - \delta}^{1 - \delta} \left| \xi_{\theta,nm}(u) \right| d\mu_N(\theta).
\]

It will now suffice to bound the second term of the above display, and a similar bound will hold for the first. Note that

\[
\int_{\mathcal{S}^d - \delta}^{1 - \delta} \left| \xi_{nm}(u) \right| d\mu_N(\theta) \leq r(\mathcal{I} + \mathcal{J}),
\]

46
where,

\[ I = \int_{\mathbb{R}^{d-1}} \int_\delta^{1-\delta} \left[ F_{\theta,n}(u) - G_{\theta,m}(u) \right] \left[ F_{\theta,n}(\gamma_{a,n}(u)) - F_{\theta}(u) \right] du d\mu_N(\theta) \]

\[ J = \int_{\mathbb{R}^{d-1}} \int_\delta^{1-\delta} \left[ F_{\theta,n}(u) - G_{\theta,m}(u) \right] \left[ G_{\theta,m}(\eta_{a,n}(u)) - G_{\theta}(u) \right] du d\mu_N(\theta). \]

It will suffice to prove that \( I \lesssim (1 - 2\delta) \psi_{\epsilon,m} \) and \( J \lesssim (1 - 2\delta) \psi_{\epsilon,m} \), up to terms depending only on \( N \). We consider the cases \( SJ_{r,\delta}(P) \backslash SJ_{r,\delta}(Q) \rangle > s \) and \( SJ_{r,\delta}(P) \backslash SJ_{r,\delta}(Q) \rangle \leq s \) separately.

**Step 2: Bounding \( I \) and \( J \) when \( SJ_{r,\delta}(P) \backslash SJ_{r,\delta}(Q) \rangle > s \).** We have,

\[ I \lesssim \int_{\mathbb{R}^{d-1}} \left( \sup_{\delta \leq u \leq 1-\delta} \left| F_{\theta,n}(u) - G_{\theta,m}(u) \right| \right) \left( \int_{\delta}^{1-\delta} \left| F_{\theta,n}(\gamma_{a,n}(u)) - F_{\theta}(u) \right| du \right) d\mu_N(\theta). \]

We will bound each factor in the above integral, beginning with the second. Using inequality (46), since \( e \geq \varepsilon \), we have for all \( u \in [\delta,1-\delta) \) and \( \theta \in \{\theta_1,\ldots,\theta_N\} \),

\[ |F_{\theta,n}(\gamma_{a,n}(u)) - F_{\theta}(u)| \leq \left| F_{\theta}(\gamma_{\varepsilon,n}(\gamma_{a,n}(u))) - F_{\theta}(u) \right| \right] \left( F_{\theta,n}(\gamma_{a,n}(u)) - F_{\theta}(u) \right) d\mu_N(\theta). \]

Now, write \( x_n = \gamma_{\varepsilon,n}(\gamma_{a,n}(\delta)) \) and \( y_n = \gamma_{\varepsilon,n}(\gamma_{a,n}(1-\delta)) \), which are respectively bounded away from 0 and 1 by condition \( A1(\delta/2;\alpha/N) \) and \( B1 \). Assumption \( B1 \) furthermore implies \( \delta \leq x_n \leq 1-\delta \leq y_n \), so for all \( \theta \in \{\theta_1,\ldots,\theta_N\} \),

\[ \int_{\mathbb{R}^{d-1}} \int_\delta^{1-\delta} \left[ F_{\theta}(\gamma_{\varepsilon,n}(\gamma_{a,n}(u))) - F_{\theta}(u) \right] du d\mu_N(\theta) \]

\[ = \int_{\mathbb{R}^{d-1}} \int_{x_n}^{y_n} F_{\theta}(u) \left( \partial_{\gamma_{\varepsilon,n}}(\gamma_{a,n}(u)) \right) du d\mu_N(\theta) - \int_{\mathbb{R}^{d-1}} \int_\delta^{1-\delta} F_{\theta}(u) du d\mu_N(\theta) \]

\[ \leq \int_{\mathbb{R}^{d-1}} \int_{x_n}^{y_n} F_{\theta}(u) (1 + \kappa_{\varepsilon,n}) du d\mu_N(\theta) - \int_{\mathbb{R}^{d-1}} \int_\delta^{1-\delta} F_{\theta}(u) du d\mu_N(\theta) \quad (\text{By B3}) \]

\[ \leq \int_{\mathbb{R}^{d-1}} \int_{1-\delta}^{y_n} F_{\theta}(u) du d\mu_N(\theta) \]

\[ - \int_{\mathbb{R}^{d-1}} \int_{\delta}^{x_n} F_{\theta}(u) du d\mu_N(\theta) + \kappa_{\varepsilon,n} \int_{\mathbb{R}^{d-1}} F_{\theta}(y_n) du d\mu_N(\theta) \]

\[ \leq (y_n - (1-\delta)) \int_{\mathbb{R}^{d-1}} F_{\theta}(y_n) du d\mu_N(\theta) \]

\[ - (x_n - \delta) \int_{\mathbb{R}^{d-1}} F_{\theta}(\delta) du d\mu_N(\theta) + \kappa_{\varepsilon,n} \int_{\mathbb{R}^{d-1}} F_{\theta}(y_n) du d\mu_N(\theta) \]

\[ \lesssim (y_n - (1+\delta) + (x_n - \delta) + \kappa_{\varepsilon,n} + \mathbb{E}[\varepsilon_{N}^{(1)}], \]

where \( \varepsilon_{N}^{(1)} = \max_{\theta \in \{\delta,1-\delta,x_n,y_n\}} \int_{\mathbb{R}^{d-1}} F_{\theta}(u) du d\mu_N(\theta) \). Notice that by assumption \( A1(\delta/2;\alpha/N) \) and Lemma 3, we have \( \mathbb{E}[\varepsilon_{N}^{(1)}] \lesssim N^{-1/2} I(d \geq 2) \). Therefore, applying the Mean Value Theorem to each of
We now bound the second factor in the above display. Since 
\[ \tilde{\mu} \] we similarly have,
\[ \| \tilde{\theta} - \mu \| \leq N(2) = (2) \leq \tilde{\theta} - \mu \] Combining these facts, we obtain
\[ I \leq (\kappa_{n} + \zeta_{N}^{(1)}) \int \sup_{\delta - 1 \leq u \leq 1 - \delta} \left| \tilde{F}_{\theta}^{-1}(u) - \tilde{G}_{\theta}^{-1}(u) \right| d\mu_{N}(\theta) \] (52)

We now bound the second factor in the above display. Since \( \tilde{F}_{\theta}^{-1}(u) \in [F_{\theta,n}^{-1}(\eta_{a,n}(u)), F_{\theta}^{-1}(u)] \) and \( \tilde{G}_{\theta}^{-1}(u) \in [G_{\theta,n}^{-1}(u), G_{\theta,m}(\eta_{a,m}(u))] \), we have for any \( u \in [\delta, 1 - \delta] \),
\[ \left| \tilde{F}_{\theta,n}^{-1}(u) - \tilde{G}_{\theta,m}^{-1}(u) \right| \leq \left| G_{\theta,n}^{-1}(\eta_{a,n}(u)) - F_{\theta,n}^{-1}(\gamma_{a,n}(u)) \right| \vee \left| F_{\theta,n}^{-1}(u) - G_{\theta}^{-1}(u) \right| \leq \left\{ \left| G_{\theta,n}^{-1}(\gamma_{a,n}(u)) - G_{\theta,m}(\eta_{a,m}(u)) \right| \right. \] (53)

It follows that
\[ \int \sup_{\delta - 1 \leq u \leq 1 - \delta} \left| \tilde{F}_{\theta,n}^{-1}(u) - \tilde{G}_{\theta,m}^{-1}(u) \right| d\mu_{N}(\theta) \lesssim (1 - 2\delta) \left\{ \left[ SW_{\infty,\delta}^{(N)}(P,Q) \right]^{r-1} + U_{\infty,n}^{(N)}(P) + U_{\infty,m}^{(N)}(Q) \right\} \] (54)

where the functional \( U_{\infty,n}^{(N)}(P) \) is defined identically as \( U_{\infty,n}(P) \) up to replacing the measure \( \mu \) by \( \mu_{N} \), and \( \zeta_{N}^{(2)} = |SW_{\infty,\delta}^{(N)}(P,Q) - SW_{\infty,\delta}(P,Q)| \vee |U_{\infty,n}^{(N)}(P) - U_{\infty,n}(P)| \vee |U_{\infty,m}^{(N)}(Q) - U_{\infty,m}(Q)| \). Notice that \( \mathbb{E}[\zeta_{N}^{(2)}] \leq N^{-1/2} I(d \geq 2) \), since \( P, Q \in \mathcal{K}_{2}(\theta) \). We conclude this section of the proof by combining the above display with equation (52). We then have
\[ \frac{I}{1 - 2\delta} \lesssim (\kappa_{n} + \zeta_{N}^{(1)}) \left\{ SW_{\infty,\delta}^{(N)}(P,Q) + U_{\infty,n}(P) + U_{\infty,m}(Q) + \zeta_{N}^{(2)} \right\} \lesssim \psi_{n,m} + (\zeta_{N}^{(1)} \vee \zeta_{N}^{(2)}) , \]
and by a symmetric argument,
\[ \mathcal{J} \leq \varphi_{r,m} + (\gamma_N^{(1)} \vee \gamma_N^{(2)}). \]

**Step 3: Bounding \( I \) and \( J \) when \( \text{SJ}_{r,\delta}(P) \vee \text{SJ}_{r,\delta}(Q) \leq s \).** By means of Holder’s inequality, we have

\[
I \leq \int_{\mathcal{S}_{d-1}} 1^{1-\delta} \left( F_{\theta,n}^{-1}(u) - \tilde{G}_{\theta,m}^{-1}(u) \right)^{r-1} \left| F_{\theta,n}^{-1}(\gamma_{a,n}(u)) - F_{\theta}^{-1}(u) \right| dud\nu_N(\theta)
\]

\[
\leq \left( \int_{\mathcal{S}_{d-1}} \int_{\delta}^{1-\delta} \left( F_{\theta,n}^{-1}(u) - \tilde{G}_{\theta,m}^{-1}(u) \right)^{r} dud\nu_N(\theta) \right)^{\frac{1}{r}} \left( \int_{\mathcal{S}_{d-1}} \int_{\delta}^{1-\delta} \left| F_{\theta,n}^{-1}(\gamma_{a,n}(u)) - F_{\theta}^{-1}(u) \right| dud\nu_N(\theta) \right)^{\frac{1}{p}}
\]

\[
\leq \left( \int_{\mathcal{S}_{d-1}} \int_{\delta}^{1-\delta} \left( F_{\theta,n}^{-1}(u) - \tilde{G}_{\theta,m}^{-1}(u) \right)^{r} dud\nu_N(\theta) \right)^{\frac{1}{r}} \times \left( \left( \int_{\mathcal{S}_{d-1}} \int_{\delta}^{1-\delta} \left| F_{\theta,n}^{-1}(\gamma_{a,n}(u)) - F_{\theta}^{-1}(u) \right| dud\nu_N(\theta) \right)^{\frac{1}{p}} \right)^{\frac{1}{p}}
\]

\[
= \left( \int_{\mathcal{S}_{d-1}} \int_{\delta}^{1-\delta} \left( F_{\theta,n}^{-1}(u) - \tilde{G}_{\theta,m}^{-1}(u) \right)^{r} dud\nu_N(\theta) \right)^{\frac{1}{r}} \times \left( \left( \int_{\mathcal{S}_{d-1}} \int_{\delta}^{1-\delta} \left| F_{\theta,n}^{-1}(\gamma_{a,n}(u)) - F_{\theta}^{-1}(u) \right| dud\nu_N(\theta) \right)^{\frac{1}{p}} \right)^{\frac{1}{p}}
\]

By \( B_2, \sup |\partial \gamma_{a,n}^{-1}(u)/\partial u| \leq 1 + \kappa_{r,n} \). By combining this fact with Lemma 6, the above display leads to

\[
\frac{I}{1-2\delta} \leq \left( \frac{1}{1-2\delta} \int_{\mathcal{S}_{d-1}} \int_{\delta}^{1-\delta} \left( F_{\theta,n}^{-1}(u) - \tilde{G}_{\theta,m}^{-1}(u) \right)^{r} dud\nu_N(\theta) \right)^{\frac{1}{r}} \left\{ V_{\epsilon,n}(P)(1 + o(1)) + \tilde{\gamma}_N \right\}^{\frac{1}{p}}
\]

\[
\leq \left( \frac{1}{1-2\delta} \int_{\mathcal{S}_{d-1}} \int_{\delta}^{1-\delta} \left( F_{\theta,n}^{-1}(u) - \tilde{G}_{\theta,m}^{-1}(u) \right)^{r} dud\nu_N(\theta) \right)^{\frac{1}{r}} \left\{ V_{\epsilon,n}(P) \right\}^{\frac{1}{p}} + \tilde{\gamma}_N^{\frac{1}{p}}. \tag{54}
\]

By using similar calculations as in equations (53) and (54) to bound the first factor in the above display,
we have
\[
\frac{1}{1-2\delta} \int_{\mathbb{R}^{d-1}} \int_{\delta}^{1-\delta} (\tilde{F}_{\theta,n}^{-1}(u) - \tilde{G}_{\theta,m}^{-1}(u))^r \, d\mu_N(\theta)
\]
\[
\lesssim [SW_{r,\delta}(P,Q)]^r + \frac{1}{1-2\delta} \int_{\mathbb{R}^{d-1}} \int_{\delta}^{1-\delta} |F_{\theta}^{-1}(u) - F_{\theta}^{-1}(\gamma_{\eta,n}(\gamma_{a,n}(u)))|^r \, d\mu_N(\theta)
\]
\[
+ \frac{1}{1-2\delta} \int_{\mathbb{R}^{d-1}} \int_{\delta}^{1-\delta} |G_{\theta}^{-1}(\gamma_{\varepsilon,m}(\gamma_{a,m}(u))) - G_{\theta}^{-1}(u)|^r \, d\mu_N(\theta)
\]
\[
\lesssim SW_{r,\delta}(P,Q) + V_{\varepsilon,n}(P) + V_{\varepsilon,m}(Q) + (\zeta_N^3),
\]
where \(\zeta_N^3 = |SW_{r,\delta}(P,Q) - [SW_{r,\delta}^{(N)}(P,Q)]^r|\), and thus \(E[\zeta_N^3] \leq N^{-1/2}I(d \geq 2)\). Define
\[
\zeta_N := \zeta_N^{(1)} \vee \zeta_N^{(2)} \vee \zeta_N^{(3)} \vee \zeta_N^3,
\]
so that \(E[\zeta_N] \lesssim N^{-1/2r}I(d \geq 2)\). Putting these facts together with equation (54), we arrive at
\[
\frac{I}{1-2\delta} \lesssim (SW_{r,\delta}(P,Q) + V_{\varepsilon,n}(P) + V_{\varepsilon,m}(Q))^{\frac{r-1}{r}} [V_{\varepsilon,n}(P)]^{\frac{1}{r}} + \zeta_N = \psi_{\varepsilon,\eta} + \zeta_N.
\]
Finally, by a symmetric argument, we also have \(\frac{I}{1-2\delta} \lesssim \psi_{\varepsilon,\eta} + \zeta_N\).

**Step 4: Conclusion.** Returning to equation (48), we have shown
\[
\frac{1}{1-2\delta} \int_{\mathbb{R}^{d-1}} \int_{\delta}^{1-\delta} |\kappa_{nm}(u)| \, d\mu_N(\theta) \lesssim \psi_{\varepsilon,\eta} + \varphi_{\varepsilon,\eta} + \zeta_N.
\]
By the same arguments, we may obtain the same upper bound, up to universal constant factors, on the integral \(\frac{1}{1-2\delta} \int_{\mathbb{R}^{d-1}} \int_{\delta}^{1-\delta} |\kappa_{nm}(u)| \, d\mu_N(\theta)\) in equation (48). We deduce that, for some \(c_1 > 0\) (possibly depending on \(s,b,\delta,r\)), we have
\[
\left( \frac{1}{1-2\delta} \int_{\mathbb{R}^{d-1}} \int_{\delta}^{1-\delta} A_{nm}^r(u) \, d\mu_N(\theta) \right)^{\frac{1}{r}} \leq \left\{ [SW_{r,\delta}^{(N)}(P,Q)]^r - c_1 (\psi_{\varepsilon,\eta} + \varphi_{\varepsilon,\eta} + \zeta_N) \right\}^{\frac{1}{r}}.
\]
The definition of \(\kappa_N\) in terms of \(\kappa_N^{(3)}\) then implies that for a different constant \(c_2 > 0\) (again, depending on \(s,b,\delta,r\)),
\[
\left( \frac{1}{1-2\delta} \int_{\mathbb{R}^{d-1}} \int_{\delta}^{1-\delta} A_{nm}^r(u) \, d\mu_N(\theta) \right)^{\frac{1}{r}} \leq \left\{ SW_{r,\delta}^r(P,Q) - c_2 (\psi_{\varepsilon,\eta} + \varphi_{\varepsilon,\eta} + \zeta_N) \right\}^{\frac{1}{r}}.
\]
By the same arguments, there exists a constant \(c_3 > 0\) (again, depending on \(s,b,\delta,r\)), such that
\[
\left( \frac{1}{1-2\delta} \int_{\mathbb{R}^{d-1}} \int_{\delta}^{1-\delta} B_{nm}^r(u) \, d\mu_N(\theta) \right)^{\frac{1}{r}} \leq \left\{ SW_{r,\delta}(P,Q) + c_3 (\psi_{\varepsilon,\eta} + \varphi_{\varepsilon,\eta} + \zeta_N) \right\}^{\frac{1}{r}}.
\]
The claim follows by choosing \(c = c_2 \vee c_3\).

It now remains to prove Lemma 6.
D.1 Proof of Lemma 6

Since $S_{r,\delta}(P) \leq s < \infty$, it follows from Lemma 2 that $F_{\theta}^{-1}$ is Lebesgue-almost everywhere differentiable for $\mu$-almost every $\theta \in S^{d-1}$. Thus by a first-order Taylor expansion,

$$\int_{S^{d-1}} \int_{h(\delta)}^{h(1-\delta)} \left[ F_{\theta}^{-1}(f^{-1}(u)) - F_{\theta}^{-1}(g^{-1}(u)) \right]^r d\mu_N(\theta) \leq \int_{S^{d-1}} \int_{h(\delta)}^{h(1-\delta)} \left( \frac{1}{p_\theta(F_{\theta}^{-1}(u))} \left| f^{-1}(u) - g^{-1}(u) \right| + o\left( |f^{-1}(u) - g^{-1}(u)| \right) \right)^r d\mu_N(\theta)$$

$$\leq \int_{S^{d-1}} \int_{h(\delta)}^{h(1-\delta)} \left( \frac{\kappa_{\varepsilon,n}(u)}{p_\theta(F_{\theta}^{-1}(u))} + o(\kappa_{\varepsilon,n}) \right)^r d\mu_N(\theta)$$

$$\leq \int_{S^{d-1}} \int_{\delta - \kappa_{\varepsilon,n}}^{\delta + \kappa_{\varepsilon,n}} \left( \frac{\kappa_{\varepsilon,n}(u)}{p_\theta(F_{\theta}^{-1}(u))} + o(\kappa_{\varepsilon,n}) \right)^r d\mu_N(\theta)$$

$$= (1 + O(\kappa_{\varepsilon,n})) \int_{S^{d-1}} \int_{h(\delta)}^{1-\delta} \left( \frac{\kappa_{\varepsilon,n}(u)}{p_\theta(F_{\theta}^{-1}(u))} + o(\kappa_{\varepsilon,n}) \right)^r d\mu_N(\theta)$$

$$= (1 + o(1)) \int_{S^{d-1}} \int_{\delta}^{1-\delta} \left( \frac{\kappa_{\varepsilon,n}(u)}{p_\theta(F_{\theta}^{-1}(u))} \right)^r d\mu_N(\theta).$$

Since $S_{r,\delta}(P) \leq s$ and $P \in \mathcal{K}_{2r}(b)$, the integrand in the above display is bounded by a constant depending on $s$ and $b$, uniformly in $P$. It follows that for some $K > 0$ depending on $s, r, \delta$, we have

$$\mathbb{E} [\kappa_N] \leq KN^{-1/2} I(d \geq 2), \quad \text{where} \quad \kappa_N := \int_{S^{d-1}} \int_{\delta}^{1-\delta} \left( \frac{\kappa_{\varepsilon,n}(u)}{p_\theta(F_{\theta}^{-1}(u))} \right)^r d\mu_N(\mu_N - \mu)(\theta).$$

The claim follows.

□

E Proof of Theorem 3

The proof of this result has two main components. In Lemma 8 we show that the Sliced Wasserstein distance is Hadamard differentiable under certain conditions. Theorem 3 then follows via an application of the delta method for the bootstrap.

Hadamard Differentiability of the Sliced Wasserstein distance. Throughout this subsection, for a metric space $(T, \rho)$, $C[T]$ denotes the set of real-valued continuous functions defined on $T$, endowed with the supremum norm, and $\ell^\infty(T) = \{ f : T \to \mathbb{R} : \sup_{t \in T} |f(t)| < \infty \}$. Let $D[I]$ denote the Skorokhod space of càdlàg functions defined over an interval $I = [a_1, a_2] \subseteq \mathbb{R}$, endowed with the supremum norm. We will make use of the following result from van der Vaart & Wellner (1996) (Lemma 3.9.20), regarding the Hadamard differentiability of the quantile function at a fixed point $u \in (a_1, a_2)$. Let $D_\psi$ denote the set of nondecreasing maps $A \in D[I]$ such that the set $\{ x \in I : A(x) \geq u \}$ is nonempty for any given $u \in (0,1)$, and define the map

$$\psi : D_\psi \subseteq D[I] \to \mathbb{R}, \quad \psi : A \mapsto A^{-1}(u) = \inf \{ x \in I : A(x) \geq u \}. \quad (55)$$
Lemma 7 (van der Vaart & Wellner (1996)). Let $A \in \mathbb{D}_\psi$ satisfy the following two properties.

(i) $A$ is differentiable at a point $\xi_u \in (a_1,a_2)$ such that $A(\xi_u) = u$.

(ii) $A$ has strictly positive derivative at $\xi_u$.

Then, $\psi$ is Hadamard-differentiable at $A$ tangentially to the set of functions $H \in D[I]$ which are continuous at $\xi_u$, with Hadamard derivative given by

$$\psi'_A(H) = \frac{H(\xi_u)}{A'(\xi_u)}.$$ 

Now, define $\mathcal{H} = \mathbb{R} \times \mathbb{S}^{d-1}$, identified with the set of half-spaces in $\mathbb{R}^d$. Let $\mathbb{D}_0$ denote the set of maps $F : \mathcal{H} \to \mathbb{R}$ such that $F(\cdot, \theta) \in C[\mathbb{R}]$ for $\mu$-almost all $\theta \in \mathbb{S}^{d-1}$, and such that $F(x, \cdot)$ is Borel-measurable for all $x \in \mathbb{R}$. Furthermore, define $\mathbb{D}_\theta$ as the set of maps $F : \mathcal{H} \to \mathbb{R}$ such that $F(\cdot, \theta) \in D[I]$ is a CDF for all $\theta \in \mathbb{S}^{d-1}$, and $F(x, \cdot)$ is Borel-measurable for all $x \in \mathbb{R}$. Define the map

$$\phi : \mathbb{D}_\theta^2 \to \mathbb{R}_+,$$

$$(i) \quad A \in \mathbb{D}_\theta^2 = \mathbb{D}_\theta^2,$$  

$$(ii) \quad \phi((F,G)) = \frac{1}{1-2\delta} \int_{\mathbb{S}^{d-1}} \int_0^{1-\delta} |F^{-1}(u,\theta) - G^{-1}(u,\theta)|^r d\mu(\theta),$$

for a fixed constant $\delta \in (0,1/2)$, where we employ the notation $F^{-1}(u,\theta) = F^\theta_1(u) = \inf\{x \in \mathbb{R} : F_\theta(x) \geq u\}$ and $F^\theta_1(u) = F_\theta(\cdot)$ in this section only, to maximize clarity.

The Hadamard differentiability of $\phi$, tangentially to $\mathbb{D}_0$, is established below.

Lemma 8. Let $P,Q \in K(b)$ for some $b > 0$, and assume the measures $\pi_\theta \# P$ and $\pi_\theta \# Q$ are absolutely continuous for $\mu$-almost every $\theta \in \mathbb{S}^{d-1}$, with respective densities $p_\theta$ and $q_\theta$, and respective CDFs $F(\cdot, \theta)$ and $G(\cdot, \theta)$. Then, the map $\phi$ is Hadamard differentiable at $(F,G)$, tangentially to $\mathbb{D}_0$, with Hadamard derivative given by

$$\phi' : \mathbb{D}_\theta^2 \to \mathbb{R},$$

$$(i) \quad 1/t_k \to \infty, \quad \text{and define for all } k \geq 1,$$

$$(ii) \quad A \in \mathbb{D}_\theta^2 = \mathbb{D}_\theta^2,$$

$$\Delta_k = \frac{1}{t_k(1-2\delta)} \int_{\mathbb{S}^{d-1}} \int_0^{1-\delta} \left\{ \left| (F+t_k H_{1k})^{-1} - (G+t_k H_{2k})^{-1} \right|^r - \left| F^{-1} - G^{-1} \right|^r \right\},$$

We will prove that the limit of $\Delta_k$ exists when taking $k \to \infty$. For all $r > 1$, the map $(x,y) \in \mathbb{R}^2 \to |x-y|^r$ is differentiable. Thus, for all $u \in [\delta, 1-\delta]$ and all $\theta \in \mathbb{S}^{d-1}$, there exists $\tilde{F}^{-1}_k(u,\theta)$ (resp. $G^{-1}_k(u,\theta)$) on the line joining $F^{-1}(u,\theta)$ (resp. $G^{-1}(u,\theta)$) and $(F+t_k H_{1k})^{-1}(u,\theta)$ (resp. $(G+t_k H_{2k})^{-1}(u,\theta)$) such that

$$\Delta_k = \frac{1}{t_k(1-2\delta)} \int_{\mathbb{S}^{d-1}} \int_0^{1-\delta} \varphi(\tilde{F}^{-1}_k, G^{-1}_k) \left\{ \left| (F+t_k H_{1k})^{-1} - (G+t_k H_{2k})^{-1} \right|^r - \left| F^{-1} - G^{-1} \right|^r \right\},$$

(56)

52
where \( \varphi(x,y) = r \operatorname{sgn}(x-y)|x-y|^{r-1} \). We will now argue that each of the limits

\[
B_1(u, \theta) = \lim_{k \to \infty} \varphi\left( \tilde{F}_k^{-1}(u, \theta); \tilde{G}_k^{-1}(u, \theta) \right),
\]

\[
B_2(u, \theta) = \lim_{k \to \infty} \frac{(F + t_k H_{1k})^{-1}(u, \theta) - F^{-1}(u, \theta)}{t_k}, \quad B_3(u, \theta) = \lim_{k \to \infty} \frac{(G + t_k H_{2k})^{-1}(u, \theta) - G^{-1}(u, \theta)}{t_k},
\]

exists for \((\lambda \otimes \mu)\)-almost every \((u, \theta) \in [\delta, 1 - \delta] \times \mathbb{S}^{d-1}\).

Regarding \(B_1\), notice that the map \( \varphi \) is continuous in both of its arguments. By the definitions of \( \tilde{F}_k^{-1}, \tilde{G}_k^{-1} \), it thus suffices to argue that \((F + t_k H_{1k})^{-1}(u, \theta) \to F^{-1}(u, \theta)\), and \((G + t_k H_{2k})^{-1}(u, \theta) \to G^{-1}(u, \theta)\) for almost every \((u, \theta)\). To this end, we reason similarly as in the proof of Lemma 7. Given a sequence \( \epsilon_k = o(t_k) \), the definition of quantile implies for all \( u \in [\delta, 1 - \delta] \),

\[
(F + t_k H_{1k})((F + t_k H_{1k})^{-1}(u, \theta) - \epsilon_k, \theta) \leq u \leq (F + t_k H_{1k})((F + t_k H_{1k})^{-1}(u, \theta), \theta),
\]

whence, by uniform boundedness of the sequence \((H_{1k})\),

\[
F((F + t_k H_{1k})^{-1}(u, \theta) - \epsilon_k, \theta) + O(t_k) \leq u \leq F((F + t_k H_{1k})^{-1}(u, \theta), \theta) + O(t_k).
\]

Now, for \( \mu \)-almost every \( \theta \in \mathbb{S}^{d-1} \), \( \pi_{\theta \#} P \) is absolutely continuous, implying that for almost every \( u \in [\delta, 1 - \delta] \), the density \( p_{\theta} \) is positive at \( F^{-1}(u, \theta) \). For such \( u \), \( F(\cdot, \theta) \) is strictly monotonic on any interval of the form \((F^{-1}(u, \theta) - \epsilon, F^{-1}(u, \theta) + \epsilon)\), for any \( \epsilon > 0 \), and is hence strictly bounded away from \( u \) away from any such interval. Comparing with the previous display, we deduce that for large enough \( k \), \((F + t_k H_{1k})^{-1}(u, \theta) \geq F^{-1}(u, \theta) - \epsilon \), and \((F + t_k H_{1k})^{-1}(u, \theta) - \epsilon_k \leq F^{-1}(u, \theta) + \epsilon \). It readily follows that \((F + t_k H_{1k})^{-1}(u, \theta) \to F^{-1}(u, \theta)\), and upon applying a similar argument to \((G + t_k H_{2k})^{-1}(u, \theta)\), we obtain

\[
B_1(u, \theta) = \varphi(F^{-1}(u, \theta); G^{-1}(u, \theta)) = \operatorname{sgn}(F^{-1}(u, \theta) - G^{-1}(u, \theta))|F^{-1}(u, \theta) - G^{-1}(u, \theta)|^{r-1},
\]

for \((\lambda \otimes \mu)\)-almost every \((u, \theta)\).

We now turn to the limit \(B_2\). Recall that \( F(\cdot, \theta) \) is continuous for almost every \( \theta \in \mathbb{S}^{d-1} \). Fixing such a choice of \( \theta \), let \( A = F(\cdot, \theta) \). For any fixed \( u \in [\delta, 1 - \delta] \), the existence of \( B_2(u, \theta) \) would be implied by the Hadamard differentiability of the map \( \psi \) in equation (55) at \( A \), tangentially to \( \mathcal{D}_0 \), sufficient conditions for which are given by conditions (i) and (ii) of Lemma 7. In particular, notice that we do not require the Hadamard differentiability of the map \( A \to A^{-1} \), viewed as mapping into a space of functions. Furthermore, we note that Lemma 3 implies the existence of universal constants \( a_1, a_2 \in \mathbb{R} \), not depending on \( F \), such that \( a_1 \leq F^{-1}(\delta) \leq F^{-1}(1 - \delta) \leq a_2 \), thus we may set \( I = [a_1, a_2] \) when applying Lemma 7.

Now, by choice of \( \theta \), condition (i) of Lemma 7 is immediately satisfied by \( A \). Furthermore, notice that almost every point in the set \( A^{-1}(\delta, 1 - \delta) \) is contained in the set of points at which the density \( p_{\theta} \) is nonzero. It follows that \( A \) is differentiable at \( A^{-1}(u) \) for almost every \( u \in [\delta, 1 - \delta] \), implying that condition (ii) of Lemma 1 is satisfied for all such \( u \). We deduce from Lemma 1 the limit

\[
B_2(u, \theta) = -\frac{H_1(A^{-1}(u))}{p_{\theta}(A^{-1}(u))} = -\frac{H_1(F^{-1}(u, \theta))}{p_{\theta}(F^{-1}(u, \theta))},
\]
for $(\lambda \otimes \mu)$-almost every $(u, \theta)$. We similarly obtain, almost everywhere,

$$B_3(u, \theta) = \frac{H_2(G^{-1}(u, \theta))}{q_0(G^{-1}(u, \theta))}.$$  

Since $P,Q \in \mathcal{K}(b)$, Lemma 2 implies the integrability of the sequences in the limits $B_j, j = 1, 2, 3$. Taking $k \to \infty$ in equation (56), an application of the Dominated Convergence Theorem is thus valid. Together with the definitions of the $B_j$ the claim follows. □

We now turn to the proof of the main result of this subsection.

**Proof of Theorem 3.** Since the set of half-spaces $\mathcal{H}$ forms a separable Vapnik-Chervonenkis class, it is Donsker, implying that the empirical process $G_{nm} = \sqrt{\frac{nm}{n+m}}(P_n - P_mQ_m - Q)$ converges weakly in $\mathbb{D} = \ell^\infty(\mathcal{H}) \times \ell^\infty(\mathcal{H})$,

$$\sup_{h \in \text{BL}_1(\mathbb{D})} \left| \mathbb{E}[h(G_{nm})] - \mathbb{E}[h(G_{(P,Q)})] \right| \to 0, \quad (57)$$

to a process $G_{(P,Q)} := (\sqrt{\sigma}G_P, \sqrt{1-\sigma}G_Q)$, where $G_P$ and $G_Q$ denote $P$- and $Q$- Brownian bridges respectively, and where we identify the set $\mathcal{H}$ with the set of indicator functions over $\mathcal{H}$. Under this abuse of notation, notice that the process $G_P$ takes the form $G_P(x, \theta) = G \circ F(x, \theta)$ for a standard Brownian Bridge $G$, for all $(x, \theta) \in \mathcal{H}$. By assumption, for almost all $\theta \in \mathbb{S}^{d-1}, F(\cdot, \theta)$ is continuous, and since almost all sample paths of $G$ are continuous, we deduce that almost every sample path of $G_P(\cdot, \theta)$ is also continuous. We deduce that $G_P$ takes values in $\mathbb{D}_0$ almost surely, and similarly for $G_Q$.

Furthermore, Theorem 3.6.3 of *van der Vaart & Wellner (1996)* implies the same conditional limiting distribution for the bootstrap empirical process $G_{nm}^* = \sqrt{\frac{nm}{n+m}}(P_n^* - P_nQ_m^* - Q_m)$,

$$\sup_{h \in \text{BL}_1(\mathbb{D})} \left| \mathbb{E}[h(G_{nm}^*)]X_1, \ldots, X_n, Y_1, \ldots, Y_m \right| - \mathbb{E}[h(G_{(P,Q)})] \to 0, \quad (58)$$

in outer probability, where $h$ ranges over $\text{BL}_1(\mathbb{D})$. Now, viewing $\phi$ as a map over $\mathbb{D}$, the Hadamard differentiability of $\phi$ (Lemma 8) together with equation (57) and the functional delta method (see for instance Theorems 3.9.4 of *van der Vaart & Wellner (1996)*) implies

$$\sup_{h \in \text{BL}_1(\mathbb{R})} \left| \mathbb{E}\left[h\left(\sqrt{\frac{nm}{n+m}}(\phi(P_nQ_m) - \phi(P_nQ_m))\right)\right] - \mathbb{E}[h(\phi(G_{(P,Q)}))] \right| \to 0. \quad (59)$$

Likewise, the delta method for the bootstrap (Theorem 3.9.11 of *van der Vaart & Wellner (1996)*) and equations (57) and (58) imply

$$\sup_{h \in \text{BL}_1(\mathbb{R})} \left| \mathbb{E}\left[h\left(\sqrt{\frac{nm}{n+m}}(\phi(P_nQ_m^*) - \phi(P_nQ_m))\right)\right] X_1, \ldots, X_n, Y_1, \ldots, Y_m \right| - \mathbb{E}[h(\phi'(G_{(P,Q)}))] \to 0, \quad (60)$$

in outer probability. A combination of equations (59) and (60) readily leads to the theorem. □
F Proof of Theorem 4

If \( P \neq Q \), and \( P_\theta \) and \( Q_\theta \) are absolutely continuous for all \( \theta \in \mathbb{S}^{d-1} \), the assumption-light finite-sample interval \( C_{nm}^\dagger \) and the bootstrap interval \( C_{nm}^* \), have asymptotic coverage of at least \( 1 - \alpha \), and so we only focus on the case when one of these conditions fails.

First, let us consider the case when \( P = Q \). In this case, we have that \( \text{SW}_{r,\delta}(P,Q) = 0 \) and so,

\[
\mathbb{P}(\text{SW}_{r,\delta}(P,Q) \notin C_{nm}) \\
\leq \mathbb{P}(\text{SW}_{r,\delta}(P,Q) \notin C_{nm}^*) + \mathbb{P}(\text{SW}_{r,\delta}(P,Q) \notin C_{nm}^*, 0 \in C_{nm}^*) \\
\leq \mathbb{P}(0 \notin C_{nm}^*) \\
\leq \alpha + o(1),
\]

by Proposition 5.

Now suppose there exists a set \( S \subseteq \mathbb{S}^{d-1} \) of positive \( \mu \)-measure such that one of the distributions \( P_\theta \) and \( Q_\theta \) is not absolutely continuous for all \( \theta \in S \). We will assume this property for \( P_\theta \) without loss of generality. Up to modifying \( S \) on a subset of \( \mu \)-measure zero, \( P_\theta \) admits no singular component for all \( \theta \in S \) by assumption, thus the Lebesgue Decomposition Theorem implies that \( P_\theta \) admits a nonzero atomic component. Since \( S \) has positive \( \mu \) measure, it follows that \( P \) itself admits a nonzero atomic component, thus, there exists \( a \in \mathbb{R}^d \) such that \( P(\{a\}) = \epsilon > 0 \) for all \( \theta \in S \). We deduce,

\[
\mathbb{P}(D_n > 0) = \prod_{i \neq j} [1 - \mathbb{P}(X_i = X_j)] \leq \prod_{i \neq j} [1 - \mathbb{P}(X_i = a)\mathbb{P}(X_j = a)] = (1 - \epsilon)^2 \frac{n(n-1)}{2}.
\]

Thus, \( \mathbb{P}(D_n > 0) = o(1) \), and we obtain

\[
\mathbb{P}(\text{SW}_{r,\delta}(P,Q) \notin C_{nm}) \leq \mathbb{P}(\text{SW}_{r,\delta}(P,Q) \notin C_{nm}^*, D_n > 0) + \mathbb{P}(\text{SW}_{r,\delta}(P,Q) \notin C_{nm}^*, D_n = 0) \\
\leq \mathbb{P}(D_n > 0) + \mathbb{P}(\text{SW}_{r,\delta}(P,Q) \notin C_{nm}^*) \\
\leq \alpha + o(1).
\]

The claim follows. \( \square \)

G Proofs of Additional Results

G.1 Proof of Proposition 5

Given \( \theta \sim \mu \), we have

\[
W_r^\ast(P_\theta,Q_\theta) = \frac{1}{1 - 2\delta} \int_\delta^{1-\delta} \left[ F_{\theta}^{-1}(u) - G_{\theta}^{-1}(u) \right]^r du \leq \max_{a \in \{\delta,1-\delta\}} \left| F_{\theta}^{-1}(a) \right|^r + \max_{a \in \{\delta,1-\delta\}} \left| G_{\theta}^{-1}(a) \right|^r.
\]

Therefore, it follows from Lemma 3 that

\[
\sup_{P,Q \in K_{2r}(b)} \int_{\mathbb{S}^{d-1}} W_r^2(P_\theta,Q_\theta) d\mu(\theta) \leq B,
\]

55
for some universal constant \( B > 0 \) depending only on \( b, \delta, r \). The claim now follows by the Central Limit Theorem.

\[ \square \]

\[ \text{G.2 Proof of Corollary 1} \]

The proof is a straightforward consequence of Lemma 1.

\[ \square \]

\[ \text{G.3 Proof of Corollary 2} \]

By Theorem 2, we have with probability at least \( 1 - \epsilon \),

\[
\lambda(C_{nm}) \leq \left\{ \text{SW}_{r, \delta}(P,Q) + c\left( \psi_{\epsilon, nm} + \varphi_{\epsilon, nm} + \kappa_N \right) \right\}^{1/r} - \text{SW}_{r, \delta}(P,Q)
\]

\[
= \text{SW}_{r, \delta}(P,Q) \left\{ \frac{1 + c}{\text{SW}_{r, \delta}(P,Q)} \left( \psi_{\epsilon, nm} + \varphi_{\epsilon, nm} + \kappa_N \right) \right\}^{1/r} - 1
\]

\[
\leq \text{SW}_{r, \delta}(P,Q) \left\{ \left[ 1 + \frac{c}{\Gamma} \left( \psi_{\epsilon, nm} + \varphi_{\epsilon, nm} + \kappa_N \right) \right]^{1/r} - 1 \right\}
\]

\[
\lesssim \frac{\text{SW}_{r, \delta}(P,Q)}{\Gamma} \left( \psi_{\epsilon, nm} + \varphi_{\epsilon, nm} + \kappa_N \right)(1 + o(1)),
\]

by a first-order Taylor expansion. The claim follows.

\[ \square \]

\[ \text{G.4 Proof of Proposition 6} \]

The proof is straightforward. We have,

\[
P\left( \theta_0 \notin C_{nm} \right) \leq P\left( \ell_{nm}(\theta_0) > \epsilon \right) \leq P\left( \ell_{nm}(\theta_0) > \epsilon_0 \right) = P\left( \ell_{nm}(\theta_0) > \text{SW}_{r, \delta}(P,P_0) \right).
\]

The claim now follows from Proposition 5.

\[ \square \]

\[ \text{G.5 Proof of Example 2} \]

We begin by proving the validity of the inequality in equation (27). Let \( \mathcal{A} \) be a collection of sets, and let \( S_A(n) \) denote the shattering number (Vapnik 2013) of \( \mathcal{A} \). The relative VC inequality is then given by

\[
P\left( \sup_{A \in \mathcal{A}} \frac{|P_n(A) - P(A)|}{\sqrt{P_n(A)}} \geq t \right) \leq 4S_A(2n)e^{-nt^2/4}, \quad t > 0.
\]
Letting $A=\{(-\infty,x] : x \in \mathbb{R}\}$ and $A=[x,\infty) : x \in \mathbb{R}\}$ respectively, we obtain

$$
P\left( \sup_{x \in \mathbb{R}} \frac{|F_n(x) - F(x)|}{\sqrt{F_n(x)}} \geq t \right) \leq 4(2n+1)e^{-nt^2/4},$$

$$
P\left( \sup_{x \in \mathbb{R}} \frac{|F_n(x) - F(x)|}{\sqrt{1-F_n(x)}} \geq t \right) \leq 4(2n+1)e^{-nt^2/4},$$

for all $t > 0$. By a union bound and the fact that $u(1-u) \geq \frac{1}{2}(u \wedge (1-u))$ for all $u \in [0,1]$, we arrive at

$$
P\left( \sup_{x \in \mathbb{R}} \frac{|F_n(x) - F(x)|}{\sqrt{F_n(x)(1-F_n(x))}} \geq t \right) \leq 8(2n+1)e^{-nt^2/16}.$$  

Setting $t = \nu_{\alpha,n} := \sqrt{\frac{16}{n} \log(16/\alpha) + \log(2n+1)}$, we deduce that with probability at least $1 - \alpha/2$,

$$|F_n(x) - F(x)| \leq \nu_{\alpha,n} \sqrt{F_n(x)(1-F_n(x))}, \quad \forall x \in \mathbb{R}. \quad (61)$$

This proves the validity of equation (27).

We now invert equation (61) to obtain the functions $\gamma_{\alpha,n}$ and $\eta_{\alpha,n}$ which lead to a quantile confidence band. We will require the following definitions of lower CDF and upper quantile function,

$$F(x) := \lim_{y \to x^-} F(x) = \mathbb{P}(X_1 \leq x), \quad F^{-1}(u) = \inf \{ x \in \mathbb{R} : F(x) > u \},$$

with empirical analogues given by

$$F_n(x) := \lim_{y \to x^-} F_n(y) = \frac{1}{n} \sum_{i=1}^{n} I(X_i < x), \quad F^{-1}_n(u) = \inf \{ x \in \mathbb{R} : F_n(x) > u \}.$$  

Notice that $F$ and $F^{-1}$ are right continuous, whereas $\overline{F}$ and $\overline{F}^{-1}$ are left continuous. Furthermore, we make use of the following elementary inequalities relating quantile functions and CDFs,

$$F_n(x) \geq u \iff x \geq F^{-1}_n(u), \quad (62)$$

$$F_n(x) \leq u \iff x \leq F^{-1}_n(u), \quad (63)$$

$$F(x) \geq u \iff x \geq F^{-1}(u), \quad (64)$$

$$F(x) \leq u \iff x \leq F^{-1}(u). \quad (65)$$

We now turn to the proof. The calculations which follow are elementary, but tedious. Let $v = F(x)$. By
We now turn to an upper confidence bound on

\[ F_n(x) + \nu_{\alpha,n} \sqrt{F_n(x)(1-F_n(x))} \geq v \]

\[ \implies (F_n(x) - F_n(x)^2)\nu_{\alpha,n}^2 \geq v^2 - 2vF_n(x) + F_n(x)^2 \]

\[ \implies F_n(x)^2(1+\nu_{\alpha,n}^2) - F_n(x)(2v+\nu_{\alpha,n}^2) + v^2 \leq 0 \]

\[ \implies F_n(x) \geq \frac{2v+\nu_{\alpha,n}^2}{2(1+\nu_{\alpha,n}^2)} - \frac{\sqrt{[2v+\nu_{\alpha,n}^2]^2 - 4(1+\nu_{\alpha,n}^2)v^2}}{2(1+\nu_{\alpha,n}^2)} \]

\[ \implies F_n(x) \geq \frac{2v+\nu_{\alpha,n}^2}{2(1+\nu_{\alpha,n}^2)} - \frac{\nu_{\alpha,n}\nu_{\alpha,n}^2 + 4v(1-v)}{2(1+\nu_{\alpha,n}^2)} = \gamma_{\alpha,n}(v) \]

\[ \implies x \geq F_n^{-1}(\gamma_{\alpha,n}(v)) \] (By (62))

\[ \implies x \geq F_n^{-1}(\gamma_{\alpha,n}(F(x))). \]

Now, let \( u \in (0,1) \). Setting \( x = F^{-1}(u) \) and using the fact that \( F \circ F^{-1}(u) \geq u \) by equation (64), the above display implies

\[ F^{-1}(u) \geq F_n^{-1}(\gamma_{\alpha,n}(F \circ F^{-1}(u))) \geq F_n^{-1}(\gamma_{\alpha,n}(u)), \]

uniformly in \( u \in (0,1) \), with probability at least \( 1-\alpha/2 \).

We now turn to an upper confidence bound on \( F^{-1}(u) \). Upon taking limits from the left in equation (61), we obtain

\[ F_n(x) - \nu_{\alpha,n} \sqrt{F_n(x)(1-F_n(x))} \leq F(x) \]

university in \( x \in \mathbb{R} \), on the same event of probability at least \( 1-\alpha/2 \). Thus, letting \( v = F(x) \), we have

\[ F_n(x) - \nu_{\alpha,n} \sqrt{F_n(x)(1-F_n(x))} \leq v \]

\[ \implies \nu_{\alpha,n}^2 F_n(x)(1-F_n(x)) \geq (F_n(x)-v)^2 \]

\[ \implies \nu_{\alpha,n}^2 F_n(x) - \nu_{\alpha,n}^2 F_n(x)^2 \geq F_n(x)^2 - 2vF_n(x) + v^2 \]

\[ \implies F_n(x)^2(1-\nu_{\alpha,n}^2) - (\nu_{\alpha,n}^2 + 2v)F_n(x) + v^2 \leq 0 \]

\[ \implies F_n(x) \leq \frac{\nu_{\alpha,n}^2 + 2v}{2(1-\nu_{\alpha,n}^2)} - \frac{\sqrt{[\nu_{\alpha,n}^2 + 2v]^2 - 4(1-\nu_{\alpha,n}^2)v^2}}{2(1-\nu_{\alpha,n}^2)} = \eta_{\alpha,n}(v) \]

\[ \implies x \leq F_n^{-1}(\eta_{\alpha,n}(v)) \]

\[ \implies x \leq F_n^{-1}(\eta_{\alpha,n}(F(x))). \] (By (63))

Therefore, setting \( x = F^{-1}(u) \) for \( u \in (0,1) \), and using the fact that \( F \circ F^{-1}(u) \leq u \) by equation (65), we obtain

\[ F^{-1}(u) \leq F_n^{-1}(\eta_{\alpha,n}(F(F^{-1}(u)))) \leq F_n^{-1}(\eta_{\alpha,n}(u)). \]

Upon taking limits from the right, this implies

\[ F^{-1}(u) \leq F_n^{-1}(\eta_{\alpha,n}(u)). \]
uniformly in $u$ with probability at least $1 - \alpha/2$. We conclude that

$$\mathbb{P}\left(F_n^{-1}(\gamma_{\alpha,n}(u)) \leq F^{-1}(u) \leq F_n^{-1}(\eta_{\alpha,n}(u)), \forall u \in (0,1) \right) \geq 1 - \alpha/2.$$ 

The validity of equation (28) follows. \(\square\)

H Interpreting Theorem 2 when $\text{SJ}_{r,\delta}(P) < \infty$ but $\text{SJ}_{r,\delta}(Q) = \infty$

We now discuss the situation where $\text{SJ}_{r,\delta}(P) \leq s$ for a fixed constant $s > 0$, but $\text{SJ}_{r,\delta}(Q) = \infty$. We focus on the one-dimensional case for simplicity, and we omit superscripts depending on $N=1$ below.

**Corollary 3.** Let $P,Q \in \mathcal{P}(\mathbb{R})$ be absolutely continuous with respect to the Lebesgue measure, and assume the same conditions as Theorem 2. Assume there exists a fixed constant $s > 0$ such that $J_{r,\delta}(P) \leq s$, and let $\epsilon > 0$. If $W_{r,\delta}(P,Q) \lesssim \kappa_{\epsilon,n}^{-1}$, then with probability at least $1 - \epsilon$, we have

$$\lambda(C^{(N)}) \lesssim \left[\kappa_{\epsilon,n} \wedge m\right]^r.$$

For example, when $\gamma_{\epsilon,n}, \eta_{\epsilon,n}$ are based on the DKW inequality as in Example 1, and $r=2$, Corollary 3 implies the near-parametric rate $n^{-\frac{r}{2}} + m^{-\frac{r}{2}}$, which is strictly improves the $n^{-\frac{r}{2}} + m^{-\frac{r}{2}}$ rate implied by Corollary 1. Moreover, the rate implied by Corollary 3 approaches the parametric rate for large $r$, which we conjecture to be the minimax rate for this scenario.

**Proof of Corollary 3.** Let $P^\delta, Q^\delta$ the $\delta$-trimmings of $P$ and $Q$, in the sense of equation (7). The quantile functions of $P^\delta, Q^\delta$ are given by

$$F_\delta^{-1}(u) = F^{-1}(u(1-2\delta)+\delta), \quad G_\delta^{-1}(u) = G^{-1}(u(1-2\delta)+\delta).$$

Now, write $h = \kappa_{\epsilon,n} \wedge m$ for simplicity, and let $P_{h}^\delta, Q_{h}^\delta$ denote distributions with quantile functions $u \mapsto F_\delta^{-1}(u+h)$ and $u \mapsto G_\delta^{-1}(u+h)$ respectively, which are well defined since $h \leq \delta$ under assumption $A(\delta/2;\alpha)$. Then, is is a straightforward observation that

$$U_{\epsilon,m}(Q) \leq W_{r,\delta}^{-1}(Q^\delta, Q_{h}^\delta).$$

Furthermore, the finiteness of $J_{r,\delta}(P)$ implies $W_{\infty}(P^\delta, P_{h}^\delta) \lesssim h$, and we have

$$W_{r}(P_{h}^\delta, Q_{h}^\delta) = W_{r}(P^\delta, Q^\delta) + \int_{\delta+h}^{1-\delta+h} |F^{-1} - G^{-1}|^r - \int_{\delta}^{1-\delta} |F^{-1} - G^{-1}|^r \lesssim W_{r}(P^\delta, Q^\delta) + h.$$
Combining these facts, together with equation \(6\) and the bound \(J_{r,\delta}(P) \leq s\), we arrive at

\[
U_{\varepsilon,m}^r(Q) = W_\infty(Q^\delta, Q_h^\delta)
\leq W_\infty(Q^\delta, P^\delta) + W_\infty(P^\delta, P_h^\delta) + W_\infty(P_h^\delta, Q_h^\delta)
\lesssim W_{r,\delta}^r(P^\delta, Q^\delta) + h + W_{r,\delta}^r(P_h^\delta, Q_h^\delta)
\lesssim W_{r,\delta}^r(P^\delta, Q^\delta) + h^{\frac{r-1}{r+1}}
\lesssim W_{r,\delta}^r(P, Q) + h^{\frac{r-1}{r+1}}.
\]

Thus, by Theorem 2, and under the assumption \(W_{r,\delta}(P, Q) \lesssim h^{\frac{r-1}{r+1}}\), we have with probability at least \(1 - \epsilon\),

\[
\lambda(C_{nm}) \lesssim \left\{ W_{r,\delta}^r(P, Q) + h \left( W_{r,\delta}^{-1}(P, Q) + U_{\varepsilon,n}(P) + U_{\varepsilon,m}(Q) \right) \right\}^{1/r} - W_{r,\delta}(P, Q)
\lesssim \left\{ W_{r,\delta}^r(P, Q) + h \left( W_{r,\delta}^{-1}(P, Q) + h^{-1} + W_{r,\delta}^{\frac{r-1}{r+1}}(P, Q) + h^{\frac{r(r-1)}{r+1}} \right) \right\}^{1/r} - W_{r,\delta}(P, Q)
\lesssim W_{r,\delta}(P, Q) \left\{ 1 + h \left( W_{r,\delta}^{\frac{r-1}{r+1}}(P, Q) + W_{r,\delta}^{\frac{r-1}{r+1}}(P, Q) + h^{\frac{r(r-1)}{r+1}} W_{r,\delta}^{\frac{r-1}{r+1}}(P, Q) \right) \right\}^{1/r} - 1
\asymp W_{r,\delta}(P, Q) \left( \frac{h^{\frac{r+1}{r+1}}}{W_{r,\delta}^{\frac{r-1}{r+1}}(P, Q)} \right)^{1/r}
\asymp h^{\frac{r+1}{r(r+1)}}.
\]

The claim follows. \(\square\)