

Minimax confidence intervals for the Sliced Wasserstein distance*

Tudor Manole, Sivaraman Balakrishnan, Larry Wasserman

*Department of Statistics and Data Science,
Carnegie Mellon University*

e-mail: tmanole@andrew.cmu.edu; siva@stat.cmu.edu; larry@stat.cmu.edu

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.

MSC2020 subject classifications: Primary 62G15; Secondary 62G05, 62C20.

Keywords and phrases: Optimal transport, Sliced Wasserstein distance, nonparametric inference, minimax lower bound, likelihood-free inference.

Received April 2021.

1. Introduction

The Wasserstein distance is a metric between probability distributions which has received a surge of interest in statistics and machine learning (Panaretos and 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

arXiv: 1909.07862

*This work was supported in part by the National Science Foundation grants DMS-1713003, DMS-2113684 and CIF-1763734. TM acknowledges the support of the Natural Sciences and Engineering Research Council of Canada (NSERC), through a PGS D scholarship. SB was supported by a Google Research Scholar Award and an Amazon Research Award.

admitting at least $r \geq 1$ moments, with support in $\mathbb{R}^d$ for $d \geq 1$, the r -th order Wasserstein distance between P and Q is defined by

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

where $\Pi(P, Q)$ denotes the set of joint probability distributions with marginals P and Q , known as couplings. Any minimizer γ 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 $\mathbb{R}^d$.

The Wasserstein distance has broadly served two uses in the statistics literature (see the review article of Panaretos and Zemel (2019b) and references therein). On the one hand, it has been used as a theoretical tool for asymptotic theory (Shorack and Wellner, 2009; Shao and Tu, 2012), since convergence in the r -Wasserstein distance is equivalent to weak convergence of probability measures and their r -th moments (Villani, 2003). Wasserstein distances also play a prominent role in the analysis of mixture models (Nguyen, 2013; Ho, Yang and Jordan, 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 ℓ_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); del Barrio, Giné and Utzet (2005); Courty et al. (2016); Ramdas, Trillos and Cuturi (2017); Arjovsky, Chintala and Bottou (2017); Ho et al. (2017); Bernton et al. (2019a,b); Verdinelli and Wasserman (2019) and references therein.

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 and 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 a generally 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 (9). This robustification of the Sliced Wasserstein distance compares distributions up to a 2δ fraction

of their probability mass, thereby generalizing the one-dimensional trimmed Wasserstein distance introduced by Munk and Czado (1998). 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 $X_1, \dots, X_n \sim P$ and $Y_1, \dots, Y_m \sim Q$, we derive confidence sets $C_{nm} \subseteq \mathbb{R}$ such that

$$\inf_{P, Q} \mathbb{P}(\text{SW}_{r, \delta}(P, Q) \in C_{nm}) \geq 1 - \alpha, \quad (2)$$

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 and Czado, 1998; Freitag, Munk and Vogt, 2003; Freitag, Czado and Munk, 2007; Freitag and 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 $\text{SJ}_{r, \delta}$ introduced formally in Section 3.1 (equation (15)). The magnitude of $\text{SJ}_{r, \delta}(P)$ 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 and Ledoux (2019) who showed that when this functional is finite, the empirical measure of $X_1, \dots, X_n$ converges to P under the Wasserstein distance at the fast rate of $O(1/\sqrt{n})$, assuming $d = 1$. On the other hand, when this functional is infinite Bobkov and 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 $\text{SJ}_{r, \delta}$ functional in inference is more nuanced. When $\text{SJ}_{r, \delta}(P)$ and $\text{SJ}_{r, \delta}(Q)$ are finite, 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 and Ledoux (2019). On the other hand, when these values are infinite, a dichotomy arises: in full generality, when $\text{SW}_{r, \delta}(P, Q)$ 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 $\text{SJ}_{r, \delta}$ functional is infinite, accurate $O(1/\sqrt{n \wedge m})$ -inference is possible so long as $\text{SW}_{r, \delta}(P, Q)$ is bounded away from 0. We emphasize that the intervals we construct are *adaptive*, i.e. they have small lengths under appropriate conditions on the $\text{SJ}_{r, \delta}$ functional and $\text{SW}_{r, \delta}(P, Q)$, 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 $\text{SJ}_{r, \delta}(P)$.

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 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, Fan and Beaumont, 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 and Czado, 1998; Freitag, Munk and Vogt, 2003; Freitag, Czado and Munk, 2007; Freitag and Munk, 2005) are typically unverifiable in such applications.

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

- We define the δ -trimmed Sliced Wasserstein distance $\text{SW}_{r,\delta}$, and the functional $\text{SJ}_{r,\delta}$, generalizing the functional J_r of Bobkov and Ledoux (2019). We show that the finiteness of $\text{SJ}_{r,\delta}(P)$ is a sufficient condition for the empirical measure to estimate P at the parametric rate under the trimmed Sliced Wasserstein distance, and we prove corresponding minimax lower bounds. We also derive minimax rates of estimating the Sliced Wasserstein distance between two distributions, both in the trimmed and untrimmed settings. These rates are sensitive to the magnitude of the $\text{SJ}_{r,\delta}$ functional.
- We propose two-sample confidence intervals for $\text{SW}_{r,\delta}(P, Q)$ which have finite-sample coverage under minimal moment assumptions. We bound the length of our confidence intervals, showing that they are adaptive both to the magnitude of $\text{SJ}_{r,\delta}(P)$, $\text{SJ}_{r,\delta}(Q)$ and to whether or not $P = Q$. These lengths achieve the minimax rate of estimating the Sliced Wasserstein distance, up to polylogarithmic factors.
- We further contrast our finite-sample confidence intervals with asymptotic methods. In particular, under certain regularity conditions, we derive limit laws and show that the bootstrap is consistent in estimating the distribution of the empirical r -Sliced Wasserstein distance for all $r > 1$, whenever $P \neq Q$. We then show how this last assumption may be removed by combining 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.

Notation In what follows, given a vector $x = (x_1, \dots, x_d) \in \mathbb{R}^d$, $\|x\| = (\sum_{i=1}^d x_i^2)^{1/2}$ denotes the ℓ_2 norm of x . For any $a, b \in \mathbb{R}$, $a \vee b$ denotes the max-

imum of a and b , and $a \wedge b$ 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 universal constant $C > 0$ such that $a_n \leq Cb_n$, and we write $a_n \asymp b_n$ if $a_n \lesssim b_n \lesssim a_n$. When the constant C depends on another numerical constant $x \in \mathbb{R}$, we either state this dependence explicitly, or write $a_n \lesssim_x b_n$ and $a_n \asymp_x b_n$. δ_x denotes the Dirac delta measure at a point $x \in \mathbb{R}^d$. The Lebesgue measure on $\mathbb{R}^k$, for an integer $k \geq 1$ to be understood from context, is denoted λ . Given a map $T : \mathbb{R}^d \rightarrow \mathbb{R}^d$ 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 $\mathcal{X} \subseteq \mathbb{R}^d$. For all $r \geq 1$, let $\mathcal{P}_r(\mathcal{X})$ denote the subset of measures in $\mathcal{P}(\mathcal{X})$ admitting finite r -th moment.

The one-dimensional Wasserstein distance The infimum in the definition of the Wasserstein distance (1) is always achieved in the setting of this paper (cf. Theorem 4.1, Villani (2003)). 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 $P, Q \in \mathcal{P}_r(\mathcal{X})$. 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 and Ledoux, 2019)

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

The ∞ -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,

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

exists, and defines a new metric W_∞ 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 $W_r(P, Q) \leq W_\infty(P, Q)$ shows that W_∞ is a stronger metric than W_r for any $r \geq 1$. In fact, it is strictly stronger: for instance, $W_r(\delta_0, (1 - \epsilon)\delta_0 + \epsilon\delta_1) \rightarrow 0$ as $\epsilon \rightarrow 0$ if and only if r is finite. In contrast, the metrics W_r induce the same (weak) topology for all finite r , when $\text{diam}(\mathcal{X}) < \infty$ (Villani, 2003).

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 and Czado (1998) define the δ -trimmed Wasserstein distance (up to rescaling) by

$$W_{r,\delta}(P, Q) = \left(\frac{1}{1 - 2\delta} \int_\delta^{1-\delta} |F^{-1}(u) - G^{-1}(u)|^r du \right)^{\frac{1}{r}}. \quad (5)$$

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 2δ fraction of their tail mass. Specifically, let P^δ denote the distribution with CDF $F^\delta(x) = (F(x) - \delta)/(1 - 2\delta)$, for all $F^{-1}(\delta) \leq x \leq F^{-1}(1 - \delta)$, and similarly for Q^δ . Then, Álvarez-Esteban et al. (2008) note that $W_{r,\delta}(P, Q) = W_r(P^\delta, Q^\delta)$.

In addition, we define the trimmed ∞ -Wasserstein distance by

$$W_{\infty,\delta}(P, Q) := \lim_{r \rightarrow \infty} W_{r,\delta}(P, Q) = \sup_{\delta \leq u \leq 1-\delta} |F^{-1}(u) - G^{-1}(u)|.$$

2.2. The Sliced Wasserstein distance

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

$$\text{SW}_r(P, Q) = \left(\int_{\mathbb{S}^{d-1}} W_r^r(P_\theta, Q_\theta) d\mu(\theta) \right)^{\frac{1}{r}}. \quad (6)$$

Since P_θ and Q_θ are one-dimensional distributions, equation (6) admits the closed form

$$\text{SW}_r(P, Q) = \left(\int_{\mathbb{S}^{d-1}} \int_0^1 |F_\theta^{-1}(u) - G_\theta^{-1}(u)|^r du d\mu(\theta) \right)^{\frac{1}{r}}, \quad (7)$$

where F_θ^{-1} and G_θ^{-1} are the respective quantile functions of P_θ and Q_θ . Both integrals of the above expression can be approximated via Monte Carlo sampling from $\mathbb{S}^{d-1}$ 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 for compactly-supported 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

$$\text{SW}_r^r(P, Q) \leq c_{d,r} W_r^r(P, Q) \leq C_{d,r} M^{r-1/(d+1)} \text{SW}_r^{1/(d+1)}(P, Q), \quad (8)$$

where $C_{d,r} > 0$ is a constant depending on d and r , but not depending on M , and $c_{d,r} = \frac{1}{d} \int_{\mathbb{S}^{d-1}} \|\theta\|_r^r d\mu(\theta)$, 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 .

Though the original Sliced Wasserstein distance of Rabin et al. (2011) was defined in terms of the uniform distribution μ over $\mathbb{S}^{d-1}$, a straightforward adaptation of Proposition 5.12 of Bonnotte (2013) shows that SW_r remains a metric over $\mathcal{P}(\mathbb{R}^d)$ when μ is replaced by any probability measure which is absolutely continuous with respect to the Hausdorff measure on $\mathbb{S}^{d-1}$. We allow μ to be any such measure throughout the sequel.

The trimmed Sliced Wasserstein distance In analogy to the trimmed Wasserstein distance in equation (5), we further define

$$\text{SW}_{r,\delta}(P, Q) = \left(\frac{1}{1-2\delta} \int_{\mathbb{S}^{d-1}} \int_{\delta}^{1-\delta} |F_{\theta}^{-1}(u) - G_{\theta}^{-1}(u)|^r du d\mu(\theta) \right)^{\frac{1}{r}}, \quad (9)$$

for some $\delta \in [0, 1/2)$. We also define the trimmed ∞ -Sliced Wasserstein distance by $\text{SW}_{\infty,\delta}(P, Q) = \int_{\mathbb{S}^{d-1}} W_{\infty,\delta}(P_{\theta}, Q_{\theta}) d\mu(\theta)$, and more generally, we write

$$\text{SW}_{\infty,\delta}^{(r)}(P, Q) = \int_{\mathbb{S}^{d-1}} W_{\infty,\delta}^r(P_{\theta}, Q_{\theta}) d\mu(\theta).$$

When $\delta > 0$, the trimmed-Sliced Wasserstein distance is well-defined and finite for all distributions $P, Q \in \mathcal{P}(\mathbb{R}^d)$, including those admitting fewer than r moments. Nevertheless, it can be easily seen that $\sup_{P, Q \in \mathcal{P}(\mathbb{R}^d)} \text{SW}_{r,\delta}(P, Q) = \infty$. It will be fruitful in our development to impose moment conditions which ensure that the quantity $\text{SW}_{r,\delta}(P, Q)$ is uniformly bounded—one such condition is given in terms of the class

$$\mathcal{K}_{r,\rho}(b) = \left\{ P \in \mathcal{P}(\mathbb{R}^d) : \int_{\mathbb{S}^{d-1}} \mathbb{E}_{X \sim P} [|X^{\top} \theta|^{\rho}]^{\frac{r}{\rho}} d\mu(\theta) \leq b \right\}, \quad b, \rho, r \geq 1. \quad (10)$$

We shall use the special case $\rho = 2$ most often, in which case we drop the subscript ρ and simply write $\mathcal{K}_r(b) := \mathcal{K}_{r,2}(b)$. It follows from Lemma 3 in Appendix A that $\text{SW}_{r,\delta}(P, Q)$ is uniformly bounded over distributions $P, Q \in \mathcal{K}_r(b)$, by a constant depending only on b, r and δ . Notice that if $\bar{b} = b^{\rho/r}$, then $\mathcal{K}_{r,\rho}(b)$ contains the class

$$\bar{\mathcal{K}}_{\rho}(\bar{b}) = \{ P \in \mathcal{P}_{\rho}(\mathbb{R}^d) : \mathbb{E}_{X \sim P} [\|X\|^{\rho}] \leq \bar{b} \}. \quad (11)$$

Finally, we write $\mathcal{K}_{r,\rho} = \bigcup_{b \geq 1} \mathcal{K}_{r,\rho}(b)$, $\mathcal{K}_r = \bigcup_{b \geq 1} \mathcal{K}_r(b)$ and $\bar{\mathcal{K}}_{\rho} = \bigcup_{\bar{b} \geq 1} \bar{\mathcal{K}}_{\rho}(\bar{b})$.

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 and Czado (1998) study limiting distributions of the empirical (plug-in) Wasserstein distance estimator, and Freitag, Munk and Vogt (2003); Freitag, Czado and Munk (2007); Freitag and Munk (2005) establish sufficient conditions for the validity of the bootstrap in estimating the distribution of the empirical second-order trimmed 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 $\text{SJ}_{r,\delta}(P)$, $\text{SJ}_{r,\delta}(Q)$, 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 and Cuturi, 2019; Niles-Weed and Rigollet, 2022), 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—see Kim, Balakrishnan and Wasserman (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 and Munk (2018); Tameling, Sommerfeld and Munk (2019); Klatt, Tameling and Munk (2020); Klatt, Munk and Zemel (2020). For distributions with multidimensional support, Rippl, Munk and Sturm (2016) consider the situation where P and Q only differ by a location-scale transformation. Imaizumi, Ota and Hamaguchi (2019) study the validity of the multiplier bootstrap for estimating the distribution of the plug-in estimator of an approximation of the Wasserstein distance. Central Limit Theorems for empirical Wasserstein distances in general dimension have been established by del Barrio and Loubes (2019); del Barrio, González-Sanz and Loubes (2021), 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); Boissard and Le Gouic (2014); Fournier and Guillin (2015); Bobkov and Ledoux (2019); Weed and Bach (2019); Singh and Póczos (2019); Lei (2020), and references therein) 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 and Rigollet (2022), 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 and Rigollet (2022) by showing that faster rates can be achieved depending on the finiteness of the $\text{SJ}_{r,\delta}$ functional, and on the magnitude of $\text{SW}_{r,\delta}(P, Q)$.

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). In contrast to these methods, both of which employ approximate Bayesian computation, our work provides frequentist coverage guarantees under minimal assumptions.

3. 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

$$\mathcal{R}_{nm} \equiv \mathcal{R}_{nm}(\mathcal{O}; r) = \inf_{\hat{S}_{nm}} \sup_{(P, Q) \in \mathcal{O}} \mathbb{E}_{P^{\otimes n} \otimes Q^{\otimes m}} |\hat{S}_{nm} - \text{SW}_{r,\delta}(P, Q)|, \quad (12)$$

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 $\mathcal{O} \subseteq \mathcal{P}(\mathbb{R}^d) \times \mathcal{P}(\mathbb{R}^d)$ is a collection of pairs of distributions. Our motivation for studying this quantity is the observation that $\mathcal{R}_{nm}$ 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 (12) 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

$$\mathcal{M}_n \equiv \mathcal{M}_n(\mathcal{J}; r) = \inf_{\hat{P}_n} \sup_{P \in \mathcal{J}} \mathbb{E}_{P^{\otimes n}} \left\{ \text{SW}_{r,\delta}(\hat{P}_n, P) \right\}, \quad (13)$$

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 $\mathcal{J} \subseteq \mathcal{P}(\mathbb{R}^d)$. Problems (12) and (13) 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 (13), we have, by the triangle inequality,

$$\begin{aligned} \mathcal{R}_{nm}(\mathcal{O}; r) &\lesssim \mathbb{E} |\text{SW}_{r,\delta}(\hat{P}_n, \hat{Q}_m) - \text{SW}_{r,\delta}(P, Q)| \\ &\leq \mathbb{E} \text{SW}_{r,\delta}(\hat{P}_n, P) + \mathbb{E} \text{SW}_{r,\delta}(\hat{Q}_m, Q) \lesssim \mathcal{M}_{n \wedge m}(\mathcal{J}; r), \end{aligned} \quad (14)$$

for suitable families $\mathcal{J}$ and $\mathcal{O}$ (typically $\mathcal{O} \subseteq \mathcal{J} \times \mathcal{J}$). Inequality (14) implies that estimating a distribution under $\text{SW}_{r,\delta}$ is a more challenging problem, statistically, than that of estimating the Sliced Wasserstein distance between two distributions. It is unclear, however, whether the rate $\mathcal{M}_{n \wedge m}$ is a tight upper

bound on $\mathcal{R}_{nm}$, or whether the latter can be further reduced. For the Wasserstein distance W_r in general dimension, Liang (2019) and Niles-Weed and Rigollet (2022) showed that there is no gap between these minimax risks (ignoring polylogarithmic factors) for compactly supported distributions.

Let us now briefly summarize the main results of this section. We bound $\mathcal{M}_n$ and $\mathcal{R}_{nm}$, and show that there can be a large gap between these minimax risks when the pairs of distributions in $\mathcal{O}$ are appropriately separated. In the special case $d = 1$, $\text{SW}_{r,\delta}$ reduces to the (trimmed) Wasserstein distance, and our results imply faster rates than those of Liang (2019) and Niles-Weed and Rigollet (2022), for estimating the Wasserstein distance between distributions bounded away from each other. Furthermore, in contrast to the minimax risk for estimating the Wasserstein distance and estimating under 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 $\mathcal{R}_{nm}$ (due to its direct connection to confidence intervals) we begin by studying $\mathcal{M}_n$ to motivate our choices of families $\mathcal{O}$. Inspired by Bobkov and Ledoux (2019), in Section 3.1 we define a functional $\text{SJ}_{r,\delta}$, whose magnitude is related to the regularity of the supports of P and Q , and whose finiteness implies improved rates of decay for $\mathcal{M}_n$. We then study the minimax risk $\mathcal{R}_{nm}$ over various families $\mathcal{O}$ in Section 3.2.

3.1. Minimax estimation under the Sliced Wasserstein distance

Let $\delta \in [0, 1/2)$, $P \in \mathcal{P}(\mathbb{R}^d)$, and let $X_1, \dots, X_n \sim P$ 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, extending the comprehensive treatment by Bobkov and Ledoux (2019). We then provide corresponding minimax lower bounds on $\mathcal{M}_n$.

For any $\theta \in \mathbb{S}^{d-1}$, let p_θ denote the density of the absolutely continuous component in the Lebesgue decomposition of the measure $P_\theta = \pi_{\theta\#}P$. Define the functional

$$\text{SJ}_{r,\delta}(P) = \frac{1}{1-2\delta} \int_{\mathbb{S}^{d-1}} \int_\delta^{1-\delta} \left(\frac{\sqrt{u(1-u)}}{p_\theta(F_\theta^{-1}(u))} \right)^r du d\mu(\theta), \quad (15)$$

with the convention that $0/0 = 0$. When $d = 1$, we write $J_{r,\delta}$ instead of $\text{SJ}_{r,\delta}$, and in the untrimmed case $\delta = 0$, we omit the subscript δ and write SJ_r or J_r . When $d = 1$ and $\delta = 0$, Bobkov and 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 $\text{SJ}_{r,\delta}$ functional with respect to trimmed Sliced Wasserstein distances, using distinct proof techniques.

It can be seen that a necessary condition for the finiteness of $\text{SJ}_{r,\delta}(P)$ is that for μ -almost every $\theta \in \mathbb{S}^{d-1}$, the density p_θ is supported on a (possibly

infinite) interval. When δ vanishes, the value of $\text{SJ}_{r,\delta}(P)$ also depends 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 $\text{SJ}_{r,\delta}(P) < \infty$ whenever $\delta > 0$, whereas for $\delta = 0$, $\text{SJ}_r(P) < \infty$ if and only if $1 \leq r < 2$ by a similar argument as Bobkov and 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 $(a, b) \subseteq \mathbb{R}$, one has $\text{SJ}_{r,\delta}(P) < \infty$ if and only if $\Delta_1 = \Delta_2$, for every $\delta \in [0, 1/2)$.

We now provide two upper bounds on $\mathbb{E}[\text{SW}_{r,\delta}(P_n, P)]$, which are effective when $\text{SJ}_{r,\delta}(P) < \infty$ and $\text{SJ}_{r,\delta}(P) = \infty$ respectively. Recall the class $\mathcal{K}_r(b)$ from equation (10).

Proposition 1. *Let $b, r \geq 1$ and $\delta \in (0, 1/2)$. Assume $P \in \mathcal{K}_r(b)$, and that $\delta \geq 2(r+2)/n$.*

(i) *There exist constants $c_r, c'_r > 0$ depending only on r such that*

$$\mathbb{E}[\text{SW}_{r,\delta}(P_n, P)] \leq c_r \text{SJ}_{r, \frac{\delta}{2}}^{\frac{1}{r}}(P) \sqrt{\frac{\log n}{n}} + \frac{c_r (b n e^{-c'_r n \delta})^{\frac{1}{r}}}{\sqrt{\delta}}. \quad (16)$$

(ii) *There exists a constant $k_r > 0$ depending only on r such that*

$$\mathbb{E}[\text{SW}_{r,\delta}(P_n, P)] \leq \frac{k_r}{\sqrt{\delta}} \left(\frac{b}{1-2\delta} \right)^{\frac{1}{r}} n^{-1/2r}. \quad (17)$$

Proposition 1 provides two upper bounds on the rate of convergence of the empirical measure under $\text{SW}_{r,\delta}$, which are closely related to those of Theorem 5.3 and Theorem 7.16 of Bobkov and Ledoux (2019) for the one-dimensional untrimmed Wasserstein distance. Bobkov and Ledoux (2019) established these results by hinging upon a representation of the empirical one-dimensional Wasserstein distance in terms of the so-called mean square beta distribution, coupled with Poincaré-type inequalities for such measures. While extensions of these techniques to the *untrimmed* Sliced Wasserstein distance are straightforward, and will be stated for completeness in Section 3.3, it was unclear to us whether they may be adapted to the *trimmed* setting $\delta > 0$. Our proof of Proposition 1 is instead based on uniform concentration inequalities for the empirical quantile process—an approach which we now describe, and which foreshadows the construction of our confidence intervals in Section 4.

Proposition 1(i) is proved using a uniform bound for self-normalized empirical processes, known as the relative Vapnik-Chervonenkis (VC) inequality (Vapnik, 2013; Bousquet, Boucheron and Lugosi, 2003), which will be further described in Example 2 below. This result shows that the empirical measure converges at the parametric rate under $\text{SW}_{r,\delta}$, up to a polylogarithmic factor, provided $\text{SJ}_{r,\delta/2}(P)$ is bounded, and provided, for instance, that the trimming constant δ does not vanish at a rate faster than $n^{-\beta}$ for some $\beta \in (0, 1)$. We emphasize

that this convergence is uniform in P —for instance, one has that for all $s \geq 1$,

$$\sup_{\substack{P \in \mathcal{K}_r(b) \\ \text{SJ}_{r,\delta/2}(P) \leq s}} \mathbb{E}[\text{SW}_{r,\delta}(P_n, P)] \lesssim_{b,r} s^{\frac{1}{r}} \sqrt{\frac{\log n}{n}}, \text{ when } \delta \asymp n^{-\beta}, \text{ for some } \beta \in (0, 1).$$

In fact, the above bound continues to hold when $s \leq 1$ —a regime relevant for distributions with vanishing variances—so long as $s^{1/r} \sqrt{\log n/n}$ remains greater than the second term in equation (16). Finally, we note that the polylogarithmic factor in the above display arises from the relative VC inequality. Example 2.7 of Giné and Koltchinskii (2006) suggests that this factor may be improved to $\sqrt{\log \log n}$, but not further if $\delta \gtrsim 1/n$. We do not know whether this factor can be omitted if stronger conditions are placed on δ while allowing it to vanish.

Proposition 1(ii) is primarily of interest for distributions such that $\text{SJ}_{r,\delta}(P) = \infty$, and is proved using the Dvoretzky-Kiefer-Wolfowitz inequality (Dvoretzky, Kiefer and Wolfowitz, 1956; Massart, 1990). This result shows that the empirical measure converges to P at the nonparametric rate $n^{-1/2r}$ under no assumptions on P apart from the mild moment assumption $P \in \mathcal{K}_r(b)$. In contrast to Proposition 1(i), however, this result suffers from a markedly worse dependence on δ . Indeed, the resulting rate of convergence deteriorates as soon as $\delta = o(1)$, and we conjecture that this behaviour is necessary under the stated assumptions.

The rates in Proposition 1 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 and Guillin (2015), Weed and Bach (2019)) shows that $\mathbb{E}[W_r(P_n, P)] \lesssim n^{-1/d}$ whenever $d > 2r$, and this rate is known to be tight (Dudley, 1969; Singh and Póczos, 2019). Thus, 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 (8).

For completeness, we close this subsection by stating a lower bound on the minimax risk $\mathcal{M}_n$ in equation (13). In view of Proposition 1, it is natural to carry out our analysis over the class of distributions

$$\mathcal{J}(s) = \{P \in \mathcal{K}_r(b) : \text{SJ}_{r,\delta}(P) \leq s\}, \quad s \in [0, \infty].$$

Proposition 2. *Let $b, r \geq 1$ and $\delta \in (0, 1/2)$. Then, there exist constants $C_1, C_2 > 0$, possibly depending on b, r, δ , such that for all $s > 0$ satisfying $b \geq (2s)^{1/r}$,*

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

Proposition 2 implies that the rates achieved by the empirical measure in Proposition 1 are minimax optimal over the classes considered above, up to polylogarithmic factors. The proof of this result will follow as a special case of our bounds on the minimax risk $\mathcal{R}_{nm}$, to which we turn our attention next.

3.2. Minimax estimation of the Sliced Wasserstein distance

In this section, we bound the minimax risk $\mathcal{R}_{nm}$ defined in equation (12). 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 $X_1, \dots, X_n \sim P$ and $Y_1, \dots, Y_m \sim Q$, respectively.

Proposition 3. *Let $b, r \geq 1$ and $\delta \in (0, 1/2)$. Assume $P, Q \in \mathcal{K}_r(b)$, and that $\delta \geq 2(r+2)/(n \wedge m)$.*

(i) *There exists a constant $c_r > 0$, possibly depending on r , such that*

$$\begin{aligned} & \mathbb{E}|\text{SW}_{r,\delta}(P_n, Q_m) - \text{SW}_{r,\delta}(P, Q)| \\ & \lesssim_r \left(\frac{b}{1-2\delta} \right)^{\frac{1}{r}} \frac{n^{-\frac{1}{2r}}}{\sqrt{\delta}} \wedge \left(\text{SJ}_{r,\frac{\delta}{2}}^{\frac{1}{r}}(P) \sqrt{\frac{\log n}{n}} + \frac{(bne^{-c_r n \delta})^{\frac{1}{r}}}{\sqrt{\delta}} \right) \\ & \quad + \left(\frac{b}{1-2\delta} \right)^{\frac{1}{r}} \frac{m^{-\frac{1}{2r}}}{\sqrt{\delta}} \wedge \left(\text{SJ}_{r,\frac{\delta}{2}}^{\frac{1}{r}}(Q) \sqrt{\frac{\log m}{m}} + \frac{(bme^{-c_r m \delta})^{\frac{1}{r}}}{\sqrt{\delta}} \right). \end{aligned}$$

(ii) *Suppose $\text{SW}_{r,\delta}(P, Q) \geq \Gamma$, for some real number $\Gamma > 0$. Then,*

$$\mathbb{E}|\text{SW}_{r,\delta}(P_n, Q_m) - \text{SW}_{r,\delta}(P, Q)| \lesssim_{\Gamma,r} \frac{b}{\delta^{r/2}(1-2\delta)} \left(n^{-\frac{1}{2}} + m^{-\frac{1}{2}} \right).$$

Proposition 3(i) is an immediate consequence of inequality (14), 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 measures under $\text{SW}_{r,\delta}$ given in Proposition 1. In particular, these results show that the parametric rate for estimating $\text{SW}_{r,\delta}$ is achievable for distributions satisfying $\text{SJ}_{r,\delta/2}(P), \text{SJ}_{r,\delta/2}(Q) < \infty$, while the rate $n^{-1/2r} + m^{-1/2r}$ is otherwise achievable. 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 under $\text{SW}_{r,\delta}$. This fast rate of convergence is obtained irrespective of the values of $\text{SJ}_{r,\delta}(P)$ and $\text{SJ}_{r,\delta}(Q)$. Discrepancies between rates of convergence at the null ($P = Q$) and away from the null ($P \neq Q$) have previously been noted by Sommerfeld and Munk (2018) for Wasserstein distances over finite spaces—indeed, their rates match those of Proposition 3 when $\text{SJ}_{r,\delta/2}(P), \text{SJ}_{r,\delta/2}(Q) = \infty$. Finally, we note that the natural estimator $\text{SW}_{r,\delta}(P_n, Q_m)$ is *adaptive* to the typically unknown quantities $\text{SJ}_{r,\delta/2}(P)$ and $\text{SJ}_{r,\delta/2}(Q)$, 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 $\text{SJ}_{r,\delta/2}$ functionals are finite, or when P and Q are sufficiently well-separated.

We now provide corresponding lower bounds on the minimax risk $\mathcal{R}_{nm}$. Inspired by Proposition 3, we define the following collection of pairs of distributions,

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

where $s_1, s_2 \in [0, \infty]$ and $\Gamma \geq 0$. To ensure that the class $\mathcal{O}(\Gamma; s_1, s_2)$ is nonempty, we assume in what follows that $\Gamma^r \leq c_r b$, for some sufficiently small constant $c_r > 0$ depending only on r . With these definitions in place we now state our minimax lower bounds on the risk $\mathcal{R}_{nm}$.

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

- (i) *There exists a constant $C_1 > 0$, possibly depending on δ, r, b , such that for any $s_1, s_2 \in [0, \infty]$,*

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

- (ii) *For any $\Gamma > 0$ such that $\Gamma^r \leq c_r b$, there exists a constant $C_2 > 0$ possibly depending on δ, r, b, Γ such that*

$$\mathcal{R}_{nm}(\mathcal{O}(\Gamma; \infty, \infty); r) \geq C_2 \left(n^{-\frac{1}{2}} + m^{-\frac{1}{2}} \right).$$

Theorem 1 implies that the rates achieved by the empirical Sliced Wasserstein distance $\text{SW}_r(P_n, Q_m)$ in Proposition 3, including their dependence on the $\text{SJ}_{r,\delta}$ functional, are minimax optimal (ignoring polylogarithmic factors). We defer its proof to Appendix C. This result is proved by a standard information-theoretic technique of constructing pairs of distributions which are statistically indistinguishable but have very different Sliced Wasserstein distances. We then obtain lower bounds via an application of Le Cam's Lemma (see, for instance, Theorem 2.2 of Tsybakov (2008)). Beyond this careful choice of distributions, the bulk of our technical effort lies in tightly bounding the various Sliced Wasserstein distances (see Lemma 10 in Appendix C).

In Section 4, 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. Before turning to these results, we discuss estimation rates in the untrimmed case when $\delta = 0$.

3.3. Minimax estimation of the untrimmed Sliced Wasserstein distance

Though our results below on finite-sample and asymptotic inference will focus on the trimmed Sliced Wasserstein distance, as it is an estimand for which assumption-free inference is possible, we end this section by deriving convergence rates for estimating the untrimmed Sliced Wasserstein distance. In this setting, a straightforward extension of Theorem 5.3 and Theorem 7.16 of Bobkov and Ledoux (2019) already leads to the following untrimmed analogue of Proposition 1, which we state for completeness. We recall that the class $\mathcal{K}_{r,\rho}(b)$ is defined in equation (10).

Proposition 4. *Let $r \geq 1$ and $s > 0$. Then,*

$$\sup_{\substack{P \in \mathcal{P}(\mathbb{R}^d) \\ \text{SJ}_r(P) \leq s}} \mathbb{E} \text{SW}_r(P_n, P) \lesssim_r s^{\frac{1}{r}} n^{-\frac{1}{2}}.$$

Furthermore, for any $\rho > 2r$ and $b > 0$,

$$\sup_{P \in \mathcal{K}_{r,\rho}(b)} \mathbb{E} \text{SW}_r(P_n, P) \lesssim_{b,\rho,r} n^{-\frac{1}{2r}}.$$

Convergence rates for estimating $\text{SW}_r(P, Q)$ immediately follow from Proposition 4. For example, we obtain

$$\sup_{P, Q \in \mathcal{K}_{r,\rho}(b)} \mathbb{E} |\text{SW}_r(P_n, Q_m) - \text{SW}_r(P, Q)| \lesssim_{b,\rho,r} n^{-\frac{1}{2r}} + m^{-\frac{1}{2r}}.$$

By analogy with Proposition 3(ii), it is natural to ask whether the above convergence rate may be improved to the parametric rate when P and Q are separated in Sliced Wasserstein distance. Such an assertion cannot be deduced from the work of Bobkov and Ledoux (2019), and is the subject of the following main result.

Theorem 2. *For any $r, b \geq 1$, $\Gamma > 0$, and any $\rho > 2r$, it holds that*

$$\sup_{\substack{P, Q \in \mathcal{K}_{r,\rho}(b) \\ \text{SW}_r(P, Q) \geq \Gamma}} \mathbb{E} |\text{SW}_r(P_n, Q_m) - \text{SW}_r(P, Q)| \lesssim_{\Gamma,\rho,r} b \left(\sqrt{\frac{\log n}{n}} + \sqrt{\frac{\log m}{m}} \right). \quad (18)$$

Theorem 2 proves that the parametric rate for estimating the Sliced Wasserstein distance between well-separated distributions continues to hold in the absence of trimming, at the price of a polylogarithmic factor. In fact, our proof shows more generally that the following bound holds without separation conditions on P and Q ,

$$\sup_{P, Q \in \mathcal{K}_{r,\rho}(b)} \mathbb{E} |\text{SW}_r^r(P_n, Q_m) - \text{SW}_r^r(P, Q)| \lesssim_{\rho,r} b \left(\sqrt{\frac{\log n}{n}} + \sqrt{\frac{\log m}{m}} \right). \quad (19)$$

The only regularity condition required for these bounds is that $P, Q \in \mathcal{K}_{r,\rho}(b)$, for some $\rho > 2r$. When $d = 1$, this condition is equivalent to assuming that P and Q have finite moments of order ρ , and is otherwise weaker when $d > 1$. The threshold $\rho > 2r$ appears to be nearly sharp, at least for the conclusion of equation (19) to hold. One clearly requires $\rho \geq r$, as otherwise $\text{SW}_r(P, Q)$ may be infinite. In the range $r < \rho < 2r$, for the special case $d = 1$ and $P = Q$, Fournier and Guillin (2015) argue that the rate in equation (19) cannot be improved beyond $n^{-\frac{\rho-r}{\rho}}$, which is polynomially slower than the parametric rate. While we do not know the sharp rate in this regime when $P \neq Q$, we expect that the parametric rate is not achievable even under this restriction, for $\rho < 2r$.

Theorem 2 is proved using a peeling argument, coupled with a uniform self-normalized concentration inequality for the empirical quantile process, which was already discussed following the statement of Proposition 1. Unlike the latter result, where this inequality allowed us to obtain rates which adapt to the $\text{SJ}_{r,\delta}$ functional, its use here is essential for obtaining a nearly sharp rate without unnecessary moment assumptions, as it allows us to tightly control the behaviour of extremal empirical quantiles. We defer the proof to Appendix D.

4. Finite-sample confidence intervals

4.1. Finite-sample confidence intervals in dimension one

Throughout this subsection, let $r \geq 1$ and $\delta \in [0, 1/2)$ be given, let $P, Q \in \mathcal{P}(\mathbb{R})$ be probability distributions with respective CDFs F, G , and let $X_1, \dots, X_n \sim P$ and $Y_1, \dots, Y_m \sim Q$ 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 δ -trimmed Wasserstein distance, with the following non-asymptotic coverage guarantee

$$\inf_{P, Q \in \mathcal{P}(\mathbb{R})} \mathbb{P}(W_{r,\delta}(P, Q) \in C_{nm}) \geq 1 - \alpha, \quad (20)$$

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^r norm of the quantile functions of P and Q (cf. equation (3)), 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

$$\inf_{P \in \mathcal{P}(\mathbb{R})} \mathbb{P}\left(F_n^{-1}(\gamma_{\alpha,n}(u)) \leq F^{-1}(u) \leq F_n^{-1}(\eta_{\alpha,n}(u)), \forall u \in (0, 1)\right) \geq 1 - \alpha/2, \quad (21)$$

for some sequences of functions $\gamma_{\alpha,n}, \eta_{\alpha,n} : (0, 1) \rightarrow \mathbb{R}$. The study of uniform quantile bounds of the form (21) is a classical topic (see for instance, the book of Shorack and 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, Kiefer and Wolfowitz, 1956; Massart, 1990), we have

$$\mathbb{P}\left(|F_n(x) - F(x)| \leq \beta_n, \forall x \in \mathbb{R}\right) \geq 1 - \frac{\alpha}{2}, \quad \beta_n = \sqrt{\frac{1}{2n} \log(4/\alpha)}. \quad (22)$$

Inverting this inequality leads to the choice

$$\gamma_{\alpha,n}(u) = u - \beta_n, \quad \eta_{\alpha,n}(u) = u + \beta_n, \quad u \in (0, 1). \quad (23)$$

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

$$\mathbb{P}\left(|F_n(x) - F(x)| \leq \nu_{\alpha,n} \sqrt{F_n(x)(1 - F_n(x))}, \forall x \in \mathbb{R}\right) \geq 1 - \frac{\alpha}{2}, \quad (24)$$

where $\nu_{\alpha,n} := \sqrt{\frac{16}{n} [\log(16/\alpha) + \log(2n + 1)]}$. As shown in Appendix G.6, inverting inequality (24) leads to the following choice, for all $u \in (0, 1)$,

$$\begin{aligned} \gamma_{\alpha,n}(u) &= \frac{2u + \nu_{\alpha,n}^2 - \nu_{\alpha,n} \sqrt{\nu_{\alpha,n}^2 + 4u(1-u)}}{2(1 + \nu_{\alpha,n}^2)}, \\ \eta_{\alpha,n}(u) &= \frac{2u + \nu_{\alpha,n}^2 + \nu_{\alpha,n} \sqrt{\nu_{\alpha,n}^2 + 4u(1-u)}}{2(1 + \nu_{\alpha,n}^2)}. \end{aligned} \quad (25)$$

Given sequences of functions $\gamma_{\alpha,n}, \eta_{\alpha,n}$ satisfying equation (21), one has with probability at least $1 - \alpha$, $A_{nm}(u) \leq |F^{-1}(u) - G^{-1}(u)| \leq B_{nm}(u)$ uniformly in $u \in [\delta, 1 - \delta]$, where,

$$\begin{aligned} A_{nm}(u) &= [F_n^{-1}(\gamma_{\alpha,n}(u)) - G_m^{-1}(\eta_{\alpha,m}(u))] \vee [G_m^{-1}(\gamma_{\alpha,m}(u)) - F_n^{-1}(\eta_{\alpha,n}(u))] \vee 0, \\ B_{nm}(u) &= [F_n^{-1}(\eta_{\alpha,n}(u)) - G_m^{-1}(\gamma_{\alpha,m}(u))] \vee [G_m^{-1}(\eta_{\alpha,m}(u)) - F_n^{-1}(\gamma_{\alpha,n}(u))]. \end{aligned}$$

This observation readily leads to the following Proposition.

Proposition 5. Let $\delta \in [0, 1/2]$ and $r \geq 1$. Then, the interval

$$C_{nm} = \left[\left(\frac{1}{1 - 2\delta} \int_{\delta}^{1-\delta} A_{nm}^r(u) du \right)^{\frac{1}{r}}, \left(\frac{1}{1 - 2\delta} \int_{\delta}^{1-\delta} B_{nm}^r(u) du \right)^{\frac{1}{r}} \right], \quad (26)$$

satisfies $\inf_{P, Q \in \mathcal{P}(\mathbb{R})} \mathbb{P}(W_{r,\delta}(P, Q) \in C_{nm}) \geq 1 - \alpha$.

Proposition 5 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($\delta; \alpha$) We have $\gamma_{\alpha,n \wedge m}(\delta) > 0$ and $\eta_{\alpha,n \wedge m}(1 - \delta) < 1$.

If $\gamma_{\alpha,n}, \eta_{\alpha,n}$ are chosen via the DKW inequality (23), these inequalities imply the choice $\delta \gtrsim (n \wedge m)^{-1/2}$, while if they are chosen via the relative VC inequality (25), one must take $\delta \gtrsim \log(n \wedge m)(n \wedge m)^{-1}$. These choices exclude 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 . If explicit bounds on the quantile functions of P and Q are known near the boundary of the unit interval—which is for instance the case when an upper bound on the moments of P and Q is known—these may be used to replace the

confidence band $[F_n^{-1}(\gamma_{\alpha,n}(u)), F_n^{-1}(\eta_{\alpha,n}(u))]$ by one of finite length for values of $u \in [0, 1]$ satisfying $\gamma_{\alpha,n}(u) < 0$ and $\eta_{\alpha,n}(u) > 1$. Doing so would lead to a confidence interval C_{nm} of finite length for the untrimmed Wasserstein distance. Since our goal is assumption-free inference, however, we do not pursue this avenue here and we therefore assume **A1**($\delta; \alpha$) holds throughout the sequel.

4.2. Finite-sample confidence intervals in general dimension

We now use Proposition 5 to derive a confidence interval for $\text{SW}_{r,\delta}(P, Q)$, where $P, Q \in \mathcal{P}(\mathbb{R}^d)$. In analogy to Section 4.1, a natural approach is to choose functions $\bar{\gamma}_{\alpha,n}$ and $\bar{\eta}_{\alpha,n}$ such that

$$\mathbb{P}\left(F_{\theta,n}^{-1}(\bar{\gamma}_{\alpha,n}(u)) \leq F_{\theta}^{-1}(u) \leq F_{\theta,n}^{-1}(\bar{\eta}_{\alpha,n}(u)), \forall u \in (0, 1), \theta \in \mathbb{S}^{d-1}\right) \geq 1 - \frac{\alpha}{2}, \quad (27)$$

uniformly in $P \in \mathcal{P}(\mathbb{R}^d)$, where $F_{\theta,n}(x) = (1/n) \sum_{i=1}^n I(X_i^\top \theta \leq x)$ for all $x \in \mathbb{R}$ and $\theta \in \mathbb{S}^{d-1}$, and F_{θ}^{-1} denotes the quantile function of $P_{\theta} = \pi_{\theta\#}P$. Such a bound can be obtained, for instance, by applying the VC inequality (Vapnik, 2013) to the empirical process indexed by the set of half-spaces in $\mathbb{R}^d$. An assumption-free confidence interval for $\text{SW}_{r,\delta}(P, Q)$ with finite-sample coverage may then be constructed by following the same lines as in the previous section. Due to the uniformity of equation (27) 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 μ .

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

$$\text{SW}_{r,\delta}^{(N)}(P, Q) = \left(\int_{\mathbb{S}^{d-1}} W_{r,\delta}^r(P_{\theta}, Q_{\theta}) d\mu_N(\theta) \right)^{\frac{1}{r}} = \left(\frac{1}{N} \sum_{j=1}^N W_{r,\delta}^r(P_{\theta_j}, Q_{\theta_j}) \right)^{\frac{1}{r}}.$$

For any $\theta \in \mathbb{S}^{d-1}$, let $[\ell_{N,nm}(\theta), u_{N,nm}(\theta)]$ be the confidence interval in equation (26) for $W_{r,\delta}(P_{\theta}, Q_{\theta})$, at level $1 - \alpha/N$. Let

$$L_{N,nm} = \int_{\mathbb{S}^{d-1}} \ell_{N,nm}^r(\theta) d\mu_N(\theta), \quad U_{N,nm} = \int_{\mathbb{S}^{d-1}} u_{N,nm}^r(\theta) d\mu_N(\theta),$$

and set

$$C_{nm}^{(N)} = \left[L_{N,nm}^{\frac{1}{r}}, U_{N,nm}^{\frac{1}{r}} \right]. \quad (28)$$

By a Bonferroni correction, we obtain conditional coverage of $\text{SW}_{r,\delta}^{(N)}(P, Q)$, i.e. almost surely,

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

We further obtain finite-sample coverage of the Sliced Wasserstein distance itself by the following small enlargement of $C_{nm}^{(N)}$.

Proposition 6. *Let $b, r \geq 1$ and $\delta \in (0, 1/2)$. Let $(M_N)_{N=1}^\infty$ be a nonnegative sequence such that $M_N \rightarrow \infty$ as $N \rightarrow \infty$. Define*

$$\bar{C}_{nm}^{(N)} = \left[\left(L_{N,nm} - M_N/\sqrt{N} \right)^{\frac{1}{r}}, \left(U_{N,nm} + M_N/\sqrt{N} \right)^{\frac{1}{r}} \right]. \quad (29)$$

Then, there is a constant $c > 0$ depending only on r such that

$$\inf_{P, Q \in \mathcal{K}_{2r}(b)} \mathbb{P} \left(\text{SW}_{r,\delta}(P, Q) \in \bar{C}_{nm}^{(N)} \right) \geq 1 - \alpha - \frac{bc}{M_N^2 \delta^r}.$$

Proposition 6 ensures that an enlargement of the interval $C_{nm}^{(N)}$, of size less than $(M_N/\sqrt{N})^{\frac{1}{r}}$, will cover $\text{SW}_{r,\delta}(P, Q)$ at level $1 - \alpha - O(M_N^{-2})$, for any fixed sample sizes n and m . Notice that N is chosen by the practitioner, so that this enlargement can be made to be of lower order than the length of $C_{nm}^{(N)}$. We shall therefore focus our analysis and numerical studies on the interval $C_{nm}^{(N)}$ rather than $\bar{C}_{nm}^{(N)}$.

Although the coverage of the above intervals requires no assumptions on P and Q , apart from the mild moment condition $P, Q \in \mathcal{K}_{2r}(b)$, we now show that their length achieves the minimax rates established in Theorem 1, and is adaptive to the magnitude of $\text{SJ}_{r,\delta}(P), \text{SJ}_{r,\delta}(Q)$.

4.2.1. Bounds on the confidence interval length

In this section, we state a general upper bound (Theorem 3) on the length of $C_{nm}^{(N)}$, depending on $\gamma_{\alpha,n}, \eta_{\alpha,n}$. We subsequently specialize this result through Corollaries 1, 2 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 $(0, 1)$, and are respectively increasing and decreasing as functions of α . Given $\epsilon \in (0, 1)$, for notational convenience we write $\epsilon := (\epsilon \wedge \alpha)/N$ and $a = \alpha/N$. In the sequel, we also omit explicitly indexing various quantities by the number of Monte Carlo samples N . Our upper bounds on the length of $C_{nm}^{(N)}$ will depend on the function

$$\tilde{\kappa}_{\epsilon,n}(u) = \max \left\{ |f^{-1}(u) - g^{-1}(u)| : f, g \in \{\gamma_{a,n}, \gamma_{\epsilon,n}, \eta_{a,n}, \eta_{\epsilon,n}\} \right\}, \quad u \in (0, 1),$$

as measured by the following two sequences,

$$\kappa_{\epsilon,n} = \sup_{\frac{\delta}{2} \leq u \leq 1 - \frac{\delta}{2}} \tilde{\kappa}_{\epsilon,n}(u), \quad V_{\epsilon,n}(P) = \frac{1}{1 - 2\delta} \int_{\mathbb{S}^{d-1}} \int_{\delta/2}^{1-\delta/2} \left[\frac{\tilde{\kappa}_{\epsilon,n}(u)}{p_\theta(F_\theta^{-1}(u))} \right]^r d\mu(\theta).$$

Here, recall p_θ denotes the density of the absolutely continuous component of P_θ . Additional technical assumptions **B1-B3** regarding $\gamma_{\alpha,n}, \eta_{\alpha,n}, \kappa_{\varepsilon,n}$, appear in Appendix E. For appropriate choices of δ and ε , these assumptions are satisfied by the choices of $\gamma_{\alpha,n}$ and $\eta_{\alpha,n}$ described in Examples 1 and 2, for which the corresponding values of $\kappa_{\varepsilon,n}$ and $V_{\varepsilon,n}(P)$ are derived in the following simple Lemma.

Lemma 1. *Let $\varepsilon \in (0, 1)$.*

1. *If $\gamma_{\varepsilon,n}$ and $\eta_{\varepsilon,n}$ are chosen as in equation (23), then there exist constants $c_1, c_2 > 0$ depending only on r such that*

$$\kappa_{\varepsilon,n} \leq c_1 \sqrt{\frac{\log(4/\varepsilon)}{n}}, \quad \text{and} \quad V_{\varepsilon,n}(P) \leq c_2 \left(\frac{\kappa_{\varepsilon,n}}{\sqrt{\delta}} \right)^r \text{SJ}_{r, \frac{\delta}{2}}(P).$$

2. *If $\gamma_{\varepsilon,n}$ and $\eta_{\varepsilon,n}$ are chosen as in equation (25), then there exist constants $k_1, k_2 > 0$ depending only on r such that*

$$\kappa_{\varepsilon,n} \leq k_1 \nu_{\varepsilon,n}, \quad \text{and} \quad V_{\varepsilon,n}(P) \leq k_2 \nu_{\varepsilon,n}^r \text{SJ}_{r, \frac{\delta}{2}}(P).$$

The proof is a straightforward consequence of Examples 1, 2, together with the derivations in Appendices B and G.6, and is therefore omitted. We now define the functional

$$U_{\varepsilon,n}(P) = \frac{1}{1-2\delta} \int_{\mathbb{S}^{d-1}} \left(\sup_{\substack{\delta \leq u \leq 1-\delta \\ |h| \leq \kappa_{\varepsilon,n}}} |F_\theta^{-1}(u+h) - F_\theta^{-1}(u)|^{r-1} \right) d\mu(\theta).$$

$U_{\varepsilon,n}(P)$ is an upper bound on the magnitude of the largest jump discontinuity of the quantile function F_θ^{-1} , averaged over directions $\theta \in \mathbb{S}^{d-1}$. When $\text{SJ}_{r,\delta}(P) < \infty$, the quantile function F_θ^{-1} is absolutely continuous for almost all $\theta \in \mathbb{S}^{d-1}$ (see Lemma 2 in Appendix A), implying that $U_{\varepsilon,n}(P)$ decays to zero as $n \rightarrow \infty$. The lengths of our confidence intervals will now depend on the quantities

$$\psi_{\varepsilon,nm} = \begin{cases} (\text{SW}_{\infty,\delta}^{(r-1)}(P, Q) + U_{\varepsilon,n}(P) + U_{\varepsilon,m}(Q)) \frac{\sqrt{b\kappa_{\varepsilon,n}}}{\sqrt{\delta}}, & \text{SJ}_{r,\frac{\delta}{2}}(P) \vee \text{SJ}_{r,\frac{\delta}{2}}(Q) = \infty \\ (\text{SW}_{r,\delta}^r(P, Q) + V_{\varepsilon,n}(P) + V_{\varepsilon,m}(Q)) \frac{r-1}{r} [V_{\varepsilon,n}(P)]^{\frac{1}{r}}, & \text{otherwise,} \end{cases}$$

and,

$$\varphi_{\varepsilon,nm} = \begin{cases} (\text{SW}_{\infty,\delta}^{(r-1)}(P, Q) + U_{\varepsilon,n}(P) + U_{\varepsilon,m}(Q)) \frac{\sqrt{b\kappa_{\varepsilon,m}}}{\sqrt{\delta}}, & \text{SJ}_{r,\frac{\delta}{2}}(P) \vee \text{SJ}_{r,\frac{\delta}{2}}(Q) = \infty \\ (\text{SW}_{r,\delta}^r(P, Q) + V_{\varepsilon,n}(P) + V_{\varepsilon,m}(Q)) \frac{r-1}{r} [V_{\varepsilon,m}(Q)]^{\frac{1}{r}}, & \text{otherwise.} \end{cases}$$

With this notation in place, we arrive at the following upper bound on the length of $C_{nm}^{(N)}$. Recall that λ denotes the Lebesgue measure on $\mathbb{R}$. To simplify our statement, we shall only consider the case where δ is bounded away from $1/2$.

Theorem 3. Let $r, b \geq 1$ and $\alpha, \epsilon \in (0, 1)$. Let $P, Q \in \bar{\mathcal{K}}_2(b)$, and define $\delta \in (0, \delta_0)$ for some $\delta_0 \in (0, 1/2)$. Recall that $\varepsilon = (\epsilon \wedge \alpha)/N$, and assume $\kappa_{\varepsilon, n \wedge m} \leq \frac{\delta}{2} \wedge (1 - 2\delta)$. Assume further that conditions **B1-B3** hold for some constants $K_1, K_2 > 0$. Then, there exists $c > 0$ depending only on K_1, K_2, δ_0, r such that with probability at least $1 - \epsilon$,

$$\lambda(C_{nm}^{(N)}) \leq \left\{ \text{SW}_{r,\delta}^r(P, Q) + c(\psi_{\varepsilon, nm} + \varphi_{\varepsilon, nm} + \varkappa_N) \right\}^{1/r} - \text{SW}_{r,\delta}(P, Q).$$

Here, $\varkappa_N$ denotes a random variable depending only on μ_N , such that $\mathbb{E}|\varkappa_N| \leq c_1 N^{-1/2r} I(d \geq 2)$, for a constant $c_1 > 0$ depending on b, r, δ and $K_1 - K_2$.

The proof of Theorem 3 appears in Appendix E. As we shall see, the presence of $\text{SW}_{\infty, \delta}^{(r-1)}(P, Q)$ or $\text{SW}_{r,\delta}(P, Q)$ in the definition of $\varphi_{\varepsilon, nm}$ and $\psi_{\varepsilon, nm}$ implies distinct rates of decay for the confidence interval length, depending on whether P, Q approach each other under the Sliced Wasserstein distance. The fact that $\text{SW}_{\infty, \delta}$ is a stronger metric than $\text{SW}_{r,\delta}$, and the presence of the functional $U_{\varepsilon, n}$, will imply a second dichotomy in the rate of decay of the confidence interval length, based on whether or not $\text{SJ}_{r,\delta/2}(P) \vee \text{SJ}_{r,\delta/2}(Q) < \infty$.

The following result specializes Theorem 3 to Examples 1 and 2.

Corollary 1. Let $r, b \geq 1$ and $\alpha, \epsilon \in (0, 1)$. Let $P, Q \in \bar{\mathcal{K}}_2(b)$, and define $\delta \in (0, \delta_0)$ for some $\delta_0 \in (0, 1/2)$.

- (i) Suppose $\gamma_{\alpha, n}, \eta_{\alpha, n}$ are chosen as in Example 1. Then, there is a constant $c_\alpha > 0$ such that whenever $\delta \wedge (1 - 2\delta) \geq \sqrt{c_\alpha \log(N/\epsilon)/(n \wedge m)}$, we have with probability at least $1 - \epsilon$,

$$\lambda(C_{nm}^{(N)}) \underset{\alpha, b, \delta_0, r}{\lesssim} \varkappa_N^{\frac{1}{r}} + \begin{cases} \delta^{-\frac{1}{2}} \log(N/\epsilon)^{\frac{1}{2r}} \left(n^{-\frac{1}{2r}} + m^{-\frac{1}{2r}} \right), & \text{SJ}_{r, \frac{\delta}{2}}(P) \vee \text{SJ}_{r, \frac{\delta}{2}}(Q) = \infty \\ \delta^{-\frac{1}{2}} \log(N/\epsilon)^{\frac{1}{2}} \left(\frac{\text{SJ}_{r, \delta/2}^{\frac{1}{r}}(P)}{\sqrt{n}} + \frac{\text{SJ}_{r, \delta/2}^{\frac{1}{r}}(Q)}{\sqrt{m}} \right), & \text{otherwise.} \end{cases}$$

- (ii) Suppose $\gamma_{\alpha, n}, \eta_{\alpha, n}$ are chosen as in Example 2. Let $\beta_{\epsilon, nm} = \log(n \wedge m) + \log(N/\epsilon)$. Then, there is a constant $k_\alpha > 0$ such that whenever $\delta \wedge (1 - 2\delta) \geq \sqrt{k_\alpha \beta_{\epsilon, nm}/(n \wedge m)}$, we have with probability at least $1 - \epsilon$,

$$\lambda(C_{nm}^{(N)}) \underset{\alpha, b, \delta_0, r}{\lesssim} \varkappa_N^{\frac{1}{r}} + \begin{cases} \delta^{-\frac{1}{2}} \beta_{\epsilon, nm}^{\frac{1}{2r}} \left(n^{-\frac{1}{2r}} + m^{-\frac{1}{2r}} \right), & \text{SJ}_{r, \frac{\delta}{2}}(P) \vee \text{SJ}_{r, \frac{\delta}{2}}(Q) = \infty \\ \beta_{\epsilon, nm}^{\frac{1}{2}} \left(\frac{\text{SJ}_{r, \delta/2}^{\frac{1}{r}}(P)}{\sqrt{n}} + \frac{\text{SJ}_{r, \delta/2}^{\frac{1}{r}}(Q)}{\sqrt{m}} \right), & \text{otherwise.} \end{cases}$$

Whenever the trimming sequence is chosen as $\delta \asymp (n \wedge m)^{-a}$ for some $a \in (0, 1/2)$, notice that one may allow ϵ to vanish at an exponentially fast rate with respect to $n \wedge m$, in both cases of Corollary 1. The high-probability bounds in this result may then be turned into bounds on the expected confidence interval lengths, similarly as in Proposition 3, though we avoid doing so here for brevity.

Corollary 1(i) shows that the length of the DKW-based interval achieves the minimax lower bound of Theorem 1(i), up to a polylogarithmic factor in N and the approximation error $\varkappa_N$. It does not, however, achieve the optimal dependence on δ . This is a consequence of the DKW inequality not adapting to the variance of the distributions therein. Corollary 1(ii) instead shows that the relative VC-based interval has length depending on δ solely through the magnitude of the $\text{SJ}_{r,\delta/2}$ functional, at the expense of a polylogarithmic term in n, m . In both cases, the confidence interval length scales polynomially with δ^{-1} when $\text{SJ}_{r,\delta/2}(P) \vee \text{SJ}_{r,\delta/2}(Q) = \infty$, suggesting that in this case, the practitioner should not let δ vanish with $n \wedge m$ at a rate faster than logarithmic, to guarantee consistent inference.

When the distributions P and Q are assumed to be bounded away from each other in $\text{SW}_{r,\delta}$, 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 $P, Q \in \bar{\mathcal{K}}_2(b)$ satisfy $\text{SW}_{r,\delta}(P, Q) \geq \Gamma$ for some constant $\Gamma > 0$. Then, under the assumptions of Theorem 3, we have with probability at least $1 - \epsilon$,*

$$\lambda(C_{nm}^{(N)}) \underset{\Gamma, \delta_0, b, r}{\lesssim} \varkappa_N + \begin{cases} \delta^{-\frac{r}{2}} (\kappa_{\epsilon, n} + \kappa_{\epsilon, m}), & \text{SJ}_{r,\delta/2}(P) \vee \text{SJ}_{r,\delta/2}(Q) = \infty \\ V_{\epsilon, n}^{1/r}(P) + V_{\epsilon, m}^{1/r}(Q), & \text{otherwise.} \end{cases}$$

For example, when $\gamma_{\epsilon, n}, \eta_{\epsilon, n}$ are based on the DKW inequality (Example 1), Corollary 2 implies that the length of $C_{nm}^{(N)}$ achieves the parametric rate $n^{-\frac{1}{2}} + m^{-\frac{1}{2}}$ with high probability (ignoring factors depending only on N and δ), 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 $\text{SJ}_{r,\delta/2}(P), \text{SJ}_{r,\delta/2}(Q)$, without further dependence on δ , is available using the relative VC-based interval in Example 2. Further implications of Theorem 3 are discussed in Appendix H.

5. Asymptotic confidence intervals

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.

In the context of goodness-of-fit testing, Munk and Czado (1998) prove central limit theorems of the form $\sqrt{\frac{nm}{n+m}} \{W_{2,\delta}^2(P_n, Q_m) - W_{2,\delta}^2(P, Q)\} \rightsquigarrow N(0, \sigma^2)$, where P and Q are one-dimensional distributions, and $\sigma > 0$. They also construct a consistent estimator of σ . These results assume that $P \neq Q$, and that each of P and Q satisfy the following condition,

(C) F is twice continuously differentiable, with density p , which is strictly

positive over the real line. Moreover,

$$\sup_{x \in \mathbb{R}} F(x)(1 - F(x)) \left| \frac{p'(x)}{p^2(x)} \right| < \infty.$$

Assumption (C) originates from strong approximation theorems for the empirical quantile process (Csorgo and 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 continuous and positive densities on the real line, and still retaining the assumption that $P \neq Q$, Freitag, Munk and Vogt (2003); Freitag and Munk (2005); Freitag, Czado and Munk (2007) prove the consistency of the bootstrap in estimating the distribution of $W_{2,\delta}^2(P_n, Q_m)$ in the one-dimensional case.

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^r 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 those of Munk and Czado (1998) above, or those based on the bootstrap, may have shorter length than those described in Section 4.

We show in Section 5.1 that under some regularity conditions, 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, Munk and Vogt (2003); Freitag and Munk (2005); Freitag, Czado and Munk (2007) from the case $d = 1$ and $r = 2$. We then illustrate in Section 5.2, how the strengths of bootstrap can be combined with those of the finite-sample confidence intervals of Section 4.

5.1. Bootstrapping the Sliced Wasserstein distance

Let $P, Q \in \mathcal{P}(\mathbb{R}^d)$, and let $X_1, \dots, X_n \sim P$, $Y_1, \dots, Y_m \sim Q$ be i.i.d. samples which are independent of each other. Furthermore, let P_n and Q_m denote their 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 13 in Appendix F establishes the Hadamard differentiability of the Sliced Wasserstein distance at pairs of distributions (P, Q) satisfying certain regularity conditions. Limit laws for the empirical Sliced Wasserstein distance, together with consistency of the bootstrap, then follow from the functional delta method (van der Vaart and Wellner, 1996), as outlined in Theorem 4 below. We first introduce some notation. Denote by BL_1 the set of 1-Lipschitz functions $f : \mathbb{R} \rightarrow \mathbb{R}$, such that $\|f\|_\infty \leq 1$. We also write for all $u \in [\delta, 1 - \delta]$ and all $\theta \in \mathbb{S}^{d-1}$,

$$w(u, \theta) = \frac{r}{1 - 2\delta} \text{sgn}(F_\theta^{-1}(u) - G_\theta^{-1}(u)) |F_\theta^{-1}(u) - G_\theta^{-1}(u)|^{r-1},$$

as well as,

$$\begin{aligned} \sigma_P^2 &= \int_0^{1-\delta} \left(\int_{\mathbb{S}^{d-1}} \int_{F_\theta^{-1}(\delta \vee t)}^{F_\theta^{-1}(1-\delta)} w(F_\theta(x), \theta) dx d\mu(\theta) \right)^2 dt \\ &\quad - \left(\int_0^{1-\delta} \int_{\mathbb{S}^{d-1}} \int_{F_\theta^{-1}(\delta \vee t)}^{F_\theta^{-1}(1-\delta)} w(F_\theta(x), \theta) dx d\mu(\theta) dt \right)^2, \end{aligned}$$

and,

$$\begin{aligned} \sigma_Q^2 &= \int_0^{1-\delta} \left(\int_{\mathbb{S}^{d-1}} \int_{G_\theta^{-1}(\delta \vee t)}^{G_\theta^{-1}(1-\delta)} w(G_\theta(x), \theta) dx d\mu(\theta) \right)^2 dt \\ &\quad - \left(\int_0^{1-\delta} \int_{\mathbb{S}^{d-1}} \int_{G_\theta^{-1}(\delta \vee t)}^{G_\theta^{-1}(1-\delta)} w(G_\theta(x), \theta) dx d\mu(\theta) dt \right)^2, \end{aligned}$$

Finally, for any absolutely continuous distribution $P \in \mathcal{P}(\mathbb{R})$ with density p , we shall make use of the following trimmed version of the J_∞ functional introduced by Bobkov and Ledoux (2019),

$$J_{\infty, \delta}(P) = \operatorname{esssup}_{\delta \leq u \leq 1-\delta} \frac{1}{p(F^{-1}(u))}.$$

With this notation in place, the main result of this section is stated as follows.

Theorem 4. Let $\delta \in [0, 1/2)$ and $r > 1$. Let $P, Q \in \bar{\mathcal{K}}_2$ be distributions such that for all $\theta \in \mathbb{S}^{d-1}$, P_θ, Q_θ are absolutely continuous with respect to the Lebesgue measure, with respective families of densities $\{p_\theta\}_{\theta \in \mathbb{S}^{d-1}}, \{q_\theta\}_{\theta \in \mathbb{S}^{d-1}}$ which are uniformly integrable over $\mathbb{R}$. Assume that,

$$\sup_{\theta \in \mathbb{S}^{d-1}} J_{\infty, \delta/2}(P_\theta) \vee J_{\infty, \delta/2}(Q_\theta) < \infty. \quad (30)$$

Then, the following statements hold as $n, m \rightarrow \infty$ such that $\frac{n}{n+m} \rightarrow a \in (0, 1)$.

(i) (Central Limit Theorem) We have,

$$\sqrt{\frac{nm}{n+m}} \left(\operatorname{SW}_{r, \delta}^r(P_n, Q_m) - \operatorname{SW}_{r, \delta}^r(P, Q) \right) \rightsquigarrow N(0, a\sigma_P^2 + (1-a)\sigma_Q^2).$$

(ii) (Bootstrap Consistency) For $\mathbf{X} = (X_1, \dots, X_n), \mathbf{Y} = (Y_1, \dots, Y_m)$, we have

$$\begin{aligned} \sup_{h \in \text{BL}_1} \left| \mathbb{E} \left[h \left(\sqrt{\frac{nm}{n+m}} \{ \operatorname{SW}_{r, \delta}^r(P_n^*, Q_m^*) - \operatorname{SW}_{r, \delta}^r(P_n, Q_m) \} \right) \middle| \mathbf{X}, \mathbf{Y} \right] \right. \\ \left. - \mathbb{E} \left[h \left(\sqrt{\frac{nm}{n+m}} \{ \operatorname{SW}_{r, \delta}^r(P_n, Q_m) - \operatorname{SW}_{r, \delta}^r(P, Q) \} \right) \right] \right| \rightarrow 0, \end{aligned}$$

in outer probability.

Theorem 4(i) provides a central limit theorem for the empirical trimmed Sliced Wasserstein distance, centered at its population counterpart. The primary assumptions required for this result are (a) the existence and uniform integrability of the densities of P_θ and Q_θ along directions $\theta \in \mathbb{S}^{d-1}$, and (b) a uniform lower bound on these densities, over the compact sets $[F_\theta^{-1}(\delta/2), F_\theta^{-1}(1 - \delta/2)]$, as measured by the $J_{\infty, \delta/2}$ functional. Note that assumption (a) holds if P, Q admit upper bounded densities with respect to the Lebesgue measure over $\mathbb{R}^d$, but is strictly weaker; indeed, it can be satisfied by non-atomic measures which are singular with respect to the Lebesgue measure on $\mathbb{R}^d$. Furthermore, we note that the assumption of uniform integrability is vacuous in the special case $d = 1$. Assumption (b) requires the bulk of the supports of P_θ and Q_θ to be connected. Such a condition is necessary for the limit in Theorem 4(i) to be a mean-zero Gaussian distribution, as can be anticipated, for instance, from the lack of Hadamard differentiability of the Wasserstein distance over finite spaces (Sommerfeld and Munk, 2018). Nevertheless, our condition is stronger, since we have assumed $J_{\infty, \delta/2}(P_\theta)$ and $J_{\infty, \delta/2}(Q_\theta)$ are uniformly bounded in θ . Inspired by our results on estimation and finite sample inference for the Sliced Wasserstein distance, it is natural to ask whether this condition can be replaced by, say, $\text{SJ}_{r, \delta}(P), \text{SJ}_{r, \delta}(Q) < \infty$. Such a condition would allow the densities p_θ and q_θ to approach zero at a sufficiently slow rate, which is currently precluded by our theorem. We leave this question open for future work.

In the special case $d = 1$ and $r = 2$, the limiting variance obtained in Theorem 4(i) is equal to the one obtained by Munk and Czado (1998), up to renormalizing their definition of the trimmed Wasserstein distance, though our assumptions are significantly weaker since we do not require the aforementioned condition (C). Nevertheless, their result allows the trimming constant δ to vanish, while we require δ to be held fixed and, in fact, positive, when P and Q have unbounded support. In this regard, our result is closer to those of Freitag, Munk and Vogt (2003); Freitag and Munk (2005); Freitag, Czado and Munk (2007), who prove, in particular, the Hadamard differentiability of the functional $W_{2, \delta}^2$ for a nonvanishing trimming constant δ . Their results require P and Q to admit positive and continuously differentiable densities over the real line, which is a strictly stronger assumption than those of Theorem 4. In particular, we require no smoothness conditions on the various densities.

We next compare Theorem 4(i) to existing central limit theorems for untrimmed Wasserstein distances. Let $d = 1$, and assume P and Q are compactly-supported, so that one may take $\delta = 0$ in Theorem 4(i). In this case, the limiting variance may be reformulated in terms of the expressions

$$\sigma_P^2 = \text{Var}[\phi_0(X)], \quad \sigma_Q^2 = \text{Var}[\psi_0(Y)],$$

where $X \sim P \in \mathcal{P}(\mathbb{R})$, $Y \sim Q \in \mathcal{P}(\mathbb{R})$, and for all $x, y \in \mathbb{R}$,

$$\phi_0(x) = \int_{-\infty}^x w(F(t))dt, \quad \psi_0(y) = \int_{-\infty}^y w(G(t))dt.$$

Here, we abbreviate $w(\cdot) \equiv w(\cdot, \theta)$ in the one-dimensional case. It can be deduced from Gangbo and McCann (1996) that (ϕ_0, ψ_0) forms an optimal pair

of Kantorovich potentials in the dual $|\cdot|^r$ -optimal transport problem (Villani, 2003) from P to Q . In particular, for $r = 2$, the limiting variance in Theorem 4(i) reduces to the one obtained by del Barrio and Loubes (2019), who derive central limit theorems for $W_2^2(P_n, Q_m)$, in general dimension $d \geq 1$. Their results are not centered at $W_2^2(P, Q)$ due to the large bias of empirical Wasserstein distances in general dimension, however, it was shown by del Barrio, Gordaliza and Loubes (2019) that when $d = 1$, these limit theorems may be centered at the population Wasserstein distance under assumptions akin to condition (C). We also refer to Berthet, Fort and Klein (2020) and the recent work of Hundrieser et al. (2022) for distinct assumptions under which such a result can be obtained.

Theorem 4(ii) proves the consistency of the bootstrap in estimating the distribution of $\text{SW}_{r,\delta}^r(P_n, Q_m)$. Letting F_{nm}^* denote the CDF of $\text{SW}_{r,\delta}^r(P_n^*, Q_m^*) - \text{SW}_{r,\delta}^r(P_n, Q_m)$, it follows that an asymptotic $(1 - \alpha)$ -confidence interval for $\text{SW}_{r,\delta}(P, Q)$ is given by

$$C_{nm}^* = \left[\left(\text{SW}_{r,\delta}^r(P_n, Q_m) - F_{nm}^*(1 - \alpha/2) \right)^{\frac{1}{r}}, \left(\text{SW}_{r,\delta}^r(P_n, Q_m) + F_{nm}^*(\alpha/2) \right)^{\frac{1}{r}} \right].$$

The CDF F_{nm}^* is typically estimated via Monte Carlo simulation (Efron and Tibshirani, 1994). The assumptions for the validity of C_{nm}^* are those of Theorem 4, and in addition, the condition that $\text{SW}_{r,\delta}(P, Q) > 0$, which is necessary and sufficient for the limiting variance $a\sigma_P^2 + (1-a)\sigma_Q^2$ in Theorem 4 to be positive. Failure of the bootstrap at the null $\text{SW}_{r,\delta}(P, Q) = 0$ is due to the Sliced Wasserstein distance being a functional with first-order degeneracy (Munk and Czado, 1998), for which corrections such as those of Chen and Fang (2019), or the m -out-of- n bootstrap (Sommerfeld and Munk, 2018), yield consistent procedures, but are practically less attractive as they introduce further tuning parameters. We illustrate in the sequel how our finite sample confidence intervals can be combined with the bootstrap to relax this assumption.

5.2. A hybrid bootstrap approach

Let C_{nm}^* denote the preceding bootstrap confidence interval at level $1 - \alpha/2$, and let $C_{nm}^\dagger$ denote the assumption-light confidence interval for $\text{SW}_{r,\delta}(P, Q)$ in equation (29) at level $1 - \alpha/2$. Assume that the number of Monte Carlo replications N therein is taken to diverge as $n, m \rightarrow \infty$. We define the $(1 - \alpha)$ -hybrid confidence interval as:

$$C_{nm} = \begin{cases} C_{nm}^\dagger, & \text{if } 0 \in C_{nm}^\dagger, \\ C_{nm}^*, & \text{otherwise.} \end{cases} \quad (31)$$

Roughly, we use the bootstrap interval if we are reasonably certain that P and Q are bounded away from each other in Sliced Wasserstein distance, and fall back on the finite-sample interval otherwise. The following simple result characterizes the asymptotic coverage and length of the hybrid interval. In order to simplify our discussion, we write $C_{nm} = [a_{nm}^{1/r}, b_{nm}^{1/r}]$ and we focus on bounding

the length of the confidence interval $[a_{nm}, b_{nm}]$ for the r -th power of the r -Sliced Wasserstein distance. We also assume that the finite-sample interval $C_{nm}^\dagger$ is defined in terms of the DKW confidence band in Example 1.

Proposition 7. *Let $a, \alpha \in (0, 1)$, $\delta \in (0, 1/2)$, and assume the same conditions as Theorem 4. Then, the following holds assuming $\frac{n}{n+m} \rightarrow a$ when $n, m \rightarrow \infty$.*

(i) (Coverage) *We have,*

$$\liminf_{n, m \rightarrow \infty} \mathbb{P}(\text{SW}_{r, \delta}(P, Q) \in C_{nm}) \geq 1 - \alpha. \quad (32)$$

(ii) (Length) *Let $N \asymp n^{r^2}$, and choose $M_N \asymp \log N$ in the definition of $C_{nm}^\dagger$. Then, we have with probability at least $1 - \alpha$,*

$$(b_{nm} - a_{nm}) = O\left(\left(\frac{\log n}{n}\right)^{\frac{r}{2}} + \frac{\text{SW}_{r, \delta}(P, Q)}{\sqrt{n}}\right).$$

Proposition 7 establishes the asymptotic coverage of C_{nm} under the same conditions as Theorem 4. In particular, it removes the assumption $\text{SW}_{r, \delta}(P, Q) > 0$, which is needed for the asymptotic coverage of the bootstrap interval C_{nm}^* . We note that many other existing corrections of the bootstrap for functionals with first-order degeneracy, such as the m -out-of- n bootstrap or the procedures outlined in Section 2.1 of Verdinelli and Wasserman (2021), involve expanding the asymptotic length of the interval, leading to a loss of efficiency. In contrast, Proposition 7 shows that with high (albeit, fixed) probability, the hybrid interval achieves the rate-optimal asymptotic length both at the null $\text{SW}_{r, \delta}(P, Q) = 0$ and away from the null, up to a polylogarithmic factor in N (which can be removed when $d = 1$). We emphasize that this adaptivity is obtained without tuning parameters, apart from the sequences M_N, N whose precise choice does not greatly alter the properties of C_{nm} . Once again, the choice of these sequences is vacuous in the special case $d = 1$.

Though this methodology inherits benefits from both the bootstrap and finite-sample confidence intervals, it is not assumption-free. In principle, it is possible to extend this procedure by empirically testing whether the conditions of Theorem 4 are met, and to use the outcome of such a test in the conditions of equation (31). While doing so may allow for certain assumptions to be relaxed, it could become impractical: for instance, we do not know of a test for the finiteness of the $J_{\infty, \delta/2}$ functional which is free of tuning parameters.

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 twelve cores. Implementations for all confidence intervals described in this paper, along with code for reproducing the following simulations, are publicly available¹.

¹<https://github.com/tmanole/SW-inference>.

TABLE 1

Parameter settings for Models 1-5. For any $R > r > 0$, $\mathbb{T}(r, R)$ denotes the uniform distribution over the torus $\{((R + r \cos \theta) \cos \psi, (R + r \cos \theta) \sin \psi, r \sin \theta)^\top : 0 \leq \theta, \psi \leq 2\pi\} \subseteq \mathbb{R}^3$. We sample from $\mathbb{T}(r, R)$ using the Algorithm 1 of Diaconis, Holmes and Shahshahani (2013).

Mod.	P	Q
1	$\frac{1}{2}N((-1, -1)^\top, I_2) + \frac{1}{2}N((1, 1)^\top, I_2)$	$N(0, I_2)$
2	$\frac{1+n^{-1/2}}{2}\delta_2 + \frac{1-n^{-1/2}}{2}\delta_4$	$\frac{1}{2}\delta_2 + \frac{1}{2}\delta_5$
3	$\mathbb{T}(\frac{1}{2}, 1)$	$\mathbb{T}(\frac{1}{2}, 5)$
4	$.95N(0, 1) + .05N(0.1)$	$N(0, 1)$
5	$.55N((-5, -5)^\top, I_2) + .45N((5, 5)^\top, I_2)$	$\frac{1}{2}N((-5, -5)^\top, I_2) + \frac{1}{2}N((5, 5)^\top, I_2)$

Comparison of asymptotic and finite-sample confidence intervals

We compare the following confidence intervals: (i) The finite-sample interval in equation (28) (or (26) in the one-dimensional case), based on the DKW inequality from Example 1, (ii) The standard bootstrap confidence interval C_{nm}^* in Section 5.1, and (iii) The hybrid interval in equation (31). We also implemented the finite-sample interval (28) 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 (24), unlike those in the DKW inequality of Massart (1990), and consequently we do not consider this method in the present simulation study.

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 $B = 1,000$, and we set $r = 2$ except where otherwise specified. 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 interval does not distinguish $\text{SW}_{r,\delta}(P, Q)$ from zero on all replications, hence the hybrid confidence interval exhibits length and coverage between those of the finite-sample and bootstrap intervals. For larger sample sizes, the length of the hybrid interval essentially coincides with that of the bootstrap interval. The distributions in Model 2 are not absolutely continuous, thus the bootstrap and hybrid intervals are seen to undercover the true Wasserstein distance. Those of Model 3 are 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, with positive and distinct densities, thus ensuring that the conditions for all three methods are met. Models 4 and 5 consist of pairs of measures admitting Sliced Wasserstein distance near zero, causing the

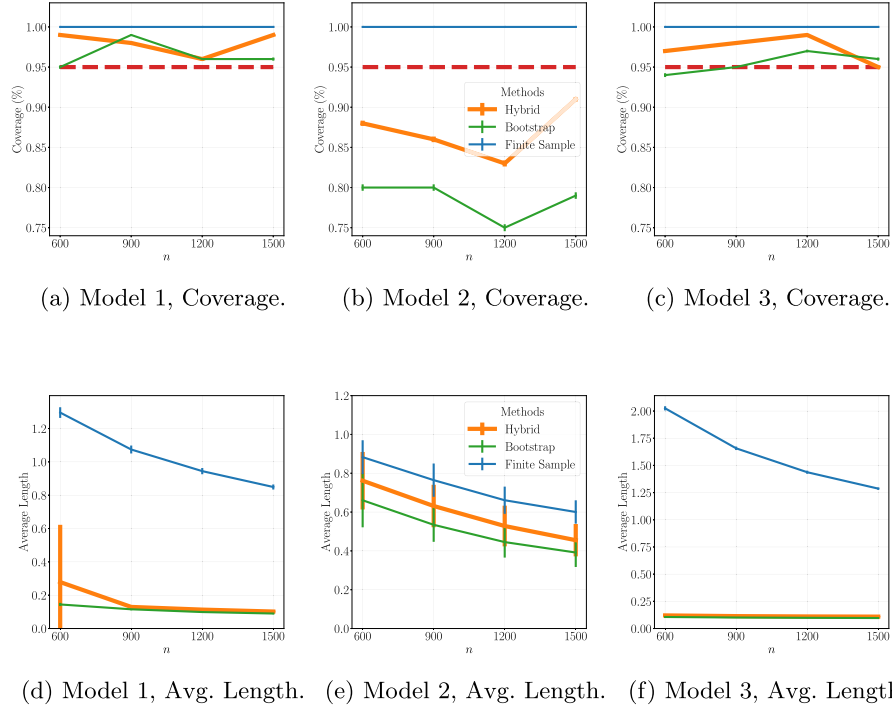


FIG 1. Average interval length and coverage for Models 1-3. Error bars represent one standard deviation. Dotted line (- -) indicate the nominal level 0.95.

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 3, and Corollaries thereafter, regarding the asymptotic length of our finite-sample confidence intervals. Consider the following two pairs of distributions,

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

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

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 (26), 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 $\Delta = 0$ under varying orders $r \in [1, 16]$ of the Wasserstein distance, while in Figures 3(c) and 3(d), we

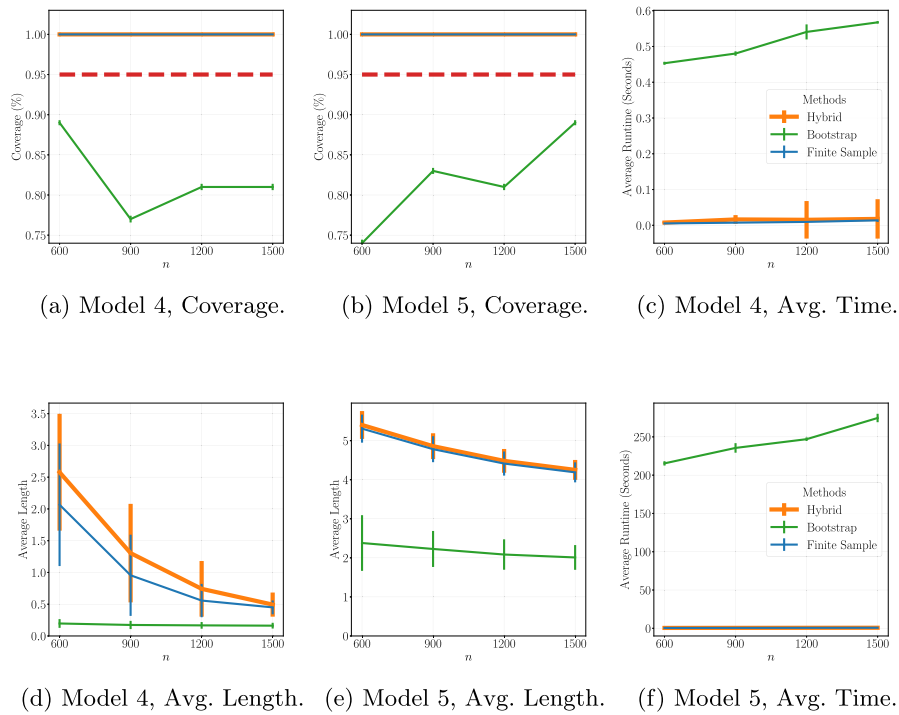


FIG 2. Average interval length, coverage, and runtime for Models 4 and 5. Error bars represent one standard deviation. Dotted lines (---) indicate the nominal level 0.95.

do so for a range of values $\Delta \in [0, .4]$ under the fixed 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 r increases, while that of the pair $(P_2, Q_{2,0})$ remains nearly unchanged. This behaviour was predicted by Corollary 1, indicating that the finite-sample interval length scales at the $n^{-1/2r}$ rate in general, but does so at the faster rate $n^{-1/2}$ for distributions admitting finite $\text{SJ}_{r,\delta}$ 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 Δ decreases to zero, yet nearly identical 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 (Brehmer et al., 2020), predator-prey dynamics in ecology (Lotka, 1920a,b), inference for cosmological parameters in astronomy (Dalmasso et al., 2020; Dalmasso, Izibicki and Lee, 2020), and network dynamics in queuing theory (Ebert

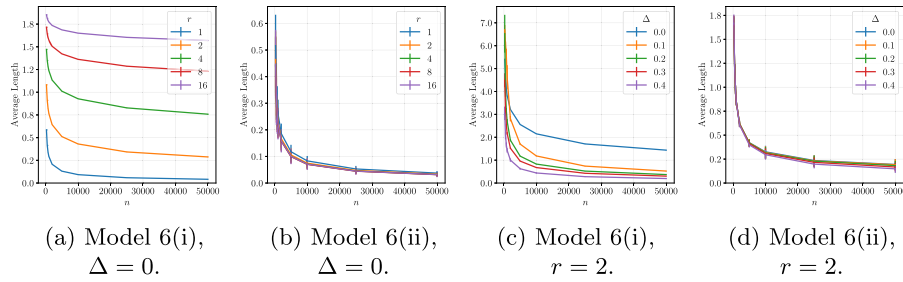


FIG 3. Average length of our finite-sample confidence interval for Models 6(i) and 6(ii), for varying values of r and Δ , based on 100 replications for each sample size considered. Error bars represent one standard deviation of the confidence interval length.

et al., 2021). In such applications, the practitioner typically has access to a parametrized stochastic simulator for the data generating process, which produces samples from a distribution $P^\eta \in \mathcal{P}(\mathbb{R}^d)$ with unknown closed form, and which depends on some physically meaningful parameters $\eta \in \Theta \subseteq \mathbb{R}^D$. The goal of *likelihood-free inference* is to characterize the values of η for which an observed sample $X_1, \dots, X_n$ is likely to have been generated by the simulator.

Approximate Bayesian computation (ABC; Sisson, Fan and Beaumont (2018)) is arguably the most popular family of methodologies for likelihood-free inference. ABC methods repeatedly simulate parameter values in Θ , and accept those for which the simulator P^η 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 η , 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, Jitkrittum and Sejdinovic, 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 $X_1, \dots, X_n$ is an i.i.d. sample from a distribution $P \in \mathcal{P}(\mathbb{R}^d)$. Fix $\eta_0 \in \arg\min_{\eta \in \Theta} \text{SW}_{r,\delta}(P, P^\eta)$, and let $\epsilon_0 = \text{SW}_{r,\delta}(P, P^{\eta_0})$. Here η_0 denotes an $\text{SW}_{r,\delta}$ -projection of the distribution P onto the family $\{P^\eta\}_{\eta \in \Theta}$. If the simulator is correctly-specified, we have $P = P^{\eta_0}$ and $\epsilon_0 = 0$, whereas if the simulator is misspecified, the set $\{\eta \in \Theta : \text{SW}_{r,\delta}(P, P^\eta) \leq \epsilon\}$, is empty for sufficiently small values of $\epsilon \geq 0$.

We propose to construct confidence sets for η_0 . For any $\eta \in \Theta$, and for any synthetic sample $Y_1^\eta, \dots, Y_m^\eta \sim P^\eta$, let $[\bar{\ell}_{nm}^{(N)}(\eta), \bar{u}_{nm}^{(N)}(\eta)]$ be a $(1 - \alpha)$ -confidence interval for $\text{SW}_{r,\delta}(P, P^\eta)$, obtained via equation (29) on the basis of $Y_1^\eta, \dots, Y_m^\eta$ and the observed sample $X_1, \dots, X_n$. A confidence set for η_0 is then easily given by the following Proposition.

Proposition 8. *Let $b, r \geq 1$ and $\delta \in (0, 1/2)$ be fixed. Given any fixed real number $\epsilon \geq \epsilon_0$, define $\bar{C}_{nm}^{(N)} = \{\eta \in \Theta : \bar{\ell}_{nm}^{(N)}(\eta) \leq \epsilon\}$. Then,*

$$\inf_{\substack{P \in \mathcal{K}_{2r}(b) \\ \{P^\eta\}_{\eta \in \Theta} \subseteq \mathcal{K}_{2r}(b)}} \mathbb{P}(\eta_0 \in \bar{C}_{nm}^{(N)}) \geq 1 - \alpha - O(M_N^{-2}).$$

Proposition 8 provides a $(1 - \alpha)$ -confidence set for the projection parameter η_0 . In the well-specified setting $\epsilon_0 = 0$, $\bar{C}_{nm}^{(N)}$ is simply a confidence set for the parameter corresponding to the data-generating distribution $P = P^{\eta_0}$. We emphasize that no assumptions were made in the statement of Proposition 8 beyond the mild moment assumption $P, P^\eta \in \mathcal{K}_{2r}(b)$ (which can be removed when $d = 1$). The intractability of the likelihood function makes such assumption-lean inference particularly attractive.

In practice, computation of the lower confidence bounds $\bar{\ell}_{nm}^{(N)}(\eta)$ in Proposition 8 may be carried out over a finite grid $\{\eta_1, \dots, \eta_M\} \subseteq \Theta$ 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 $[\bar{\ell}_{nm}^{(N)}(\eta), \bar{u}_{nm}^{(N)}(\eta)]$ can dramatically reduce the computational burden of $\bar{C}_{nm}^{(N)}$. The simulation study in Section 6 suggests that the runtime of our finite-sample 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, You and West (2011); Bonassi and 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 $i \in \{1, \dots, n\}$, and at time $t \in \{1, \dots, T\}$. Given a starting value $(U_{i,0}, V_{i,0})$ for every $i = 1, \dots, n$, the model is given by

$$\begin{cases} U_{i,t+1} = U_{i,t} + \frac{\alpha_1}{1+V_{i,t}^{\beta_1}} - (1 + 0.03U_{i,t}) + \frac{1}{2}\xi_{i,t} \\ V_{i,t+1} = V_{i,t} + \frac{\alpha_2}{1+U_{i,t}^{\beta_2}} - (1 + 0.03V_{i,t}) + \frac{1}{2}\zeta_{i,t} \end{cases}, \quad t = 1, \dots, T, \quad (33)$$

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 (33) is not observed, except for the noisy measurement

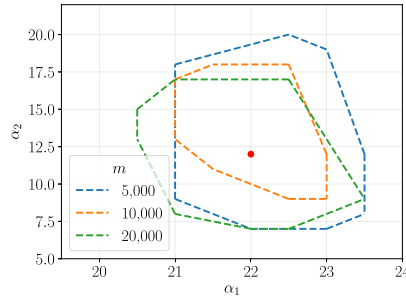


FIG 4. Confidence sets for η_0 (in red $(\bullet)$) under the well-specified setting.

$X_i = U_{i,T} + \epsilon_i$ at time T , where $\epsilon_i \sim N(\mu, \mu\sigma/U_{i,T}^\gamma)$ are drawn conditionally on a realization $U_{i,T}$ from the model (33), for $i = 1, \dots, n$, where $\mu \in \mathbb{R}$ and $\sigma, \gamma \geq 0$. $X_1, \dots, X_n$ thus forms an i.i.d. sample from a distribution P^η on $\mathbb{R}$ with respect to the parameter $\eta = (\alpha_1, \alpha_2, \beta_1, \beta_2, \mu, \sigma, \gamma) \in \mathbb{R}^7$. A closed form for P^η is unclear, but the evolution (33) makes simulation from P^η 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 β_1, β_2, γ as known, but possibly misspecified, and perform inference on $(\alpha_1, \alpha_2, \mu, \sigma)$. In what follows, we set $U_{i,0} = V_{i,0} = 10$, $i = 1, \dots, n$, and we generate $n = 2,000$ observations from P^{η_0} with $\eta_0 = (22, 12, 4, 4.5, 325, .25, .15)$, 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 C_{nm} of Proposition 8 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 (α_1, α_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, 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 ϵ , and can be seen to cover the projection parameter η_0 . The latter was approximated by $\hat{\eta}_0 = \operatorname{argmin}_\eta W_{1,\delta}(P_M, P^\eta)$ over a grid of parameters η , 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 minimax

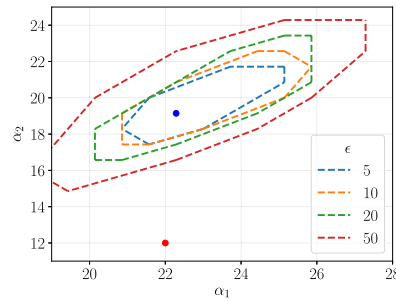


FIG 5. Confidence sets for η_0 (in blue $\bullet$) under the misspecified setting. The data-generating parameter is shown in red $\bullet$.

rates for estimating the Sliced Wasserstein distance, which are of independent interest, we bounded the length of our confidence intervals, showing that they achieve near minimax optimal length. 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 $\text{SJ}_{r,\delta}$. 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^r 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.

Appendix A: Preliminary technical 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 Appendix A of Bobkov and Ledoux (2019).

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

Furthermore, we describe the following characterization of distributions falling in the collections $\mathcal{K}_{r,\rho}(b)$ and $\bar{\mathcal{K}}_r(b)$.

Lemma 3. Let $\delta \in (0, 1/2)$ and $r, \rho, b \geq 1$. Then, for all distributions $P \in \mathcal{K}_{r,\rho}(b)$,

$$\int_{\mathbb{S}^{d-1}} |F_{\theta}^{-1}(a)|^r d\mu(\theta) \leq b(2/\delta)^{r/\rho}, \quad a \in \{\delta, 1-\delta\}.$$

Furthermore, for all distributions $P \in \bar{\mathcal{K}}_r(b)$, we have

$$\sup_{\theta \in \mathbb{S}^{d-1}} |F_{\theta}^{-1}(a)| \leq \left(\frac{2b}{\delta}\right)^{\frac{1}{r}}, \quad a \in \{\delta, 1-\delta\}.$$

Proof of Lemma 3 The claim is a simple consequence of Markov's inequality. Given $\theta \in \mathbb{S}^{d-1}$, it must hold that (a) $F_{\theta}^{-1}(\delta) < 0$ or (b) $F_{\theta}^{-1}(1-\delta) > 0$. In the former case, since $F_{\theta}(F_{\theta}^{-1}(\delta)) \geq \delta$, we have

$$\begin{aligned} \delta &\leq \mathbb{P}\{X^{\top}\theta \leq F_{\theta}^{-1}(\delta)\} = \mathbb{P}\{-X^{\top}\theta \geq |F_{\theta}^{-1}(\delta)|\} \\ &\leq \mathbb{P}\{|X^{\top}\theta| \geq |F_{\theta}^{-1}(\delta)|\} \leq \frac{\mathbb{E}_X[|X^{\top}\theta|^{\rho}]}{|F_{\theta}^{-1}(\delta)|^{\rho}}. \end{aligned} \quad (34)$$

while in the latter case, $F_{\theta}(F_{\theta}^{-1}(1-\delta)/2^{1/\rho}) \leq 1-\delta$, therefore

$$\delta \leq \mathbb{P}\left\{X^{\top}\theta \geq \frac{F_{\theta}^{-1}(1-\delta)}{2^{1/\rho}}\right\} \leq \mathbb{P}\left\{|X^{\top}\theta| \geq \frac{|F_{\theta}^{-1}(1-\delta)|}{2^{1/\rho}}\right\} \leq \frac{2\mathbb{E}[|X^{\top}\theta|^{\rho}]}{|F_{\theta}^{-1}(1-\delta)|^{\rho}} \quad (35)$$

We deduce,

$$\max_{a \in \{\delta, 1-\delta\}} |F_{\theta}^{-1}(a)| \leq \left(\frac{2\mathbb{E}[|X^{\top}\theta|^{\rho}]}{\delta}\right)^{\frac{1}{\rho}}. \quad (36)$$

Indeed, when both (a) and (b) hold, the above display follows from equations (34) and (35). When only (a) holds, it is clear that $|F_{\theta}^{-1}(\delta)| \geq |F_{\theta}^{-1}(1-\delta)|$, thus the above display follows from equation (35), and similarly when only case (b) holds. Thus, since the above display holds for any $\theta \in \mathbb{S}^{d-1}$, we deduce that for both $a \in \{\delta, 1-\delta\}$,

$$\int_{\mathbb{S}^{d-1}} |F_{\theta}^{-1}(a)|^r d\mu(\theta) \leq \left(\frac{2}{\delta}\right)^{\frac{r}{\rho}} \int_{\mathbb{S}^{d-1}} \mathbb{E}[|X^{\top}\theta|^{\rho}]^{\frac{r}{\rho}} d\mu(\theta) \leq \left(\frac{2}{\delta}\right)^{\frac{r}{\rho}} b,$$

where we used the assumption $P \in \mathcal{K}_{r,\rho}(b)$.

To prove the second claim, one similarly has the following bound when $P \in \bar{\mathcal{K}}_{\rho}(b)$, for $a \in \{\delta, 1-\delta\}$,

$$\begin{aligned} \sup_{\theta \in \mathbb{S}^{d-1}} |F_{\theta}^{-1}(a)| &\leq \sup_{\theta \in \mathbb{S}^{d-1}} \left(\frac{2\mathbb{E}[|X^{\top}\theta|^{\rho}]}{\delta}\right)^{\frac{1}{\rho}} \\ &\leq \left(\frac{2}{\delta} \mathbb{E}\left[\sup_{\theta \in \mathbb{S}^{d-1}} |X^{\top}\theta|^{\rho}\right]\right)^{\frac{1}{\rho}} = \left(\frac{2}{\delta} \mathbb{E}[\|X\|^{\rho}]\right)^{\frac{1}{\rho}} \leq \left(\frac{2b}{\delta}\right)^{\frac{1}{\rho}}. \quad \square \end{aligned}$$

Appendix B: Proof of Propositions 1 and 3

We shall begin by proving Proposition 1(i) and Proposition 3(ii). As we shall explain, Proposition 1(ii) will then follow as a special case of Proposition 3(ii), and Proposition 3(i) will follow from Proposition 1(i). Our proofs will make use of Examples 1 and 2 which appear in Section 4 of the main text, and of their corresponding proofs in Appendix G. We shall also make use of the following Lemma, which is proven in Appendix B.1.

Lemma 4. *Let $\delta \in (0, 1/2)$. Then, for any $\bar{r} \in [r, r+1]$, there exists a constant $B_r > 0$ depending only on r such that for all $\theta \in \mathbb{S}^{d-1}$ and all $\delta \geq 2(r+2)/n$,*

$$\max_{a \in \{\delta, 1-\delta\}} \mathbb{E} \left[|F_{\theta,n}^{-1}(a)|^{\bar{r}} \right] \leq B_r \left(1 + \max_{a \in \{\frac{\delta}{2}, 1-\frac{\delta}{2}\}} |F_{\theta}^{-1}(a)|^{\bar{r}} + \left(\frac{\mathbb{E}|X^\top \theta|^2}{\delta} \right)^{\frac{\bar{r}}{2}} \right).$$

In the above result and throughout this section, recall that $F_{\theta,n}^{-1}$ (resp. $G_{\theta,n}^{-1}$, F_{θ}^{-1} , G_{θ}^{-1}) denotes the quantile function of the distribution $P_{\theta,n} = \pi_{\theta\#}P_n$ (resp. $Q_{\theta,m} = \pi_{\theta\#}Q_m$, $P_{\theta} = \pi_{\theta\#}P$, $Q_{\theta} = \pi_{\theta\#}Q$), for any $\theta \in \mathbb{S}^{d-1}$. Finally, throughout the remainder of this section, the symbol “ $\lesssim$ ” is used to hide universal constants depending only on r .

Proof of Proposition 1(i) The claim is trivial if $\text{SJ}_{r,\delta/2}(P) = \infty$, thus assume otherwise. We shall begin bounding the expectation of the following quantity

$$Z_n(\theta) = W_{r,\delta}^r(P_{\theta,n}, P_{\theta}),$$

for any fixed $\theta \in \mathbb{S}^{d-1}$. The result will then follow by integration over $\mathbb{S}^{d-1}$. We begin with the following key high probability bound.

Lemma 5. *Let $y_0 = \sqrt{\delta}/4$ and $C_r > 0$ a constant depending only on r . Then, for all $y \in (0, y_0]$, we have*

$$\text{esssup}_{\theta \in \mathbb{S}^{d-1}} \mathbb{P} \left(Z_n(\theta) \geq C_r y^r J_{r,\delta/2}(P_{\theta}) \right) \leq \frac{2n+1}{16} e^{-\frac{ny^2}{16}},$$

where the essential supremum is taken with respect to the measure μ .

Lemma 5 is proven in Appendix B.2. Now, let $T_{\theta} = C_r y_0^r J_{r,\delta/2}(P_{\theta})$, so that

$$\mathbb{E}[Z_n(\theta)] = \mathbb{E}[Z_n(\theta) \cdot I(Z_n(\theta) > T_{\theta})] + \mathbb{E}[Z_n(\theta) \cdot I(Z_n(\theta) \leq T_{\theta})]. \quad (37)$$

To bound the first term, notice first that

$$Z_n(\theta) = \frac{1}{1-2\delta} \int_{\delta}^{1-\delta} |F_{\theta,n}^{-1}(u) - F_{\theta}^{-1}(u)|^r du \lesssim \max_{a \in \{\delta, 1-\delta\}} |F_{\theta,n}^{-1}(a)|^r + |F_{\theta}^{-1}(a)|^r.$$

Now, let $s = 1/(1 - 1/\eta)$, where $\eta = \bar{r}/r$ and $\bar{r} \in (r, r+1]$. Then, by Hölder’s inequality, we have uniformly in $\theta \in \mathbb{S}^{d-1}$,

$$\mathbb{E}[Z_n(\theta) \cdot I(Z_n(\theta) > T_{\theta})]$$

$$\begin{aligned}
&\leq \|Z_n(\theta)\|_{L^\eta(\mathbb{P})} \|I(Z_n(\theta) > T_\theta)\|_{L^s(\mathbb{P})} \\
&\lesssim \max_{a \in \{\frac{\delta}{2}, 1-\delta\}} \left(\|F_{\theta,n}^{-1}(a)\|_{L^{\bar{r}}(\mathbb{P})}^{\bar{r}} + |F_{\theta}^{-1}(a)|^{\bar{r}} \right)^{\frac{1}{\eta}} \|I(Z_n(\theta) > T_\theta)\|_{L^s(\mathbb{P})} \\
&\lesssim \left(1 + \max_{a \in \{\frac{\delta}{2}, 1-\frac{\delta}{2}\}} |F_{\theta}^{-1}(a)|^{\bar{r}} + \left(\frac{\mathbb{E}|X^\top \theta|^2}{\delta} \right)^{\frac{\bar{r}}{2}} \right)^{\frac{1}{\eta}} \left(\frac{2n+1}{16} \right)^{\frac{1}{s}} e^{-\frac{ny_0^2}{16s}} \\
&\leq \left(1 + \max_{a \in \{\frac{\delta}{2}, 1-\frac{\delta}{2}\}} |F_{\theta}^{-1}(a)|^r + \left(\frac{\mathbb{E}|X^\top \theta|^2}{\delta} \right)^{\frac{r}{2}} \right) \left(\frac{2n+1}{16} \right)^{\frac{1}{s}} e^{-\frac{ny_0^2}{16s}}, \tag{38}
\end{aligned}$$

where we invoked Lemmas 3–5 in equation (38). Therefore, using the fact that $s \geq 1$, $P \in \mathcal{K}_r(b)$, and invoking Lemma 3, we obtain

$$\int_{\mathbb{S}^{d-1}} \mathbb{E}[Z_n(\theta) \cdot I(Z_n(\theta) > T_\theta)] d\mu(\theta) \lesssim (b/\delta^{r/2}) n e^{-\frac{ny_0^2}{16s}} \lesssim b n e^{-c_0 n \delta} / \delta^{r/2}, \tag{39}$$

for a universal constant $c_0 > 0$ depending only on r . We further bound the second term in equation (37). Setting $t_\theta = T_\theta \wedge C_r J_{r,\delta/2}(P_\theta) \left(\frac{16}{n} \log \left(\frac{2n+1}{16} \right) \right)^{r/2}$, we arrive at

$$\begin{aligned}
&\mathbb{E}[Z_n(\theta) \cdot I(Z_n(\theta) < T_\theta)] \\
&= \int_0^\infty \mathbb{P}(Z_n(\theta) \cdot I(Z_n(\theta) < T_\theta) \geq x) dx \\
&\leq \int_0^{T_\theta} \mathbb{P}(Z_n(\theta) \geq x) dx \\
&\leq t_\theta + \int_{t_\theta}^{T_\theta} \mathbb{P}(Z_n(\theta) \geq x) dx \\
&\leq t_\theta + \frac{2n+1}{16} \int_{t_\theta}^{T_\theta} \exp \left\{ -\frac{n}{16} \left(\frac{x}{J_{r,\delta/2}(P_\theta) C_r} \right)^{2/r} \right\} dx \\
&= t_\theta + \frac{r(2n+1) C_r J_{r,\delta/2}(P_\theta)}{16} \left(\frac{4}{\sqrt{n}} \right)^r \int_{\sqrt{\log(\frac{2n+1}{16})}}^\infty e^{-y^2} y^{r-1} dy \\
&\lesssim t_\theta + \frac{r(2n+1) C_r J_{r,\delta/2}(P_\theta)}{16} \left(\frac{4}{\sqrt{n}} \right)^r \int_{\sqrt{\log(\frac{2n+1}{16})}}^\infty e^{-y^2/2} dy,
\end{aligned}$$

where we used the change of variable $y = \frac{\sqrt{n}}{4} \left(\frac{x}{J_{r,\delta/2}(P_\theta) C_r} \right)^{1/r}$, and where the final inequality holds for all n larger than a universal constant depending only on r . It follows that

$$\mathbb{E}[Z_n(\theta) \cdot I(Z_n(\theta) < T_\theta)] \lesssim t_\theta + \frac{J_{r,\delta/2}(P_\theta)}{n^{r/2}} \lesssim J_{r,\delta/2}(P_\theta) \left[\frac{1}{n} \log \left(\frac{2n+1}{16} \right) \right]^{r/2}.$$

Putting this fact together with equation (39), we have by the Fubini-Tonelli Theorem,

$$\begin{aligned}\mathbb{E}[\text{SW}_{r,\delta}^r(P_n, P)] &= \int_{\mathbb{S}^{d-1}} \mathbb{E}[Z_n(\theta)] d\mu(\theta) \\ &\lesssim \text{SJ}_{r,\delta/2}(P) \left(\frac{\log n}{n} \right)^{\frac{r}{2}} + bne^{-c_0 n \delta} / \delta^{r/2}.\end{aligned}$$

The claim now follows since $\mathbb{E}[\text{SW}_{r,\delta}(P_n, P)] \leq \left(\mathbb{E}[\text{SW}_{r,\delta}^r(P_n, P)] \right)^{\frac{1}{r}}$. $\square$

Proof of Proposition 3(ii) As we shall show, the assumption $\text{SW}_{r,\delta}(P, Q) \geq \Gamma$ implies that the deviation

$$\Delta_{nm} = \text{SW}_{r,\delta}(P_n, Q_m) - \text{SW}_{r,\delta}(P, Q)$$

is of same order as

$$D_{nm} = \text{SW}_{r,\delta}^r(P_n, Q_m) - \text{SW}_{r,\delta}^r(P, Q).$$

To bound this quantity, define for all $\theta \in \mathbb{S}^{d-1}$,

$$\begin{aligned}a_\theta &= \min \left\{ F_\theta^{-1}(\delta), F_{\theta,n}^{-1}(\delta), G_\theta^{-1}(\delta), G_{\theta,m}^{-1}(\delta) \right\}, \\ b_\theta &= \max \left\{ F_\theta^{-1}(1-\delta), F_{\theta,n}^{-1}(1-\delta), G_\theta^{-1}(1-\delta), G_{\theta,m}^{-1}(1-\delta) \right\}, \\ M(\theta) &= \max \left\{ |F_\theta^{-1}(\delta/2)|, |F_{\theta,n}^{-1}(\delta/2)|, |F_\theta^{-1}(1-\delta/2)|, |G_\theta^{-1}(1-\delta/2)|, 1 \right\},\end{aligned}$$

and

$$Z_{nm}(\theta) = (b_\theta - a_\theta)^{r-1} \int_\delta^{1-\delta} \left[|F_{\theta,n}^{-1}(u) - F_\theta^{-1}(u)| + |G_{\theta,m}^{-1}(u) - G_\theta^{-1}(u)| \right] du.$$

The bulk of our proof is now contained in the following result.

Lemma 6. *We have,*

$$|D_{nm}| \leq \frac{r}{1-2\delta} \int_{\mathbb{S}^{d-1}} Z_{nm}(\theta) d\mu(\theta).$$

Assume further that $P, Q \in \mathcal{K}_r(b)$. Let $y_0 = \delta/2$. Then, there exists a universal constant $A > 0$ such that for all $y \in (0, y_0]$

$$\sup_{\theta \in \mathbb{S}^{d-1}} \mathbb{P}(|Z_{nm}(\theta)| \geq AM^r(\theta)y) \leq 4 \exp(-2(n \wedge m)y^2).$$

Let $A, y_0 > 0$ be as in Lemma 6. Similarly as in the proof of Proposition 1, let $T_\theta = AM^r(\theta)y_0$. Then, we have,

$$\mathbb{E}[Z_{nm}(\theta)] = \mathbb{E}[Z_{nm}(\theta)I(Z_{nm}(\theta) \geq T_\theta)] + \mathbb{E}[Z_{nm}(\theta)I(Z_{nm}(\theta) < T_\theta)]. \quad (40)$$

Set $s = 1/(1 - 1/\eta)$, where again $\eta = \bar{r}/r$ and $\bar{r} \in (r, r + 1]$, so that by Hölder's inequality, we have uniformly in $\theta \in \mathbb{S}^{d-1}$,

$$\begin{aligned} & \mathbb{E}[Z_{nm}(\theta) \cdot I(Z_{nm}(\theta) \geq T_\theta)] \\ & \leq \|Z_{nm}(\theta)\|_{L^\eta(\mathbb{P})} \|I(Z_{nm}(\theta) \geq T_\theta)\|_{L^s(\mathbb{P})} \\ & \lesssim \left(\|a_\theta\|_{L^{\bar{r}}(\mathbb{P})} + \|b_\theta\|_{L^{\bar{r}}(\mathbb{P})} \right)^{\frac{1}{\eta}} \exp(-2(n \wedge m)y_0^2/s) \\ & \lesssim \left(M^{\bar{r}}(\theta) + \left(\frac{\mathbb{E}|X^\top \theta|^2}{\delta} \right)^{\frac{\bar{r}}{2}} + \left(\frac{\mathbb{E}|Y^\top \theta|^2}{\delta} \right)^{\frac{\bar{r}}{2}} \right)^{\frac{1}{\eta}} \exp(-2(n \wedge m)y_0^2/s) \\ & \lesssim \left(M^r(\theta) + \left(\frac{\mathbb{E}|X^\top \theta|^2}{\delta} \right)^{\frac{r}{2}} + \left(\frac{\mathbb{E}|Y^\top \theta|^2}{\delta} \right)^{\frac{r}{2}} \right) \exp(-2(n \wedge m)y_0^2/s), \quad (41) \end{aligned}$$

where we invoked Lemmas 4 and 6. Now, notice that $\int_{\mathbb{S}^{d-1}} M^r(\theta) d\mu(\theta) \lesssim b/\delta^{r/2}$ by Lemma 3, since $P, Q \in \mathcal{K}_r(b)$. Therefore, integrating both sides of the above display with respect to μ leads to

$$\int_{\mathbb{S}^{d-1}} \mathbb{E}[Z_{nm}(\theta) I(Z_{nm}(\theta) \geq T_\theta)] d\mu(\theta) \lesssim \exp(-c_1(n \wedge m)\delta^2)b/\delta^{r/2},$$

for a constant $c_1 > 0$. We now bound the second term in equation (40). We again use Lemma 6 to obtain

$$\begin{aligned} & \int_{\mathbb{S}^{d-1}} \mathbb{E}[Z_{nm}(\theta) \cdot I(Z_{nm}(\theta) < T_\theta)] d\mu(\theta) \\ & = \int_{\mathbb{S}^{d-1}} \int_0^\infty \mathbb{P}(Z_{nm}(\theta) \cdot I(Z_{nm}(\theta) < T_\theta) \geq x) dx d\mu(\theta) \\ & \leq \int_{\mathbb{S}^{d-1}} \int_0^{T_\theta} \mathbb{P}(Z_{nm}(\theta) \geq x) dx d\mu(\theta) \\ & \leq 4 \int_{\mathbb{S}^{d-1}} \int_0^{T_\theta} \exp \left\{ -2(n \wedge m) \left(\frac{x}{AM^r(\theta)} \right)^2 \right\} dx d\mu(\theta) \\ & \lesssim \frac{1}{\sqrt{n \wedge m}} \int_{\mathbb{S}^{d-1}} M^r(\theta) d\mu(\theta) \lesssim \frac{b}{\delta^{r/2} \sqrt{n \wedge m}}. \end{aligned}$$

Putting this fact together with equation (41), and applying the Fubini-Tonelli Theorem and Lemma 6, we arrive at

$$\begin{aligned} \mathbb{E}|D_{nm}| & \lesssim \frac{1}{1-2\delta} \mathbb{E} \left[\int_{\mathbb{S}^{d-1}} Z_{nm}(\theta) d\mu(\theta) \right] \\ & = \frac{1}{1-2\delta} \int_{\mathbb{S}^{d-1}} \mathbb{E}[Z_{nm}(\theta)] d\mu(\theta) \lesssim \frac{b}{\delta^{r/2}(1-2\delta)\sqrt{n \wedge m}}. \quad (42) \end{aligned}$$

Finally, the numerical inequality $|x^r - y^r| \geq y^{r-1}|x - y|$ for all $x, y > 0$ implies

$$\mathbb{E}|D_{nm}| \geq \text{SW}_{r,\delta}^{r-1}(P, Q) \mathbb{E}|\Delta_{nm}| \geq \Gamma^{r-1} \mathbb{E}|\Delta_{nm}|, \quad (43)$$

so that $\mathbb{E}|\Delta_{nm}| \lesssim \Gamma^{1-r} b(n \wedge m)^{-1/2}/(\delta^{r/2}(1-2\delta))$, as claimed. $\square$

Proof of Proposition 1(ii) Introduce an i.i.d. sample $X'_1, \dots, X'_n \sim P$ independent of $X_1, \dots, X_n$, and let $P_n = \frac{1}{n} \sum_{i=1}^n \delta_{X'_i}$. It follows from convexity of the mapping $(P, Q) \mapsto \text{SW}_{r,\delta}(P, Q)$, similarly as in the proof of Theorem 4.3 of Bobkov and Ledoux (2019), that

$$\mathbb{E}[\text{SW}_{r,\delta}^r(P_n, P)] \leq \mathbb{E}[\text{SW}_{r,\delta}^r(P_n, P'_n)].$$

The claim now follows from equation (42) with $P = Q$, which implies the bound

$$\mathbb{E}[\text{SW}_{r,\delta}^r(P_n, P'_n)] \lesssim \frac{bn^{-1/2}}{\delta^{r/2}(1-2\delta)},$$

so that, $\mathbb{E}[\text{SW}_{r,\delta}(P_n, P'_n)] \lesssim b^{1/r} n^{-1/2r} / (\sqrt{\delta}(1-2\delta)^{1/r})$. $\square$

Proof of Proposition 3(i) The claim is immediate from Proposition 1(i), by the triangle inequality. $\square$

It remains to prove Lemmas 4–6.

B.1. Proof of Lemma 4

We shall make use of Bennett's inequality, which we recall as follows using the notation of Pollard (2002).

Lemma 7 (Bennett's Inequality). *Let $Z_1, \dots, Z_n$ be i.i.d. random variables bounded above by 1, and such that $\mathbb{E}[Z_1] = \mu \in \mathbb{R}$, $\text{Var}[Z_1] = v > 0$. Then,*

$$\mathbb{P} \left\{ \sum_{i=1}^n (Z_i - \mu) \geq x \right\} \leq \exp \left\{ -\frac{x^2}{2W} \psi \left(\frac{x}{W} \right) \right\},$$

for all $x \geq 0$, where $W \geq nv$ is arbitrary, and where for all $t \geq -1$,

$$\psi(t) = \frac{(1+t) \log(1+t) - t}{t^2/2}, \quad \text{for } t \neq 0, \text{ and } \psi(0) = 1.$$

Turning back to the proof, fix $\theta \in \mathbb{S}^{d-1}$. Set $m_\theta = \mathbb{E}(|X^\top \theta|^2)$, and

$$x_\theta = 2[|F_\theta^{-1}(\delta/2)| \vee |F_\theta^{-1}(1-\delta/2)| \vee \ell \sqrt{m_\theta/\delta} \vee 1],$$

for a constant $\ell > 0$ to be determined below. In particular, for all $x \geq x_\theta$,

$$F_\theta(x) \geq 1 - \delta/2, \quad \text{and} \quad F_\theta(-x) \leq \delta/2.$$

Then, for all $x \geq x_\theta$,

$$\begin{aligned} \mathbb{P}(-F_{\theta,n}^{-1}(\delta) > x) &\leq \mathbb{P}(F_{\theta,n}(-x) > \delta) \\ &\leq \mathbb{P}(F_{\theta,n}(-x) - F_\theta(-x) > \delta/2) \end{aligned}$$

$$\leq \exp \left\{ -\frac{(\delta n)^2}{8W_\theta} \psi \left(\frac{n\delta}{2W_\theta} \right) \right\},$$

for any given $W_\theta \geq n \operatorname{Var}[I(X^\top \theta \leq -x)] = nF_\theta(-x)(1 - F_\theta(-x))$, by Bennett's inequality. Now, from Markov's inequality, one also has

$$F_\theta(-x) = \mathbb{P}(X^\top \theta \leq -x) = \mathbb{P}(-X^\top \theta \geq x) \leq \mathbb{P}(|X^\top \theta| \geq x) \leq m_\theta/x^2,$$

thus, we may take $W_\theta = nm_\theta/x^2$. Furthermore,

$$\begin{aligned} \frac{(\delta n)^2}{8W_\theta} \psi \left(\frac{n\delta}{2W_\theta} \right) &= \frac{(\delta n)^2}{8W_\theta} \frac{(1 + \frac{n\delta}{2W_\theta}) \log(1 + \frac{n\delta}{2W_\theta}) - \frac{n\delta}{2W_\theta}}{(\frac{n\delta}{2W_\theta})^2/2} \\ &= W_\theta \left[\left(1 + \frac{n\delta}{2W_\theta} \right) \log \left(1 + \frac{n\delta}{2W_\theta} \right) - \frac{n\delta}{2W_\theta} \right] \\ &\geq \frac{nm_\theta}{x^2} \left[\frac{x^2\delta}{2m_\theta} \log \left(1 + \frac{x^2\delta}{2m_\theta} \right) - \frac{x^2\delta}{2m_\theta} \right] \\ &\geq n \left[\frac{\delta}{2} \log \left(1 + \frac{x^2\delta}{2m_\theta} \right) - \frac{\delta}{2} \right] \\ &\geq \frac{n\delta}{4} \log \left(1 + \frac{x^2\delta}{2m_\theta} \right), \end{aligned}$$

where the last inequality holds for all $x \geq x_\theta$ upon choosing the constant $\ell = \sqrt{2(e^2 - 1)}$ in the definition of x_θ . Therefore, for all such x ,

$$\mathbb{P}(-F_{\theta,n}^{-1}(\delta) > x) \leq \left(1 + \frac{x^2\delta}{2m_\theta} \right)^{-\frac{n\delta}{4}}$$

Applying a similar argument, we obtain that for all $x \geq x_\theta$,

$$\begin{aligned} \mathbb{P}(F_{\theta,n}^{-1}(1 - \delta) > x) &\leq \mathbb{P}(1 - F_{\theta,n}(x) > \delta) \\ &\leq \mathbb{P}(F_\theta(x) - F_{\theta,n}(x) > \delta/2) \leq \left(1 + \frac{x^2\delta}{2m_\theta} \right)^{-\frac{n\delta}{4}}. \end{aligned}$$

Now notice that

$$|F_{\theta,n}^{-1}(1 - \delta)| \leq F_{\theta,n}^{-1}(1 - \delta) \vee (-F_{\theta,n}^{-1}(\delta)),$$

thus we arrive at

$$\mathbb{P}(|F_{\theta,n}^{-1}(1 - \delta)| \geq x) \leq 2 \left(1 + \frac{x^2\delta}{2m_\theta} \right)^{-\frac{n\delta}{4}}, \quad x \geq x_\theta.$$

It follows that

$$\mathbb{E}|F_{\theta,n}^{-1}(1 - \delta)|^{\bar{r}} = \bar{r} \int_0^\infty x^{\bar{r}-1} \mathbb{P}(|F_{\theta,n}^{-1}(1 - \delta)| \geq x) dx$$

$$\begin{aligned}
&\lesssim x_{\bar{\theta}}^{\bar{r}} + \int_{x_{\theta}}^{\infty} x^{\bar{r}-1} \mathbb{P}(|F_{\theta,n}^{-1}(1-\delta)| \geq x) dx \\
&\leq x_{\bar{\theta}}^{\bar{r}} + 2 \int_{x_{\theta}}^{\infty} x^{\bar{r}-1} \left(1 + \frac{x^2 \delta}{2m_{\theta}}\right)^{-\frac{n\delta}{4}} dx \\
&\leq x_{\bar{\theta}}^{\bar{r}} + 2 \int_{\sqrt{m_{\theta}/\delta}}^{\infty} x^{\bar{r}-1} \left(1 + \frac{x^2 \delta}{2m_{\theta}}\right)^{-\frac{n\delta}{4}} dx, \quad (\text{Since } \ell > 1) \\
&= x_{\bar{\theta}}^{\bar{r}} + 2 \left(\frac{m_{\theta}}{\delta}\right)^{\frac{\bar{r}}{2}} \int_1^{\infty} y^{\bar{r}-1} \left(1 + \frac{y^2}{2}\right)^{-\frac{n\delta}{4}} dy \\
&\lesssim x_{\bar{\theta}}^{\bar{r}} + \left(\frac{m_{\theta}}{\delta}\right)^{\frac{\bar{r}}{2}} \int_1^{\infty} y^{\bar{r}-1-\frac{n\delta}{2}} dy \\
&= x_{\bar{\theta}}^{\bar{r}} + \left(\frac{m_{\theta}}{\delta}\right)^{\frac{\bar{r}}{2}} \frac{1}{\frac{n\delta}{2} - \bar{r}} \leq x_{\bar{\theta}}^{\bar{r}} + \left(\frac{m_{\theta}}{\delta}\right)^{\frac{\bar{r}}{2}},
\end{aligned}$$

where we used the assumptions $\delta \geq 2(r+2)/n$ and $\bar{r} \leq r+1$ on the final line of the above display. Therefore, $\mathbb{E}|F_{\theta,n}^{-1}(1-\delta)|^{\bar{r}} \lesssim x_{\bar{\theta}}^{\bar{r}} + (m_{\theta}/\delta)^{\frac{\bar{r}}{2}}$. Upon repeating the same argument for the δ -quantile, we obtain

$$\max_{a \in \{\delta, 1-\delta\}} \mathbb{E}|F_{\theta,n}^{-1}(a)|^{\bar{r}} \lesssim x_{\bar{\theta}}^{\bar{r}} + \left(\frac{m_{\theta}}{\delta}\right)^{\frac{\bar{r}}{2}} \lesssim 1 + \max_{a \in \{\frac{\delta}{2}, 1-\frac{\delta}{2}\}} |F_{\theta}^{-1}(a)|^{\bar{r}} + \left(\frac{m_{\theta}}{\delta}\right)^{\frac{\bar{r}}{2}}.$$

This proves the claim. $\square$

B.2. Proof of Lemma 5

Since $\text{SJ}_{r,\delta/2}(P) < \infty$, it follows from Lemma 2 that F_{θ}^{-1} is absolutely continuous over $[\delta/2, 1-\delta/2]$ for μ -almost every $\theta \in \mathbb{S}^{d-1}$. We fix any such θ throughout the sequel.

To prove the claim, we shall make use of the following analogue of the relative VC inequality described in Example 2 (Bousquet, Boucheron and Lugosi, 2003). For any given $\theta \in \mathbb{S}^{d-1}$ and $\epsilon \in (0, 1)$, we have

$$\mathbb{P}\left(|F_{\theta,n}(x) - F_{\theta}(x)| \leq \nu_{\epsilon,n} \sqrt{F_{\theta}(x)(1 - F_{\theta}(x))}, \forall x \in \mathbb{R}\right) \geq 1 - \epsilon.$$

Notice here that the right-hand side of the inequality within the above probability involves the population CDF F_{θ} rather than $F_{\theta,n}$. Similarly as in Example 2 and Appendix G, the above bound implies that

$$\mathbb{P}\left(F_{\theta}^{-1}(\gamma_{\epsilon,n}(u)) \leq F_{\theta,n}^{-1}(u) \leq F_{\theta}^{-1}(\eta_{\epsilon,n}(u)), \forall u \in (0, 1)\right) \geq 1 - \epsilon,$$

where, for all $u \in (0, 1)$,

$$\begin{aligned}\gamma_{\epsilon,n}(u) &= \frac{2u + \nu_{\epsilon,n}^2 - \nu_{\epsilon,n}\sqrt{\nu_{\epsilon,n}^2 + 4u(1-u)}}{2(1 + \nu_{\epsilon,n}^2)}, \\ \eta_{\epsilon,n}(u) &= \frac{2u + \nu_{\epsilon,n}^2 + \nu_{\epsilon,n}\sqrt{\nu_{\epsilon,n}^2 + 4u(1-u)}}{2(1 + \nu_{\epsilon,n}^2)}.\end{aligned}\tag{44}$$

These functions are invertible over $[0, 1]$, with inverses given by

$$\begin{aligned}\eta_{\epsilon,n}^{-1}(t) &= t - \nu_{\epsilon,n}\sqrt{t(1-t)}, \quad t \in [\eta_{\epsilon,n}(0), \eta_{\epsilon,n}(1)] = \left[\frac{\nu_{\epsilon,n}^2}{1 + \nu_{\epsilon,n}^2}, 1\right], \\ \gamma_{\epsilon,n}^{-1}(t) &= t + \nu_{\epsilon,n}\sqrt{t(1-t)}, \quad t \in [\gamma_{\epsilon,n}(0), \gamma_{\epsilon,n}(1)] = \left[0, \frac{1}{1 + \nu_{\epsilon,n}^2}\right].\end{aligned}\tag{45}$$

We shall further make use of the following elementary Lemma, proven in Appendix B.3.

Lemma 8. Assume $\epsilon \in (0, 1)$ is chosen such that $\nu_{\epsilon,n} \leq y_0 := \sqrt{\delta}/4$. Then, for all $u \in [\delta/2, 1 - \delta/2]$,

$$\frac{\gamma_{\epsilon,n}(u)}{u} \geq \frac{1}{2}, \quad \frac{1 - \eta_{\epsilon,n}(u)}{1 - u} \geq \frac{1}{2}.$$

In particular, for all $x \in [\delta/2, 1 - \delta/2]$, and all $y \in [\gamma_{\epsilon,n}(x), \eta_{\epsilon,n}(x)]$, $\frac{x(1-x)}{y(1-y)} \leq \frac{1}{4}$.

By Lemma 8, the inequalities $\gamma_{\epsilon,n}(\delta) \geq \delta/2$ and $\eta_{\epsilon,n}(1 - \delta) \leq 1 - \delta/2$ hold whenever $\nu_{\epsilon,n} \leq y_0$, and this last inequality is satisfied whenever $\epsilon \geq \epsilon_0 := \frac{2n+1}{16} \exp(-ny_0^2/16)$. For all such ϵ , define the event

$$E_\epsilon \equiv E_\epsilon(\theta) = \left\{ F_\theta^{-1}(\gamma_{\epsilon,n}(u)) \leq F_{\theta,n}^{-1}(u) \leq F_\theta^{-1}(\eta_{\epsilon,n}(u)), \forall u \in [\delta, 1 - \delta] \right\}.$$

Over the event E_ϵ , we have

$$\begin{aligned}Z_n(\theta) &= \frac{1}{1 - 2\delta} \int_\delta^{1-\delta} |F_{\theta,n}^{-1}(u) - F_\theta^{-1}(u)|^r du \\ &\leq \frac{1}{1 - 2\delta} \int_\delta^{1-\delta} [F_\theta^{-1}(\eta_{\epsilon,n}(u)) - F_\theta^{-1}(\gamma_{\epsilon,n}(u))]^r du.\end{aligned}$$

Now, recall that θ was chosen such that F_θ^{-1} is absolutely continuous over $[\delta/2, 1 - \delta/2]$, whence

$$F_\theta^{-1}(\eta_{\epsilon,n}(u)) - F_\theta^{-1}(\gamma_{\epsilon,n}(u)) = \int_{\gamma_{\epsilon,n}(u)}^{\eta_{\epsilon,n}(u)} \frac{dt}{p_\theta(F_\theta^{-1}(t))}.$$

Now, since $\nu_{\epsilon,n} \leq y_0$, we have for all $u \in [\delta, 1 - \delta]$,

$$\eta_{\epsilon,n}(u) - \gamma_{\epsilon,n}(u) = \frac{\nu_{\epsilon,n}\sqrt{\nu_{\epsilon,n}^2 + 4u(1-u)}}{(1 + \nu_{\epsilon,n}^2)}$$

$$\leq 2\nu_{\epsilon,n}\sqrt{u(1-u)} + \nu_{\epsilon,n}^2 \leq 3\nu_{\epsilon,n}\sqrt{u(1-u)}. \quad (46)$$

Using Jensen's inequality, we deduce that, over the event E_ϵ ,

$$\begin{aligned} & (1-2\delta)Z_n(\theta) \\ & \leq \int_{\delta}^{1-\delta} \left(\int_{\gamma_{\epsilon,n}(u)}^{\eta_{\epsilon,n}(u)} \frac{1}{p_\theta(F_\theta^{-1}(t))} dt \right)^r du \\ & \leq \int_{\delta}^{1-\delta} (\eta_{\epsilon,n}(u) - \gamma_{\epsilon,n}(u))^{r-1} \int_{\gamma_{\epsilon,n}(u)}^{\eta_{\epsilon,n}(u)} \left(\frac{1}{p_\theta(F_\theta^{-1}(t))} \right)^r dt du \\ & = \int_{\delta}^{1-\delta} \frac{(\eta_{\epsilon,n}(u) - \gamma_{\epsilon,n}(u))^{r-1}}{(u(1-u))^{r/2}} \int_{\gamma_{\epsilon,n}(u)}^{\eta_{\epsilon,n}(u)} \left(\frac{\sqrt{t(1-t)}}{p_\theta(F_\theta^{-1}(t))} \right)^r \frac{(u(1-u))^{r/2}}{(t(1-t))^{r/2}} dt du \\ & \lesssim \nu_{\epsilon,n}^{r-1} \int_{\delta}^{1-\delta} \frac{1}{\sqrt{u(1-u)}} \int_{\gamma_{\epsilon,n}(u)}^{\eta_{\epsilon,n}(u)} \left(\frac{\sqrt{t(1-t)}}{p_\theta(F_\theta^{-1}(t))} \right)^r dt du \quad (\text{By (46), Lem. 8}) \\ & \leq \nu_{\epsilon,n}^{r-1} \int_{\delta/2}^{1-\delta/2} \left(\int_{\eta_{\epsilon,n}^{-1}(t)}^{\gamma_{\epsilon,n}^{-1}(t)} \frac{1}{\sqrt{u(1-u)}} du \right) \left(\frac{\sqrt{t(1-t)}}{p_\theta(F_\theta^{-1}(t))} \right)^r dt, \end{aligned}$$

where we interchanged the order of integration, and used the fact that $\delta/2 \leq \gamma_{\epsilon,n}(u) \leq \eta_{\epsilon,n}(u) \leq 1-\delta/2$ for all $u \in [\delta, 1-\delta]$. We further have

$$\int_{\eta_{\epsilon,n}^{-1}(t)}^{\gamma_{\epsilon,n}^{-1}(t)} \frac{1}{\sqrt{u(1-u)}} du \leq \frac{\gamma_{\epsilon,n}^{-1}(t) - \eta_{\epsilon,n}^{-1}(t)}{\sqrt{u_t^*(1-u_t^*)}}, \quad \text{for } u_t^* \in \underset{\eta_{\epsilon,n}^{-1}(t) \leq u \leq \gamma_{\epsilon,n}^{-1}(t)}{\operatorname{argmin}} \sqrt{u(1-u)}.$$

By again applying Lemma 8, and equation (45), we obtain

$$\frac{\gamma_{\epsilon,n}^{-1}(t) - \eta_{\epsilon,n}^{-1}(t)}{\sqrt{u_t^*(1-u_t^*)}} \leq \frac{\nu_{\epsilon,n}\sqrt{t(1-t)}}{\sqrt{u_t^*(1-u_t^*)}} \leq \frac{1}{2}\nu_{\epsilon,n}.$$

We have thus shown

$$Z_n(\theta) \lesssim \frac{\nu_{\epsilon,n}^r}{1-2\delta} \int_{\delta/2}^{1-\delta/2} \left(\frac{\sqrt{t(1-t)}}{p_\theta(F_\theta^{-1}(t))} \right)^r dt = \nu_{\epsilon,n}^r J_{r,\delta/2}(P_\theta).$$

Thus, setting $\epsilon = \frac{2n+1}{16} e^{-\frac{ny^2}{16}}$ for any $y \in (0, y_0]$, we have

$$\mathbb{P}\left(Z_n(\theta) \geq C_r y^r J_{r,\delta/2}(P_\theta)\right) \leq \epsilon,$$

for a universal constant $C_r > 0$ depending only on r . $\square$

B.3. Proof of Lemma 8

For all $u \in [\delta/2, 1-\delta/2]$, and all ϵ such that $\nu_{\epsilon,n}^2 \leq \delta/16$, we have

$$\frac{\gamma_{\epsilon,n}(u)}{u} \geq \frac{1}{1+\nu_{\epsilon,n}^2} - \frac{\nu_{\epsilon,n}\sqrt{\nu_{\epsilon,n}^2 + 4u(1-u)} - \nu_{\epsilon,n}^2}{2u(1+\nu_{\epsilon,n}^2)}$$

$$\begin{aligned}
&\geq \frac{1}{1 + \nu_{\epsilon,n}^2} - \frac{\nu_{\epsilon,n}\sqrt{1-u}}{\sqrt{u}(1 + \nu_{\epsilon,n}^2)} \\
&\geq \frac{1}{1 + \nu_{\epsilon,n}^2} - \frac{\sqrt{2}\sqrt{\delta}\sqrt{1-u}}{4\sqrt{\delta}(1 + \nu_{\epsilon,n}^2)} \\
&\geq \frac{1}{1 + \nu_{\epsilon,n}^2} - \frac{1}{2\sqrt{2}(1 + \nu_{\epsilon,n}^2)} \\
&\geq \frac{2\sqrt{2}-1}{2\sqrt{2}(1 + \nu_{\epsilon,n}^2)} \geq \frac{2\sqrt{2}-1}{2\sqrt{2}(1 + 1/16)} \geq 1/2.
\end{aligned}$$

Similarly,

$$\begin{aligned}
\frac{1 - \eta_{\epsilon,n}(u)}{1 - u} &\geq \frac{1}{1 + \nu_{\epsilon,n}^2} - \frac{\nu_{\epsilon,n}^2 + \nu_{\epsilon,n}\sqrt{\nu_{\epsilon,n}^2 + 4u(1-u)}}{2(1-u)(1 + \nu_{\epsilon,n}^2)} \\
&\geq \frac{1}{1 + \nu_{\epsilon,n}^2} - \frac{\nu_{\epsilon,n}^2 + \nu_{\epsilon,n}\sqrt{u(1-u)}}{(1-u)(1 + \nu_{\epsilon,n}^2)} \\
&= \frac{1 - \nu_{\epsilon,n}^2/(1-u)}{1 + \nu_{\epsilon,n}^2} - \frac{\nu_{\epsilon,n}\sqrt{u}}{\sqrt{1-u}(1 + \nu_{\epsilon,n}^2)} \\
&\geq \frac{1 - \nu_{\epsilon,n}^2/\delta}{1 + \nu_{\epsilon,n}^2} - \frac{\sqrt{2}\nu_{\epsilon,n}}{\sqrt{\delta}(1 + \nu_{\epsilon,n}^2)} \\
&\geq \frac{1 - 1/16}{1 + \nu_{\epsilon,n}^2} - \frac{1}{2\sqrt{2}(1 + \nu_{\epsilon,n}^2)} \\
&\geq \frac{2\sqrt{2}(1 - 1/16) - 1}{2\sqrt{2}(1 + 1/16)} \geq 1/2.
\end{aligned}$$

In particular, for all $x \in [\delta/2, 1 - \delta/2]$, and all $y \in [\gamma_{\epsilon,n}(x), \eta_{\epsilon,n}(x)]$,

$$\frac{x(1-x)}{y(1-y)} \leq \frac{x}{\gamma_{\epsilon,n}(x)} \frac{1-x}{1-\eta_{\epsilon,n}(x)} \leq \frac{1}{2}.$$

The claim follows. $\square$

B.4. Proof of Lemma 6

When $r > 1$, we have,

$$\begin{aligned}
&(1 - 2\delta)\text{SW}_{r,\delta}^r(P_n, Q_m) \\
&= \int_{\mathbb{S}^{d-1}} \int_{\delta}^{1-\delta} |F_{\theta,n}^{-1}(u) - G_{\theta,m}^{-1}(u)|^r du d\mu(\theta) \\
&= \int_{\mathbb{S}^{d-1}} \int_{\delta}^{1-\delta} \left\{ |F_{\theta}^{-1}(u) - G_{\theta}^{-1}(u)|^r + r \operatorname{sgn}(\tilde{F}_{\theta,n}^{-1}(u) - \tilde{G}_{\theta,m}^{-1}(u)) \right.
\end{aligned}$$

$$\times |\tilde{F}_{\theta,n}^{-1}(u) - \tilde{G}_{\theta,m}^{-1}(u)|^{r-1} \{ (F_{\theta,n}^{-1}(u) - F_{\theta}^{-1}(u)) - (G_{\theta,m}^{-1}(u) - G_{\theta}^{-1}(u)) \} \} dud\mu(\theta),$$

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

$$\begin{aligned} |D_{nm}| &\leq \frac{r}{1-2\delta} \int_{\mathbb{S}^{d-1}} \int_{\delta}^{1-\delta} |\tilde{F}_{\theta,n}^{-1}(u) - \tilde{G}_{\theta,m}^{-1}(u)|^{r-1} \\ &\quad \times [|F_{\theta,n}^{-1}(u) - F_{\theta}^{-1}(u)| + |G_{\theta,m}^{-1}(u) - G_{\theta}^{-1}(u)|] dud\mu(\theta) \\ &\leq \frac{r}{1-2\delta} \int_{\mathbb{S}^{d-1}} (b_{\theta} - a_{\theta})^{r-1} \\ &\quad \times \int_{\delta}^{1-\delta} [|F_{\theta,n}^{-1}(u) - F_{\theta}^{-1}(u)| + |G_{\theta,m}^{-1}(u) - G_{\theta}^{-1}(u)|] dud\mu(\theta) \\ &= \frac{r}{1-2\delta} \int_{\mathbb{S}^{d-1}} Z_{nm}(\theta) d\mu(\theta). \end{aligned}$$

This proves the first claim when $r > 1$, and the same conclusion holds trivially when $r = 1$ by the triangle inequality. To prove the second claim, given $\theta \in \mathbb{S}^{d-1}$, define the following event for any $t \in (0, \delta/2]$,

$$\begin{aligned} E_t \equiv E_t(\theta) &= \left\{ F_{\theta,n}^{-1}(u-t) \leq F_{\theta}^{-1}(u) \leq F_{\theta,n}^{-1}(u+t), \quad \forall u \in [\delta, 1-\delta] \right\} \\ &\cap \left\{ G_{\theta,m}^{-1}(u-t) \leq G_{\theta}^{-1}(u) \leq G_{\theta,m}^{-1}(u+t), \quad \forall u \in [\delta, 1-\delta] \right\}. \end{aligned}$$

A union bound together with the Dvoretzky-Kiefer-Wolfowitz inequality (Example 1) implies that, for all $t \in (0, \delta/2]$, $\mathbb{P}(E_t) \geq 1 - 4 \exp(-2(n \wedge m)t^2)$. Now, for all such t , the following inequalities hold over E_t ,

$$\begin{aligned} &\int_{\delta}^{1-\delta} |F_{\theta,n}^{-1}(u) - F_{\theta}^{-1}(u)| du \\ &\leq \int_{\delta}^{1-\delta} [F_{\theta}^{-1}(u+t) - F_{\theta}^{-1}(u-t)] du \\ &= \int_{\delta-t}^{1-\delta-t} F_{\theta}^{-1}(u) du - \int_{\delta+t}^{1-\delta+t} F_{\theta}^{-1}(u) du \\ &= \int_{\delta-t}^{\delta+t} F_{\theta}^{-1}(u) du + \int_{1-\delta-t}^{1-\delta+t} F_{\theta}^{-1}(u) du \\ &\leq 2t[|F_{\theta}^{-1}(\delta-t)| + |F_{\theta}^{-1}(\delta+t)| + |F_{\theta}^{-1}(1-\delta+t)| + |F_{\theta}^{-1}(1-\delta-t)|] \\ &\leq 2t[|F_{\theta}^{-1}(\delta/2)| + |F_{\theta}^{-1}(1-\delta/2)|]. \end{aligned}$$

Over E_t , we also have for all $t \in (0, \delta/2]$,

$$\int_{\delta}^{1-\delta} |G_{\theta,n}^{-1}(u) - G_{\theta}^{-1}(u)| du \leq 2t[|G_{\theta}^{-1}(\delta/2)| + |G_{\theta}^{-1}(1-\delta/2)|],$$

and,

$$\begin{aligned} a_\theta &\geq F_\theta^{-1}(\delta - t) \wedge G_\theta^{-1}(\delta - t) \geq F_\theta^{-1}(\delta/2) \wedge G_\theta^{-1}(\delta/2), \\ b_\theta &\leq F_\theta^{-1}(1 - \delta + t) \vee G_\theta^{-1}(1 - \delta + t) \leq F_\theta^{-1}(1 - \delta/2) \vee G_\theta^{-1}(1 - \delta/2). \end{aligned}$$

Combining these facts, we deduce that for a universal constant $A > 0$, we have with probability at least $1 - 4\exp(-2(n \wedge m)t^2)$,

$$|Z_{nm}(\theta)| \leq AtM^r(\theta),$$

as was to be shown. $\square$

Appendix C: Proof of Theorem 1

Throughout the proof, KL denotes the Kullback-Leibler divergence, and χ^2 denotes the χ^2 -divergence. In view of the identity $W_{r,\delta}(P, Q) = W_r(P^\delta, Q^\delta)$ stated in Section 2.1, 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 $(P_0, Q_0), (P_1, Q_1) \in \mathcal{O}(\Gamma; s_1, s_2)$ 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 $\text{SW}_r(P_0, Q_0)$ and $\text{SW}_r(P_1, Q_1)$ are sufficiently different. In particular, if we can show that $\text{KL}(P_0^{\otimes n} \otimes Q_0^{\otimes m}, P_1^{\otimes n} \otimes Q_1^{\otimes m}) \leq \zeta < \infty$, 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,

$$\mathcal{R}_{nm}(\mathcal{O}(\Gamma; s_1, s_2); r) \geq c_\zeta |\text{SW}_r(P_0, Q_0) - \text{SW}_r(P_1, Q_1)|, \quad (47)$$

where $c_\zeta > 0$ is a constant depending only on ζ . 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 $A = (a, 0, \dots, 0) \in \mathbb{R}^d$, and for $g > 0$, we define:

$$\begin{aligned} P_{01} &= \frac{1}{2}\delta_0 + \frac{1}{2}\delta_A, & Q_{01} &= \frac{1}{2}\delta_{gA} + \frac{1}{2}\delta_{(1+g)A} \\ P_{11} &= \left(\frac{1}{2} + \epsilon_n\right)\delta_0 + \left(\frac{1}{2} - \epsilon_n\right)\delta_A, & Q_{11} &= \frac{1}{2}\delta_{gA} + \frac{1}{2}\delta_{(1+g)A}. \end{aligned}$$

- **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 $(X, 0, \dots, 0) \in \mathbb{R}^d$, with X respectively distributed according to the distributions

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

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

- **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 $(X, 0, \dots, 0) \in \mathbb{R}^d$, 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 $P_{04}, P_{14}, Q_{04}, Q_{14} \in \mathcal{P}_r(\mathbb{R}^d)$ be the probability distributions of random vectors of the form $(X, 0, \dots, 0) \in \mathbb{R}^d$, with X respectively distributed according to the distributions

$$P_{04}^{(1)} = U\left(0, s_1^{1/r}\right), \quad Q_{04}^{(1)} = U\left(0, s_2^{1/r}\right),$$

$$P_{14}^{(1)} = (1 - \epsilon_n)U\left(0, s_1^{1/r}\right) + \epsilon_n \delta_{s_1^{1/r}}, \quad Q_{14}^{(1)} = U\left(0, s_2^{1/r}\right).$$

Construction 1 uses pairs of distributions with infinite SJ_r , while Constructions 2-4 use pairs of distributions with finite SJ_r . To compactly state our next result we define several terms,

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

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

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

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

Lemma 9. *There exists a choice of constant $k_r \in (0, 1)$ for which the following statements hold.*

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

$$\text{SW}_r^r(P_{01}, Q_{01}) = \Gamma^r,$$

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

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

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

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

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

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

$$\begin{aligned} \text{SW}_r^r(P_{03}, Q_{03}) &= \frac{s_2 t_r^r |\Delta_{\bar{\beta}}|^r}{r+1}, \\ \text{SW}_r^r(P_{13}, Q_{13}) &\geq \frac{t_r^r s_2}{r+1} \left\{ |\Delta_{\bar{\beta}}|^r + r |\Delta_{\bar{\beta}}|^{r-1} \epsilon_m \right\}. \end{aligned}$$

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

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

$$\begin{aligned} \text{SW}_r^r(P_{04}, Q_{04}) &= \frac{s_1 t_r^r |\Delta_{\beta}|^r}{r+1}, \\ \text{SW}_r^r(P_{14}, Q_{14}) &\geq \frac{t_r^r s_1}{r+1} \left\{ |\Delta_{\beta}|^r + r |\Delta_{\beta}|^{r-1} \epsilon_n \right\}. \end{aligned}$$

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

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

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

Taking this result as given, we can now complete the proof of the Theorem. Using Construction 1 with $\Gamma = 0$, we obtain from equation (47) that

$$\mathcal{R}_{nm}(\mathcal{O}(0; \infty, \infty); r) \geq c_{\zeta} |\text{SW}_r(P_{01}, Q_{01}) - \text{SW}_r(P_{11}, Q_{11})| \gtrsim \epsilon_n^{1/r} \asymp n^{-1/2r}.$$

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

$$\mathcal{R}_{nm}(\mathcal{O}(\Gamma; \infty, \infty); r) \gtrsim \Gamma \left| 1 - \left(1 + \frac{c_1 \epsilon_n}{\Gamma^r} \right)^{\frac{1}{r}} \right| = \frac{c_1 k_r n^{-1/2}}{r \Gamma^{r-1}} (1 + o(1)),$$

by a first-order Taylor expansion of the map $x \mapsto (1+x)^{\frac{1}{r}}$. The fact that Γ is bounded away from zero then implies $\mathcal{R}_{nm}(\mathcal{O}(\Gamma; \infty, \infty); r) \gtrsim n^{-1/2}$ 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: $-1 \leq \Delta_{\beta} \leq -\epsilon_n$ In this case, the condition $\Delta_{\beta} \leq 0$ implies $s_1 \geq s_2$. Since $n \leq m$, it therefore suffices to prove $\mathcal{R}_{nm}(\mathcal{O}(0, s_1, s_2); r) \gtrsim s_1^{1/r} \epsilon_n$. Furthermore, since $\beta \leq 1$, we may invoke Construction 4 to obtain

$$|\text{SW}_r(P_{04}, Q_{04}) - \text{SW}_r(P_{14}, Q_{14})| \geq \frac{s_1^{1/r} t_r |\Delta_{\beta}|}{(r+1)^{1/r}} \left[\left(1 + \frac{r \epsilon_n}{|\Delta_{\beta}|} \right)^{\frac{1}{r}} - 1 \right] \asymp s_1^{1/r} \epsilon_n,$$

where we have used the assumption $|\Delta_{\beta}| \geq \epsilon_n$ in the last order assesment of the above display. This fact together with equation (47) yields the desired lower bound for Case 1.

Case 2: $-\epsilon_n < \Delta_\beta \leq 0$ The inequality $s_1 \geq s_2$ continues to hold, thus it suffices to prove $\mathcal{R}_{nm}(\mathcal{O}(0; s_1, s_2); r) \gtrsim s_1^{1/r} \epsilon_n$. 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 $\mathcal{R}_{nm}(\mathcal{O}(0; s_1, s_2); r) \gtrsim s_2^{1/r} \epsilon_n$. We use Construction 2, and choose $\gamma_2 = s_2$, and $\Delta \in (0, 1]$ to be a constant larger than ϵ_n . We observe that all distributions have SJ_r value at most $s_2 = \min\{s_1, s_2\}$. Furthermore,

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

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

$$|\text{SW}_r(P_{02}, Q_{02}) - \text{SW}_r(P_{12}, Q_{12})| \gtrsim s_2^{1/r} \epsilon_n,$$

and this together with equation (47) yields the desired lower bound for Case 2.

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

$$|\text{SW}_r(P_{03}, Q_{03}) - \text{SW}_r(P_{13}, Q_{13})| \geq \frac{s_2^{1/r} t_r |\Delta_{\bar{\beta}}|}{(r+1)^{1/r}} \left[\left(1 + \frac{\epsilon_m}{|\Delta_{\bar{\beta}}|} \right)^{\frac{1}{r}} - 1 \right] \asymp s_2^{1/r} \epsilon_m,$$

where we have used the assumption $|\Delta_{\bar{\beta}}| \geq \epsilon_m$ in the last order assesment of the above display. This fact together with equation (47) yields the desired lower bound for Case 1.

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

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

For this case, it will thus suffice to prove $\mathcal{R}_{nm}(\mathcal{O}(0; s_1, s_2); r) \gtrsim s_1^{1/r} \epsilon_n$. Since $\Delta_{\bar{\beta}} \leq 0$, we observe that all distributions have SJ_r value 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 9 and we turn our attention to this now.

C.1. Proof of Lemma 9

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 χ^2 divergences in each construction can be computed in closed form. Doing so yields the bounds:

$$\begin{aligned} \text{KL}(P_{0i}^{\otimes n} \otimes Q_{0i}^{\otimes m}, P_{1i}^{\otimes n} \otimes Q_{1i}^{\otimes m}) &\lesssim n\epsilon_n^2, \quad i = 1, 2, 4 \\ \text{KL}(P_{03}^{\otimes n} \otimes Q_{03}^{\otimes m}, P_{13}^{\otimes n} \otimes Q_{13}^{\otimes m}) &\lesssim m\epsilon_m^2. \end{aligned}$$

Together with the definition of ϵ_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 $\mathcal{K}_r(b)$ defined in equation (10). 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 Γ) such that this expression is bounded above by $b^{1/r}$ ensures $P_{01}, Q_{01}, P_{11}, Q_{11} \in \mathcal{K}_r(b)$. We are guaranteed that such a choice exists by using the assumption $\Gamma^r \leq c_r b$, which ensures that Γ 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 $P_{ij}, Q_{ij} \in \mathcal{K}_r(b)$ 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 10. 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,

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

We prove this result in Appendix C.1.1. Taking this result as given, we can now compute the various Sliced Wasserstein distances and SJ_r evaluations.

Computing the sliced Wasserstein distances

- **Construction 1.** For any $\theta \in \mathbb{S}^{d-1}$, 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

$$\begin{aligned} F_{01,\theta}^{-1}(u) &= \begin{cases} 0 \wedge A^\top \theta, & u \in (0, 1/2) \\ 0 \vee A^\top \theta, & u \in [1/2, 1), \end{cases} \\ F_{11,\theta}^{-1}(u) &= \begin{cases} 0 \wedge A^\top \theta, & u \in (0, 1/2 + \epsilon_n) \\ 0 \vee A^\top \theta, & u \in [1/2 + \epsilon_n, 1), \end{cases} \\ G_{01,\theta}^{-1}(u) = G_{11,\theta}^{-1}(u) &= \begin{cases} gA^\top \theta \wedge (1+g)A^\top \theta, & u \in (0, 1/2) \\ gA^\top \theta \vee (1+g)A^\top \theta, & u \in [1/2, 1). \end{cases} \end{aligned}$$

Therefore,

$$\begin{aligned} & \text{SW}_r^r(P_{01}, Q_{01}) \\ &= \int_{\mathbb{S}^{d-1}} \int_0^1 |F_{01,\theta}^{-1}(u) - G_{01,\theta}^{-1}(u)|^r du d\mu(\theta) \\ &= \frac{1}{2} \int_{\{\theta \in \mathbb{S}^{d-1} : A^\top \theta \geq 0\}} |gA^\top \theta|^r d\mu(\theta) \\ &\quad + \frac{1}{2} \int_{\{\theta \in \mathbb{S}^{d-1} : A^\top \theta < 0\}} |A^\top \theta - (1+g)A^\top \theta|^r d\mu(\theta) \\ &= g^r \int_{\mathbb{S}^{d-1}} |A^\top \theta|^r d\mu(\theta) = \Gamma^r. \end{aligned}$$

Furthermore,

$$\begin{aligned} & \text{SW}_r^r(P_{11}, Q_{11}) \\ &= \int_{\mathbb{S}^{d-1}} \int_0^1 |F_{11,\theta}^{-1}(u) - G_{11,\theta}^{-1}(u)|^r du d\mu(\theta) \\ &= \int_{\mathbb{S}^{d-1}} \left(\int_0^{1/2} |0 \wedge A^\top \theta - gA^\top \theta \wedge (1+g)A^\top \theta|^r du \right. \\ &\quad + \int_{1/2}^{1/2+\epsilon_n} |0 \wedge A^\top \theta - gA^\top \theta \vee (1+g)A^\top \theta|^r du \\ &\quad \left. + \int_{1/2+\epsilon_n}^1 |0 \vee A^\top \theta - gA^\top \theta \vee (1+g)A^\top \theta|^r du \right) d\mu(\theta) \\ &= (1 - \epsilon_n) g^r \int_{\mathbb{S}^{d-1}} |A^\top \theta|^r d\mu(\theta) \\ &\quad + \epsilon_n \int_{\mathbb{S}^{d-1}} |0 \wedge A^\top \theta - gA^\top \theta \vee (1+g)A^\top \theta|^r d\mu(\theta) \\ &= (1 - \epsilon_n) g^r I_r^r + \epsilon_n \int_{\mathbb{S}^{d-1}} |0 \wedge A^\top \theta - gA^\top \theta \vee (1+g)A^\top \theta|^r d\mu(\theta) \end{aligned}$$

$$= \Gamma^r + c_1 \epsilon_n,$$

for a positive constant $c_1 > 0$. It follows that $\text{SW}_r(P_{11}, Q_{11}) \geq \text{SW}_r(P_{01}, Q_{01}) \geq \Gamma$, thus $(P_{01}, Q_{01}), (P_{11}, Q_{11}) \in \mathcal{O}(\Gamma; \infty, \infty)$, and

$$|\text{SW}_r(P_{01}, Q_{01}) - \text{SW}_r(P_{11}, Q_{11})| = \left| \Gamma - (\Gamma^r + c_1 \epsilon_n)^{\frac{1}{r}} \right|.$$

- **Construction 2.** We use the first part of Lemma 10, 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), $W_r(\pi_{\theta\#}P_{12}, \pi_{\theta\#}Q_{12}) = |\theta_1|\gamma_2^{1/r}W_r(\nu, \rho)$. Thus,

$$\begin{aligned} \text{SW}_r^r(P_{12}, Q_{12}) &= \int_{\mathbb{S}^{d-1}} W_r^r(\pi_{\theta\#}P_{12}, \pi_{\theta\#}Q_{12})d\mu(\theta) \\ &= \int_{\mathbb{S}^{d-1}} |\theta_1|^r \gamma_2 W_r^r(\nu, \rho)d\mu(\theta) \geq \gamma_2 t_r^r \left[\Delta^r + \frac{r}{4} \Delta^{r-1} \epsilon_n \right], \end{aligned}$$

by Lemma 10. Furthermore, it is easy to show that

$$\text{SW}_r^r(P_{02}, Q_{02}) = \gamma_2 \Delta^r \int_{\mathbb{S}^{d-1}} |\theta_1|^r d\mu(\theta) = \gamma_2 (t_r \Delta)^r.$$

- **Construction 3.** We use the second part of Lemma 10. We set

$$\nu = U(0, \bar{\beta}), \quad \rho = (1 - \epsilon_m)U(0, 1) + \epsilon_m \delta_1.$$

Then, for all $\epsilon \in (0, 1]$,

$$W_r(\nu, \rho) \geq \frac{1}{r+1} \left[|\Delta_{\bar{\beta}}|^r + r \epsilon_m |\Delta_{\bar{\beta}}|^{r-1} \right]$$

We then obtain

$$\begin{aligned} \text{SW}_r^r(P_{13}, Q_{13}) &= \int_{\mathbb{S}^{d-1}} |\theta_1|^r s_2 W_r^r(\nu, \rho)d\mu(\theta) \\ &\geq \frac{t_r^r s_2}{r+1} \left[|\Delta_{\bar{\beta}}|^r + r \epsilon_m |\Delta_{\bar{\beta}}|^{r-1} \right]. \end{aligned}$$

On the other hand, it is easy to see that

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

- **Construction 4.** We again use the second part of Lemma 10, setting

$$\nu = U(0, \beta), \quad \rho = (1 - \epsilon)U(0, 1) + \epsilon_n \delta_1.$$

The rest follows by the same argument as for Construction 3.

Computing the SJ_r evaluations 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

$$\begin{aligned} SJ_r(Q_{02}) &= SJ_r(Q_{12}) = SJ_r(P_{02}) \\ &\leq \int_{\mathbb{S}^{d-1}} \int_0^1 \left(\frac{\sqrt{u(1-u)}}{1/(|\theta_1|\gamma_2^{1/r})} \right)^r dud\mu(\theta) \\ &\leq \int (|\theta_1|\gamma_2^{1/r})^r d\mu(\theta) \leq \gamma_2. \end{aligned}$$

Furthermore,

$$\begin{aligned} SJ_r(Q_{13}) &\leq \int_{\mathbb{S}^{d-1}} \int_0^1 \left(\frac{\sqrt{u(1-u)}}{(1-\epsilon_n)/(|\theta_1|\gamma_2^{1/r})} \right)^r dud\mu(\theta) \\ &\leq \frac{\gamma_2}{(1-\epsilon_n)^r} \int_0^1 [u(1-u)]^{\frac{r}{2}} du. \end{aligned}$$

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 γ_2 .

To complete the proof it remains to prove Lemma 10.

C.1.1. Proof of Lemma 10

We prove each of the two claims in turn.

Proof of Claim (1) Notice that the quantile functions of ν and ρ are respectively given by

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

Thus,

$$\begin{aligned} W_r^r(\nu, \rho) &= \int_0^1 |F^{-1}(u) - G^{-1}(u)|^r du \\ &= \int_0^{\frac{1+\epsilon}{2}} \left| \Delta + u - \frac{u}{1+\epsilon} \right|^r du + \int_{\frac{1+\epsilon}{2}}^1 \left| \Delta + u - \frac{1}{2} - \frac{u-(1+\epsilon)/2}{1-\epsilon} \right|^r du \end{aligned}$$

$$= (\text{I}) + (\text{II}),$$

say. We have,

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

Also,

$$\begin{aligned} (\text{II}) &= \int_{(1+\epsilon)/2}^1 \left(\Delta + \frac{\epsilon}{1-\epsilon} - u \frac{\epsilon}{1-\epsilon} \right)^r du \\ &= -\frac{1-\epsilon}{(r+1)\epsilon} \left\{ \left(\Delta + \frac{\epsilon}{1-\epsilon} - \frac{\epsilon}{1-\epsilon} \right)^{r+1} - \left(\Delta + \frac{\epsilon}{1-\epsilon} - \frac{(1+\epsilon)\epsilon}{2(1-\epsilon)} \right)^{r+1} \right\} \\ &= -\frac{1-\epsilon}{\epsilon(r+1)} \left\{ \Delta^{r+1} - \left(\Delta + \frac{\epsilon}{2} \right)^{r+1} \right\}. \end{aligned}$$

Thus,

$$\begin{aligned} (\text{I}) + (\text{II}) &= \left\{ \left(\Delta + \frac{\epsilon}{2} \right)^{r+1} - \Delta^{r+1} \right\} \left(\frac{1+\epsilon}{\epsilon(r+1)} + \frac{1-\epsilon}{\epsilon(r+1)} \right) \\ &= \frac{2}{\epsilon(r+1)} \left\{ \left(\Delta + \frac{\epsilon}{2} \right)^{r+1} - \Delta^{r+1} \right\} \\ &= \frac{2}{\epsilon(r+1)} \left\{ \Delta^{r+1} + \frac{r+1}{2} \Delta^r \epsilon + \frac{r(r+1)}{8} (\Delta + \tilde{\epsilon})^{r-1} \epsilon^2 - \Delta^{r+1} \right\}, \end{aligned}$$

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

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

and the claim follows. $\square$

Proof of Claim (2) The respective quantile functions of ν and ρ 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,

$$\begin{aligned} \text{SW}_r^r(\nu, \rho) &= \int |F^{-1}(u) - G^{-1}(u)|^r du \\ &= \int_0^{1-\epsilon} \left| \frac{u}{1-\epsilon} - \xi u \right|^r du + \int_{1-\epsilon}^1 |1 - \xi u|^r du \\ &= \int_0^{1-\epsilon} \left[\frac{u}{1-\epsilon} - \xi u \right]^r du + \int_{1-\epsilon}^1 [1 - \xi u]^r du, \quad (\text{Since } \xi \in (0, 1]) \end{aligned}$$

$$\begin{aligned}
&= \frac{1-\epsilon}{r+1}(-\Delta_\xi + \epsilon\xi)^r + \frac{1}{\xi(r+1)} \left[(-\Delta_\xi + \epsilon\xi)^{r+1} - (-\Delta_\xi)^{r+1} \right] \\
&= \frac{1}{r+1} \left[(-\Delta_\xi + \epsilon\xi)^r \left(1 - \epsilon + \frac{-\Delta_\xi + \epsilon\xi}{\xi} \right) - \frac{|\Delta_\xi|^{r+1}}{\xi} \right] \\
&= \frac{1}{\xi(r+1)} \left[(|\Delta_\xi| + \epsilon\xi)^r - |\Delta_\xi|^{r+1} \right] \\
&= \frac{1}{\xi(r+1)} \left[(|\Delta_\xi| + \epsilon\xi)^r - |\Delta_\xi|^{r+1} \right] \\
&\geq \frac{1}{\xi(r+1)} \left[|\Delta_\xi|^r + r\epsilon\xi|\Delta_\xi|^{r-1} - |\Delta_\xi|^{r+1} \right] \\
&= \frac{|\Delta_\xi|^r}{r+1} + \frac{r\epsilon|\Delta_\xi|^{r-1}}{r+1}.
\end{aligned}$$

The claim follows. $\square$

Appendix D: Proof of Theorem 2

By the same argument as in the proof of Proposition 3, it will suffice to prove equation (19). Let $P, Q \in \mathcal{K}_{r,\rho}(b)$. We prove the claim in five steps.

Step 0: Preparation By the same argument as in the proof of Lemma 3, notice that there exists a constant $C_\rho > 0$ such that for any $\theta \in \mathbb{S}^{d-1}$,

$$|F_\theta^{-1}(u)| \vee |G_\theta^{-1}(u)| \leq C_\rho \left(\frac{b_\theta}{u(1-u)} \right)^{\frac{1}{\rho}}, \quad u \in (0, 1),$$

where for $X \sim P, Y \sim Q$ and each $\theta \in \mathbb{S}^{d-1}$,

$$b_\theta = \mathbb{E}[|X^\top \theta|^\rho] + \mathbb{E}[|Y^\top \theta|^\rho].$$

Notice that the assumption $P, Q \in \mathcal{K}_{r,\rho}(b)$ for $\rho > 2r$ implies that $\int b_\theta^{\frac{r}{\rho}} d\mu(\theta) \leq 2b$. In particular, b_θ is finite for almost all $\theta \in \mathbb{S}^{d-1}$. These statements may be applied analogously to the empirical measures P_n and Q_m . Specifically, we have

$$|F_{\theta,n}^{-1}(u)| \vee |G_{\theta,m}^{-1}(u)| \leq C_\rho \left(\frac{b_{\theta,nm}}{u(1-u)} \right)^{\frac{1}{\rho}}, \quad u \in (0, 1),$$

where we set

$$b_{\theta,nm} := \int |x^\top \theta|^\rho dP_n(x) + \int |y^\top \theta|^\rho dQ_m(y).$$

Combining these facts, we have for all $u \in (0, 1)$,

$$|F_\theta^{-1}(u)| \vee |G_\theta^{-1}(u)| \vee |F_{\theta,n}^{-1}(u)| \vee |G_{\theta,m}^{-1}(u)| \leq \psi_\theta(u) := C_\rho \left(\frac{b_\theta + b_{\theta,nm}}{u(1-u)} \right)^{\frac{1}{\rho}}.$$

We suppress the dependence of ψ_θ on n and m for ease of notation, but we emphasize that $\psi_\theta(u)$ is a random variable.

Our proof makes use of a uniform self-normalized concentration inequality for the empirical quantile process, which was introduced in Section B.2, as part of the proof of Lemma 5, and also in Example 2 of the main manuscript. Specifically, for any $\epsilon \in (0, 1)$, let $\gamma_{\epsilon,n}, \eta_{\epsilon,n}$ be the sequences given in equation (44), with inverses given in equation (45), and defined in terms of the quantity

$$\nu_{\epsilon,n} = \sqrt{\frac{16}{n} [\log(16/\epsilon) + \log(2n+1)]},$$

for any given $\epsilon \in (0, 1)$. Recall that for any $\theta \in \mathbb{S}^{d-1}$, the event

$$A_\epsilon = \left\{ u \in (0, 1) : F_\theta^{-1}(\gamma_{\epsilon,n}(u)) \leq F_{\theta,n}^{-1}(u) \leq F_\theta^{-1}(\eta_{\epsilon,n}(u)) \right\}$$

satisfies $\mathbb{P}(A_\epsilon) \geq 1 - \epsilon$.

Step 1: First reduction Apply a similar first-order Taylor expansion as in the proof of Lemma 6, to the cost function $|\cdot|^r$, to deduce that

$$\begin{aligned} & |\text{SW}_r^r(P_n, Q_m) - \text{SW}_r^r(P, Q)| \\ & \leq r \int_{\mathbb{S}^{d-1}} \int_0^1 |\tilde{F}_{\theta,n}^{-1}(u) - \tilde{G}_{\theta,m}^{-1}(u)|^{r-1} \\ & \quad \times \left[|F_{\theta,n}^{-1}(u) - F_\theta^{-1}(u)| + |G_{\theta,m}^{-1}(u) - G_\theta^{-1}(u)| \right] dud\mu(\theta), \end{aligned}$$

for some $\tilde{F}_{\theta,n}^{-1}(u)$ on the segment joining $F_\theta^{-1}(u)$ and $F_{\theta,n}^{-1}(u)$, and some $\tilde{G}_{\theta,m}^{-1}(u)$ on the segment joining $G_\theta^{-1}(u)$ and $G_{\theta,m}^{-1}(u)$. By definition of ψ_θ , we deduce

$$\begin{aligned} & |\text{SW}_r^r(P_n, Q_m) - \text{SW}_r^r(P, Q)| \\ & \lesssim \int_{\mathbb{S}^{d-1}} \int_0^1 \psi_\theta^{r-1}(u) \left[|F_{\theta,n}^{-1}(u) - F_\theta^{-1}(u)| + |G_{\theta,m}^{-1}(u) - G_\theta^{-1}(u)| \right] dud\mu(\theta) \\ & = \int_{\mathbb{S}^{d-1}} \int_0^{1/2} \psi_\theta^{r-1}(u) \left[|F_{\theta,n}^{-1}(u) - F_\theta^{-1}(u)| + |G_{\theta,m}^{-1}(u) - G_\theta^{-1}(u)| \right] dud\mu(\theta) \\ & \quad + \int_{\mathbb{S}^{d-1}} \int_{1/2}^1 \psi_\theta^{r-1}(u) \left[|F_{\theta,n}^{-1}(u) - F_\theta^{-1}(u)| + |G_{\theta,m}^{-1}(u) - G_\theta^{-1}(u)| \right] dud\mu(\theta). \end{aligned}$$

We shall bound the quantity

$$T = \int_{\mathbb{S}^{d-1}} \int_{1/2}^1 \psi_\theta^{r-1}(u) |F_{\theta,n}^{-1}(u) - F_\theta^{-1}(u)| dud\mu(\theta),$$

and a similar argument can be used to bound the remaining terms in the penultimate display.

Throughout the sequel, let $\epsilon \in (0, 1)$ be chosen such that $\nu_{\epsilon,n} \leq 1$. Let

$$\beta > \left(\frac{1}{2} - \frac{r}{\rho} \right)^{-1} > 0, \quad (48)$$

where we recall that $\rho > 2r$. Define $\delta_k = k^{-\beta}/2$ for all $k \geq 1$. Notice that $\delta_1 = 1/2$. Furthermore, let

$$K \equiv K_n = 1 \vee \left\lfloor (c\nu_{\epsilon,n})^{-\frac{\rho}{\beta(\rho-r)}} \right\rfloor$$

for a constant $c \geq 8$ to be specified below. We summarize a few algebraic facts in relation to the sequences $\delta_k, \eta_{\epsilon,n}, \gamma_{\epsilon,n}$, which we prove in Section D.1.

Lemma 11. *There exists a choice of the constant $c \geq 8$, as well as a constant $c_1 > 0$, both possibly depending on β, ρ, r , such that for all $n \geq 1$, and all $\epsilon \in (0, 1)$ for which $\nu_{\epsilon,n} \leq 1$, the following properties hold.*

- (i) $c_1 \nu_{\epsilon,n}^{\frac{\rho}{\rho-r}} \geq \delta_K$. Furthermore, if $K \geq 2$, then $\delta_K \geq (c\nu_{\epsilon,n})^{\frac{\rho}{\rho-r}}/2$.
- (ii) $1 - \gamma_{\epsilon,n}(1 - \delta_k) \leq \delta_k + \frac{\nu_{\epsilon,n}^2}{1 + \nu_{\epsilon,n}^2} + \nu_{\epsilon,n}\sqrt{\delta_k}$ for all $k = 1, \dots, K$.
- (iii) $\frac{1 - \delta_k}{1 + \nu_{\epsilon,n}^2} \leq \eta_{\epsilon,n}(1 - \delta_k) \leq 1 - \delta_k + \nu_{\epsilon,n}^2 + \nu_{\epsilon,n}\sqrt{\delta_k}$, for all $k = 1, \dots, K$.
- (iv) $\eta_{\epsilon,n}(1 - \delta_k) \leq \gamma_{\epsilon,n}(1 - \delta_{k+1})$, for all $k = 1, \dots, K - 1$, if $K \geq 2$.

With these facts in place, consider the decomposition

$$T = \int_{\mathbb{S}^{d-1}} \int_{1/2}^1 \psi_{\theta}^{r-1}(u) |F_{\theta,n}^{-1}(u) - F_{\theta}^{-1}(u)| du d\mu(\theta) = \int_{\mathbb{S}^{d-1}} T_{\theta} d\mu(\theta), \quad (49)$$

where we recall that $\delta_1 = 1/2$, and we set

$$T_{\theta} = \int_{1/2}^1 |F_{\theta,n}^{-1}(u) - F_{\theta}^{-1}(u)| \psi_{\theta}^{r-1}(u) du = T_{\theta,K}^* + \sum_{k=1}^{K-1} T_{\theta,k}, \quad \theta \in \mathbb{S}^{d-1}, \quad (50)$$

$$T_{\theta,k} = \int_{1-\delta_k}^{1-\delta_{k+1}} |F_{\theta,n}^{-1}(u) - F_{\theta}^{-1}(u)| \psi_{\theta}^{r-1}(u) du, \quad k = 1, \dots, K-1, \quad (51)$$

$$T_{\theta,K}^* = \int_{1-\delta_K}^1 |F_{\theta,n}^{-1}(u) - F_{\theta}^{-1}(u)| \psi_{\theta}^{r-1}(u) du. \quad (52)$$

In the following two steps, we bound the preceding two terms for any fixed $\theta \in \mathbb{S}^{d-1}$ and $k = 1, \dots, K-1$. The symbol “ $\lesssim$ ” will always hide constants which do not depend on θ and k .

Step 2: Bounding $T_{\theta,K}^*$ We have,

$$T_{\theta,K}^* = \int_{1-\delta_K}^1 |F_{\theta,n}^{-1}(u) - F_{\theta}^{-1}(u)| \psi_{\theta}^{r-1}(u) du \lesssim \int_{1-\delta_K}^1 \psi_{\theta}^r(u) du$$

$$\begin{aligned}
&\lesssim (b_\theta + b_{\theta,nm})^{\frac{r}{\rho}} \int_{1-\delta_K}^1 (1-u)^{-\frac{r}{\rho}} du \\
&\lesssim (b_\theta + b_{\theta,nm})^{\frac{r}{\rho}} \delta_K^{1-\frac{r}{\rho}} \lesssim (b_\theta + b_{\theta,nm})^{\frac{r}{\rho}} \nu_{\epsilon,n},
\end{aligned}$$

where the final inequality follows by Lemma 11(i).

Step 3: Bounding $T_{\theta,k}$ in probability The bulk of our work will now go into bounding $T_{\theta,k}$, for any given $\theta \in \mathbb{S}^{d-1}$ and $k = 1, \dots, K-1$. Notice that this case is vacuous when $K = 1$. The following derivations are performed over the event A_ϵ . By its definition, we have for any $k = 1, \dots, K-1$,

$$\begin{aligned}
T_{\theta,k} &\leq \int_{1-\delta_k}^{1-\delta_{k+1}} [F_\theta^{-1}(\eta_{\epsilon,n}(u)) - F_\theta^{-1}(\gamma_{\epsilon,n}(u))] \psi_\theta^{r-1}(u) du \\
&\leq \psi_\theta^{r-1}(1-\delta_{k+1}) \int_{1-\delta_k}^{1-\delta_{k+1}} [F_\theta^{-1}(\eta_{\epsilon,n}(u)) - F_\theta^{-1}(\gamma_{\epsilon,n}(u))] du \\
&= \psi_\theta^{r-1}(1-\delta_{k+1}) \left[\int_{\eta_{\epsilon,n}(1-\delta_k)}^{\eta_{\epsilon,n}(1-\delta_{k+1})} F_\theta^{-1}(u) \frac{\partial \eta_{\epsilon,n}^{-1}(u)}{\partial u} du \right. \\
&\quad \left. - \int_{\gamma_{\epsilon,n}(1-\delta_k)}^{\gamma_{\epsilon,n}(1-\delta_{k+1})} F_\theta^{-1}(u) \frac{\partial \gamma_{\epsilon,n}^{-1}(u)}{\partial u} du \right] \\
&= \psi_\theta^{r-1}(1-\delta_{k+1})(A_{\theta,k} + B_{\theta,k}), \tag{53}
\end{aligned}$$

where

$$\begin{aligned}
A_{\theta,k} &= \int_{\eta_{\epsilon,n}(1-\delta_k)}^{\eta_{\epsilon,n}(1-\delta_{k+1})} F_\theta^{-1}(u) du - \int_{\gamma_{\epsilon,n}(1-\delta_k)}^{\gamma_{\epsilon,n}(1-\delta_{k+1})} F_\theta^{-1}(u) du, \\
B_{\theta,k} &= \int_{\eta_{\epsilon,n}(1-\delta_k)}^{\eta_{\epsilon,n}(1-\delta_{k+1})} F_\theta^{-1}(u) \left(\frac{\partial \eta_{\epsilon,n}^{-1}(u)}{\partial u} - 1 \right) du \\
&\quad - \int_{\gamma_{\epsilon,n}(1-\delta_k)}^{\gamma_{\epsilon,n}(1-\delta_{k+1})} F_\theta^{-1}(u) \left(\frac{\partial \gamma_{\epsilon,n}^{-1}(u)}{\partial u} - 1 \right) du.
\end{aligned}$$

Step 3.1: Bounding $A_{\theta,k}$. Consider the decomposition

$$A_{\theta,k} = \left(\int_{\eta_{\epsilon,n}(1-\delta_k)}^{\gamma_{\epsilon,n}(1-\delta_{k+1})} + \int_{\gamma_{\epsilon,n}(1-\delta_{k+1})}^{\eta_{\epsilon,n}(1-\delta_{k+1})} - \int_{\gamma_{\epsilon,n}(1-\delta_k)}^{\eta_{\epsilon,n}(1-\delta_k)} - \int_{\eta_{\epsilon,n}(1-\delta_k)}^{\gamma_{\epsilon,n}(1-\delta_{k+1})} \right) F_\theta^{-1}(u) du.$$

Using Lemma 11(iv) and the fact that $\eta_{\epsilon,n} \geq \gamma_{\epsilon,n}$, the four lower bounds of integration in the above display are less than their respective upper bounds. Therefore,

$$A_{\theta,k} = \left(\int_{\gamma_{\epsilon,n}(1-\delta_{k+1})}^{\eta_{\epsilon,n}(1-\delta_{k+1})} - \int_{\gamma_{\epsilon,n}(1-\delta_k)}^{\eta_{\epsilon,n}(1-\delta_k)} \right) F_\theta^{-1}(u) du$$

$$\begin{aligned}
&\leq \left(\int_{\gamma_{\epsilon,n}(1-\delta_{k+1})}^{\eta_{\epsilon,n}(1-\delta_{k+1})} + \int_{\gamma_{\epsilon,n}(1-\delta_k)}^{\eta_{\epsilon,n}(1-\delta_k)} \right) \psi_\theta(u) du \\
&\leq \psi_\theta(\eta_{\epsilon,n}(1-\delta_{k+1})) \left[(\eta_{\epsilon,n}(1-\delta_{k+1}) - \gamma_{\epsilon,n}(1-\delta_{k+1})) \right. \\
&\quad \left. + (\eta_{\epsilon,n}(1-\delta_k) - \gamma_{\epsilon,n}(1-\delta_k)) \right] \\
&\leq \psi_\theta(\eta_{\epsilon,n}(1-\delta_{k+1})) \left[\frac{\nu_{\epsilon,n} \sqrt{\nu_{\epsilon,n}^2 + 4\delta_{k+1}}}{1 + \nu_{\epsilon,n}^2} + \frac{\nu_{\epsilon,n} \sqrt{\nu_{\epsilon,n}^2 + 4\delta_k}}{1 + \nu_{\epsilon,n}^2} \right] \\
&\lesssim \psi_\theta(\eta_{\epsilon,n}(1-\delta_{k+1})) \left[\nu_{\epsilon,n}^2 + \nu_{\epsilon,n} \sqrt{\delta_k} \right].
\end{aligned}$$

By Lemma 11(i), we deduce that

$$A_{\theta,k} \lesssim \psi_\theta(\eta_{\epsilon,n}(1-\delta_{k+1})) \nu_{\epsilon,n} \left[\delta_K^{\frac{\rho-r}{\rho}} + \delta_k^{\frac{1}{2}} \right] \lesssim \psi_\theta(\eta_{\epsilon,n}(1-\delta_{k+1})) \nu_{\epsilon,n} \sqrt{\delta_k},$$

where we used the assumption $\rho > 2r$. We next turn to bounding $B_{\theta,k}$.

Step 3.2: Bounding $B_{\theta,k}$ We have,

$$\begin{aligned}
\frac{\partial}{\partial u} \eta_{\epsilon,n}^{-1}(u) &= 1 - \frac{\nu_{\epsilon,n}}{2} \left[\sqrt{\frac{1-u}{u}} - \sqrt{\frac{u}{1-u}} \right], \quad u \in [\eta_{\epsilon,n}(1/2), \eta_{\epsilon,n}(1)], \\
\frac{\partial}{\partial u} \gamma_{\epsilon,n}^{-1}(u) &= 1 + \frac{\nu_{\epsilon,n}}{2} \left[\sqrt{\frac{1-u}{u}} - \sqrt{\frac{u}{1-u}} \right], \quad u \in [\gamma_{\epsilon,n}(1/2), \gamma_{\epsilon,n}(1)].
\end{aligned}$$

Since $\nu_{\epsilon,n} \leq 1$, notice that $\gamma_{\epsilon,n}(1/2)$ and $\eta_{\epsilon,n}(1/2)$ are bounded below by a positive universal constant. Therefore, in the above display, the first terms in brackets are bounded above by positive universal constants, uniformly over the stated ranges, leading to

$$\begin{aligned}
\left| \frac{\partial}{\partial u} \eta_{\epsilon,n}^{-1}(u) - 1 \right| &\lesssim \frac{\nu_{\epsilon,n}}{2} \sqrt{\frac{1}{1-u}}, \quad u \in [\eta_{\epsilon,n}(1/2), \eta_{\epsilon,n}(1)], \\
\left| \frac{\partial}{\partial u} \gamma_{\epsilon,n}^{-1}(u) - 1 \right| &\lesssim \frac{\nu_{\epsilon,n}}{2} \sqrt{\frac{1}{1-u}}, \quad u \in [\gamma_{\epsilon,n}(1/2), \gamma_{\epsilon,n}(1)].
\end{aligned}$$

Deduce that,

$$\begin{aligned}
B_{\theta,k} &= \int_{\eta_{\epsilon,n}(1-\delta_k)}^{\eta_{\epsilon,n}(1-\delta_{k+1})} F_\theta^{-1}(u) \left(\frac{\partial \eta_{\epsilon,n}^{-1}(u)}{\partial u} - 1 \right) du \\
&\quad - \int_{\gamma_{\epsilon,n}(1-\delta_k)}^{\gamma_{\epsilon,n}(1-\delta_{k+1})} F_\theta^{-1}(u) \left(\frac{\partial \gamma_{\epsilon,n}^{-1}(u)}{\partial u} - 1 \right) du \\
&\lesssim \nu_{\epsilon,n} \int_{\eta_{\epsilon,n}(1-\delta_k)}^{\eta_{\epsilon,n}(1-\delta_{k+1})} |F_\theta^{-1}(u)| \sqrt{\frac{1}{1-u}} \frac{du}{2}
\end{aligned}$$

$$\begin{aligned}
& + \nu_{\epsilon,n} \int_{\gamma_{\epsilon,n}(1-\delta_k)}^{\gamma_{\epsilon,n}(1-\delta_{k+1})} |F_\theta^{-1}(u)| \sqrt{\frac{1}{1-u}} \frac{du}{2} \\
& \leq 2\nu_{\epsilon,n} \psi_\theta(\eta_{\epsilon,n}(1-\delta_{k+1})) \sqrt{1-\gamma_{\epsilon,n}(1-\delta_k)} \\
& \lesssim \nu_{\epsilon,n} \psi_\theta(\eta_{\epsilon,n}(1-\delta_{k+1})) \sqrt{\delta_k + \nu_{\epsilon,n}^2 + \nu_{\epsilon,n} \sqrt{\delta_k}} \quad (\text{By Lemma 11(ii)}) \\
& \lesssim \nu_{\epsilon,n} \psi_\theta(\eta_{\epsilon,n}(1-\delta_{k+1})) \left[\sqrt{\delta_k} + \delta_k^{\frac{\rho-r}{\rho}} + \delta_k^{\frac{\rho-r}{2\rho} + \frac{1}{4}} \right] \quad (\text{By Lemma 11(i)}) \\
& \lesssim \nu_{\epsilon,n} \psi_\theta(\eta_{\epsilon,n}(1-\delta_{k+1})) \sqrt{\delta_k},
\end{aligned}$$

where we again used the assumption $\rho > 2r$.

Step 4: Bounding T_θ in probability Combine the conclusions of Steps 3.1 and 3.2 with equation (53) to deduce

$$\begin{aligned}
& \sum_{k=1}^{K-1} T_{\theta,k} \\
& \lesssim \nu_{\epsilon,n} \sum_{k=1}^{K-1} \sqrt{\delta_k} \psi_\theta^{r-1}(1-\delta_{k+1}) \psi_\theta(\eta_{\epsilon,n}(1-\delta_{k+1})) \\
& \lesssim \nu_{\epsilon,n} \sum_{k=1}^{K-1} \sqrt{\delta_k} \psi_\theta^r(\eta_{\epsilon,n}(1-\delta_{k+1})) \\
& \lesssim \nu_{\epsilon,n} (b_\theta + b_{\theta,nm})^{\frac{r}{\rho}} \sum_{k=1}^{K-1} \sqrt{\delta_k} [1 - \eta_{\epsilon,n}(1-\delta_{k+1})]^{-\frac{r}{\rho}} \\
& \leq \nu_{\epsilon,n} (b_\theta + b_{\theta,nm})^{\frac{r}{\rho}} \sum_{k=1}^{K-1} \sqrt{\delta_k} \left[\delta_k - \nu_{\epsilon,n}^2 - \nu_{\epsilon,n} \sqrt{\delta_k} \right]^{-\frac{r}{\rho}} \quad (\text{By Lem. 11(iii)}) \\
& \leq \nu_{\epsilon,n} (b_\theta + b_{\theta,nm})^{\frac{r}{\rho}} \sum_{k=1}^{K-1} \sqrt{\delta_k} \left[\delta_k - \left(\delta_K^{\frac{2(\rho-r)}{\rho}} + \delta_K^{\frac{\rho-r}{\rho}} \delta_k^{\frac{1}{2}} \right) / 4 \right]^{-\frac{r}{\rho}} \quad (\text{By Lem. 11(i)}) \\
& \lesssim \nu_{\epsilon,n} (b_\theta + b_{\theta,nm})^{\frac{r}{\rho}} \sum_{k=1}^{K-1} \delta_k^{\frac{1}{2} - \frac{r}{\rho}},
\end{aligned}$$

where we again used the fact that $\rho > 2r$ on the final line. By definition of β in equation (48), the sequence $\delta_k^{\frac{1}{2} - \frac{r}{\rho}}$ is summable, thus the summation in the final line of the above display is bounded above by a finite constant depending only on r, ρ, β . Combine this fact with the conclusion of Step 2 to deduce that, for a constant $C = C(\beta, \rho, r)$, we have

$$T_\theta \leq C(b_\theta + b_{\theta,nm})^{\frac{r}{\rho}} \nu_{\epsilon,n}, \quad (54)$$

over the high-probability event A_ϵ , for all ϵ such that $\nu_{\epsilon,n} \leq 1$.

Step 4: Bounding T_θ in expectation We now wish to turn the preceding display into a bound on the expectation $\mathbb{E}[T_\theta]$. Notice that $\mathbb{E}[b_{\theta,nm}] = b_\theta$, thus by Markov's inequality, we have for all $y > 0$,

$$\mathbb{P}(b_{\theta,nm}^{r/\rho} \geq b_\theta^{r/\rho} y) \leq \mathbb{E}[b_{\theta,nm}] y^{-\rho/r} / b_\theta \leq y^{-\rho/r}.$$

Furthermore, by inverting the definition of $\nu_{\epsilon,n}$ in terms of ϵ , equation (54) implies that for all $0 < u \leq 1$,

$$\mathbb{P}\left(T_\theta \geq Cu(b_\theta + b_{\theta,nm})^{r/\rho}\right) \leq \frac{2n+1}{16} \exp\left(-\frac{nu^2}{16}\right).$$

Combine the preceding two displays to deduce that for all $0 < u \leq 1$,

$$\begin{aligned} & \mathbb{P}\left\{T_\theta > Cub_\theta^{r/\rho} + Cu(b_\theta/n)^{r/\rho} \exp\left(\frac{(r/\rho)nu^2}{16}\right)\right\} \\ &= \mathbb{P}\left\{T_\theta > Cub_\theta^{\frac{r}{\rho}} + Cu\left(\frac{b_\theta}{n}\right)^{\frac{r}{\rho}} e^{\frac{(r/\rho)nu^2}{16}}, b_{\theta,nm}^{r/\rho} \leq (b_\theta/n)^{r/\rho} e^{\frac{(r/\rho)nu^2}{16}}\right\} \\ & \quad + \mathbb{P}\left\{T_\theta > Cub_\theta^{r/\rho} + Cu(b_\theta/n)^{r/\rho} e^{\frac{(r/\rho)nu^2}{16}}, b_{\theta,nm}^{r/\rho} > (b_\theta/n)^{r/\rho} e^{\frac{(r/\rho)nu^2}{16}}\right\} \\ &\leq \mathbb{P}\left\{T_\theta > Cu(b_\theta^{\frac{r}{\rho}} + b_{\theta,nm}^{\frac{r}{\rho}})\right\} + \mathbb{P}\left\{b_{\theta,nm}^{r/\rho} > (b_\theta/n)^{r/\rho} \exp\left(\frac{(r/\rho)nu^2}{16}\right)\right\} \\ &\leq \mathbb{P}\left\{T_\theta > Cu(b_\theta + b_{\theta,nm})^{\frac{r}{\rho}}\right\} + \mathbb{P}\left\{b_{\theta,nm}^{\frac{r}{\rho}} > (b_\theta/n)^{r/\rho} \exp\left(\frac{(r/\rho)nu^2}{16}\right)\right\} \\ &\lesssim n \exp\left(-\frac{nu^2}{16}\right). \end{aligned}$$

Let $f_n(u) = Cub_\theta^{r/\rho} + Cu(b_\theta/n)^{r/\rho} \exp((r/\rho)nu^2/16)$. f_n is strictly increasing over $\mathbb{R}_+$, thus it is invertible with inverse f_n^{-1} . Notice that $f_n^{-1}(0) = 0$ and $f_n^{-1}(u) \rightarrow \infty$ as $u \rightarrow \infty$. Furthermore,

$$f'_n(u) \lesssim b_\theta^{r/\rho} + (b_\theta/n)^{r/\rho} \left[\exp((r/\rho)nu^2/16) + nu^2 \exp((r/\rho)nu^2/16) \right].$$

Now, let $t_n = \sqrt{\frac{16 \log n}{n}}$ and let n be sufficiently large to ensure $t_n \leq 1$. We have,

$$\begin{aligned} & \mathbb{E}\left[T_\theta \cdot I(T_\theta \leq f_n(1))\right] \\ &= \int_0^{f_n(1)} \mathbb{P}(T_\theta \geq x) dx \\ &\leq f_n(t_n) + \int_{f_n(t_n)}^{f_n(1)} \mathbb{P}(T_\theta \geq x) dx \\ &= f_n(t_n) + \int_{t_n}^1 \mathbb{P}(T_\theta \geq f(u)) f'(u) du \end{aligned}$$

$$\begin{aligned}
&\lesssim f_n(t_n) + b_\theta^{r/\rho} n \int_{t_n}^1 \exp(-nu^2/16) du \\
&\quad + \left(\frac{b_\theta}{n}\right)^{r/\rho} n \int_{t_n}^1 \exp(-n(1-r/\rho)u^2/16) du \\
&\quad + \left(\frac{b_\theta}{n}\right)^{r/\rho} n^2 \int_{t_n}^1 u^2 \exp(-n(1-r/\rho)u^2/16) du \\
&\lesssim f_n(t_n) + b_\theta^{r/\rho} \sqrt{n} \exp(-nt_n^2/16) \\
&\quad + \left(\frac{b_\theta}{n}\right)^{r/\rho} \sqrt{n}(1 + \sqrt{nt_n}) \exp(-n(1-r/\rho)t_n^2/16),
\end{aligned}$$

where the bound on the final term can be obtained by integration by parts. Thus,

$$\begin{aligned}
&\mathbb{E}[T_\theta \cdot I(T_\theta \leq f_n(1))] \\
&\lesssim b_\theta^{\frac{r}{\rho}} \sqrt{\frac{\log n}{n}} + \frac{b_\theta^{r/\rho}}{\sqrt{n}} + \left(\frac{b_\theta}{n}\right)^{r/\rho} \sqrt{n \log n} \cdot n^{\frac{r}{\rho}-1} \lesssim b_\theta^{\frac{r}{\rho}} \sqrt{\frac{\log n}{n}}.
\end{aligned}$$

Finally, in order to control $\mathbb{E}[T_\theta \cdot I(T_\theta > f_n(1))]$, we use the naive bound

$$\begin{aligned}
T_\theta &= \int_{1/2}^1 |F_{\theta,n}^{-1}(u) - F_\theta^{-1}(u)| \psi_\theta^{r-1}(u) du \\
&\lesssim \int_{1/2}^1 \psi_\theta^r(u) du \lesssim (b_{\theta,nm} + b_\theta)^{\frac{r}{\rho}} \int_{1/2}^1 \frac{du}{(1-u)^{r/\rho}} \lesssim (b_{\theta,nm} + b_\theta)^{\frac{r}{\rho}}.
\end{aligned}$$

Using the inequality $2r < \rho$ and Jensen's inequality, we deduce

$$\mathbb{E}[T_\theta^2] \lesssim \mathbb{E}[(b_\theta + b_{\theta,nm})^{2r/\rho}] = b_\theta^{\frac{2r}{\rho}} + \mathbb{E}[b_{\theta,nm}]^{\frac{2r}{\rho}} = 2b_\theta^{\frac{2r}{\rho}}.$$

Thus, using the Cauchy-Schwarz inequality, we arrive at

$$\mathbb{E}[T_\theta \cdot I(T_\theta > f_n(1))] \leq \sqrt{\mathbb{E}[T_\theta^2] \mathbb{P}(T_\theta > f_n(1))} \lesssim b_\theta^{\frac{r}{\rho}} \sqrt{n} \exp(-n/32).$$

We deduce that

$$\mathbb{E}[T_\theta] = \mathbb{E}[T_\theta \cdot I(T_\theta > f_n(1))] + \mathbb{E}[T_\theta \cdot I(T_\theta \leq f_n(1))] \lesssim b_\theta^{\frac{r}{\rho}} (\log n/n)^{1/2}.$$

Step 5: Conclusion By the Fubini-Tonelli Theorem and the nonnegativity of T_θ , we deduce from the above display that,

$$\begin{aligned}
\mathbb{E}[T] &= \mathbb{E}\left[\int_{\mathbb{S}^{d-1}} T_\theta d\mu(\theta)\right] \\
&= \int_{\mathbb{S}^{d-1}} \mathbb{E}[T_\theta] d\mu(\theta) \lesssim (\log n/n)^{1/2} \int_{\mathbb{S}^{d-1}} b_\theta^{r/\rho} d\mu(\theta) \leq b(\log n/n)^{1/2},
\end{aligned}$$

where the final inequality follows from the assumption that $P \in \mathcal{K}_{r,\rho}(b)$. The claim follows. $\square$

D.1. Proof of Lemma 11

Part (i) is trivial from the definition of K . To prove parts (ii)–(iv), note that for all $k \geq 1$,

$$\begin{aligned}\gamma_{\epsilon,n}(1 - \delta_k) &= \frac{2(1 - \delta_k) + \nu_{\epsilon,n}^2 - \nu_{\epsilon,n}\sqrt{\nu_{\epsilon,n}^2 + 4\delta_k(1 - \delta_k)}}{2(1 + \nu_{\epsilon,n}^2)} \\ &\geq \frac{1 - \delta_k}{1 + \nu_{\epsilon,n}^2} - \nu_{\epsilon,n}\sqrt{\delta_k}.\end{aligned}\quad (55)$$

Therefore,

$$\gamma_{\epsilon,n}(1 - \delta_k) \geq 1 - \delta_k - \frac{\nu_{\epsilon,n}^2}{1 + \nu_{\epsilon,n}^2} - \nu_{\epsilon,n}\sqrt{\delta_k},$$

which proves claim (ii). Furthermore,

$$\begin{aligned}\frac{1 - \delta_k}{1 + \nu_{\epsilon,n}^2} \leq \eta_{\epsilon,n}(1 - \delta_k) &= \frac{2(1 - \delta_k) + \nu_{\epsilon,n}^2 + \nu_{\epsilon,n}\sqrt{\nu_{\epsilon,n}^2 + 4\delta_k(1 - \delta_k)}}{2(1 + \nu_{\epsilon,n}^2)} \\ &\leq (1 - \delta_k) + \frac{1}{2} \left[2\nu_{\epsilon,n}^2 + \nu_{\epsilon,n}\sqrt{4\delta_k} \right] \\ &= (1 - \delta_k) + \nu_{\epsilon,n}^2 + \nu_{\epsilon,n}\sqrt{\delta_k},\end{aligned}\quad (56)$$

which proves claim (iii). To prove claim (iv), it follows from equations (55)–(56) that it suffices to show

$$\frac{1 - \delta_{k+1}}{1 + \nu_{\epsilon,n}^2} - \nu_{\epsilon,n}\sqrt{\delta_{k+1}} \geq (1 - \delta_k) + \nu_{\epsilon,n}^2 + \nu_{\epsilon,n}\sqrt{\delta_k}.$$

This assertion is equivalent to

$$\delta_k - \delta_{k+1} \geq \nu_{\epsilon,n} \left[\sqrt{\delta_{k+1}} + \sqrt{\delta_k} \right] + \nu_{\epsilon,n}^2 \left[2 - \delta_k + \nu_{\epsilon,n}^2 + \nu_{\epsilon,n}\sqrt{\delta_k} + \nu_{\epsilon,n}\sqrt{\delta_{k+1}} \right],$$

which, in turn, will be satisfied if the following inequality holds,

$$\delta_k - \delta_{k+1} \geq 2\nu_{\epsilon,n}\sqrt{\delta_k} + 5\nu_{\epsilon,n}^2.$$

Using a first-order Taylor expansion of the map $x \mapsto x^{-\beta}$, notice that $\delta_k - \delta_{k+1} \geq \beta(k+1)^{-(1+\beta)}/2$. This fact together with property (i) implies that it is enough to show

$$\frac{\beta}{2}(k+1)^{-(1+\beta)} \geq \frac{2(2\delta_K)^{\frac{\rho-r}{\rho}}}{c} \frac{k^{-\frac{\beta}{2}}}{\sqrt{2}} + \frac{5}{c^2} (2\delta_K)^{2\frac{\rho-r}{\rho}},$$

for which, in turn, it suffices to show

$$\frac{\beta}{2}(k+1)^{-(1+\beta)} \geq \frac{2k^{-\beta(\frac{3}{2}-\frac{r}{\rho})}}{c\sqrt{2}} + \frac{5}{c^2} k^{-2\beta\frac{\rho-r}{\rho}}.$$

By definition of β in condition (48), we have

$$1 + \beta < \beta \left(\frac{3}{2} - \frac{r}{\rho} \right) \vee 2\beta \frac{\rho - r}{\rho},$$

thus for a sufficiently large choice of c , depending only on β, ρ, r , the penultimate display holds for all $k \geq 1$. The claim follows. $\square$

Appendix E: Proof of Theorem 3

We begin by formally stating assumptions **B1-B3**, referenced in the statement of Theorem 3.

- B1** $\gamma_{\epsilon,n}(u), \eta_{\epsilon,n}(u)$, viewed as functions of $u \in [0, 1]$, are nondecreasing, differentiable, invertible with differentiable inverses, and are also respectively nondecreasing and nonincreasing functions of $\epsilon \in (0, 1)$ and $n \geq 1$.
- B2** There exists a constant $K_1 > 0$ such that for all $f, g \in \{\gamma_{\tau/N,n}, \eta_{\tau/N,n}, \iota : \tau \in \{\epsilon, \epsilon \wedge \alpha\}\}$, with ι the identity function on $[0, 1]$, we have $\delta/2 \leq f(\delta)$, $f(1 - \delta) \leq 1 - \delta/2$, and

$$\sup_{\delta/2 \leq u \leq 1 - \delta/2} \left| \frac{\partial g^{-1}(f(u))}{\partial u} - 1 \right| \leq K_1 \kappa_{\epsilon,n}.$$

- B3** There exists a constant $K_2 > 1$ such that for all $t \in [\delta/2, 1 - \delta/2]$ and all $\gamma_{\epsilon,n}(t) \leq u \leq \eta_{\epsilon,n}(t)$ we have

$$\frac{1}{K_2} \leq \frac{\gamma_{\epsilon,n}^{-1}(u) - \eta_{\epsilon,n}^{-1}(u)}{\eta_{\epsilon,n}(t) - \gamma_{\epsilon,n}(t)} \leq K_2.$$

Proof of Theorem 3 Throughout the proof, the symbol “ $\lesssim$ ” is used to hide constants possibly depending on K_1, K_2, δ_0, r . Furthermore, the symbol $\varkappa_N$ is used to denote a random variable depending only on $\theta_1, \dots, \theta_N$, whose definition may change from line to line, but which always satisfies $\mathbb{E}_{\mu^{\otimes N}}[\varkappa_N] \leq CN^{-1/2r}I(d \geq 2)$, where $C > 0$ denotes a constant possibly depending on K_1, K_2, δ, b, r . We prove the claim in five steps.

Step 0: Setup With probability at least $1 - \epsilon$, uniformly in $j = 1, \dots, N$ and $u \in (0, 1)$, we have both

$$F_{\theta_j,n}^{-1}(\gamma_{\epsilon/N,n}(u)) \leq F_{\theta_j}^{-1}(u) \leq F_{\theta_j,n}^{-1}(\eta_{\epsilon/N,n}(u)), \quad (57)$$

and,

$$G_{\theta_j,m}^{-1}(\gamma_{\epsilon/N,m}(u)) \leq G_{\theta_j}^{-1}(u) \leq G_{\theta_j,m}^{-1}(\eta_{\epsilon/N,m}(u)). \quad (58)$$

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 $a = \alpha/N$, $e = \epsilon/N$, and we recall that $\varepsilon = e \wedge a$.

Recall that for all $\theta \in \{\theta_1, \dots, \theta_N\}$, $F_{\theta,n}$, $G_{\theta,m}$ denote the empirical CDFs of P_θ and Q_θ respectively, and $F_{\theta,n}^{-1}$, $G_{\theta,m}^{-1}$ their corresponding quantile functions. We may write

$$C_{nm}^{(N)} = \left[\left(\frac{1}{1-2\delta} \int_{\mathbb{S}^{d-1}} \int_{\delta}^{1-\delta} A_{\theta,nm}(u) du d\mu_N(\theta) \right)^{\frac{1}{r}}, \right. \\ \left. \left(\frac{1}{1-2\delta} \int_{\mathbb{S}^{d-1}} \int_{\delta}^{1-\delta} B_{\theta,nm}^r(u) du d\mu_N(\theta) \right)^{\frac{1}{r}} \right],$$

where

$$A_{\theta,nm}(u) = [F_{\theta,n}^{-1}(\gamma_{a,n}(u)) - G_{\theta,m}^{-1}(\eta_{a,m}(u))] \\ \vee [G_{\theta,m}^{-1}(\gamma_{a,m}(u)) - F_{\theta,n}^{-1}(\eta_{a,n}(u))] \vee 0,$$

and

$$B_{\theta,nm}(u) = [F_{\theta,n}^{-1}(\eta_{a,n}(u)) - G_{\theta,m}^{-1}(\gamma_{a,m}(u))] \vee [G_{\theta,m}^{-1}(\eta_{a,m}(u)) - F_{\theta,n}^{-1}(\gamma_{a,n}(u))].$$

We will first show that

$$\left| \frac{1}{1-2\delta} \int_{\mathbb{S}^{d-1}} \int_{\delta}^{1-\delta} A_{\theta,nm}^r(u) du d\mu_N(\theta) - [\text{SW}_{r,\delta}^{(N)}(P, Q)]^r \right| \lesssim \psi_{\varepsilon,nm} + \varphi_{\varepsilon,nm} + \varkappa_N.$$

A similar argument can be used to bound this expression with $A_{\theta,nm}$ replaced by $B_{\theta,nm}$, 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 $\theta \in \{\theta_1, \dots, \theta_N\}$. By Taylor expansion of the map $(x, y) \in \mathbb{R}^2 \mapsto (x-y)^r$, there exists $\tilde{F}_{\theta,n}^{-1}(u)$ (resp. $\tilde{G}_{\theta,m}^{-1}(u)$) on the segment joining $F_{\theta,n}^{-1}(\gamma_{a,n}(u))$ and $F_{\theta}^{-1}(u)$ (resp. $G_{\theta}^{-1}(u)$ and $G_{\theta,m}^{-1}(\eta_{a,m}(u))$) such that

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

where

$$\xi_{\theta,nm}(u) = r \left(\tilde{F}_{\theta,n}^{-1}(u) - \tilde{G}_{\theta,m}^{-1}(u) \right)^{r-1} \times \\ \left\{ \left(F_{\theta,n}^{-1}(\gamma_{a,n}(u)) - F_{\theta}^{-1}(u) \right) - \left(G_{\theta,m}^{-1}(\eta_{a,m}(u)) - G_{\theta}^{-1}(u) \right) \right\}.$$

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

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

where

$$\zeta_{\theta,nm}(u) = r \left(\bar{G}_{\theta,m}^{-1}(u) - \bar{F}_{\theta,n}^{-1}(u) \right)^{r-1} \times \left\{ \left(F_{\theta,n}^{-1}(\eta_{a,n}(u)) - F_{\theta}^{-1}(u) \right) - \left(G_{\theta,m}^{-1}(\gamma_{a,m}(u)) - G_{\theta}^{-1}(u) \right) \right\}.$$

Now, consider the numerical inequality $|(a^r + b) \vee ((-a)^r + d) \vee 0 - |a|^r| \leq 3(|b| + |d|)$, 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

$$\begin{aligned} & \left| \frac{1}{1-2\delta} \int_{\mathbb{S}^{d-1}} \int_{\delta}^{1-\delta} A_{\theta,nm}^r(u) du d\mu_N(\theta) - \left[\text{SW}_{r,\delta}^{(N)}(P, Q) \right]^r \right| \\ & \leq \frac{1}{1-2\delta} \int_{\mathbb{S}^{d-1}} \int_{\delta}^{1-\delta} \left| \left[(F_{\theta}^{-1}(u) - G_{\theta}^{-1}(u))^r + \xi_{\theta,nm}(u) \right] \right. \\ & \quad \left. \vee \left[(G_{\theta}^{-1}(u) - F_{\theta}^{-1}(u))^r + \zeta_{\theta,nm}(u) \right] \vee 0 - |F_{\theta}^{-1}(u) - G_{\theta}^{-1}(u)|^r \right| du d\mu_N(\theta) \\ & \lesssim \int_{\mathbb{S}^{d-1}} \int_{\delta}^{1-\delta} |\zeta_{\theta,nm}(u)| du d\mu_N(\theta) + \int_{\mathbb{S}^{d-1}} \int_{\delta}^{1-\delta} |\xi_{\theta,nm}(u)| du d\mu_N(\theta). \quad (59) \end{aligned}$$

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

$$\int_{\mathbb{S}^{d-1}} \int_{\delta}^{1-\delta} |\xi_{\theta,nm}(u)| du d\mu_N(\theta) \leq r(\mathcal{I} + \mathcal{J}),$$

where,

$$\mathcal{I} = \int_{\mathbb{S}^{d-1}} \int_{\delta}^{1-\delta} \left| \tilde{F}_{\theta,n}^{-1}(u) - \tilde{G}_{\theta,m}^{-1}(u) \right|^{r-1} \left| F_{\theta,n}^{-1}(\gamma_{a,n}(u)) - F_{\theta}^{-1}(u) \right| du d\mu_N(\theta) \quad (60)$$

$$\mathcal{J} = \int_{\mathbb{S}^{d-1}} \int_{\delta}^{1-\delta} \left| \tilde{F}_{\theta,n}^{-1}(u) - \tilde{G}_{\theta,m}^{-1}(u) \right|^{r-1} \left| G_{\theta,m}^{-1}(\eta_{a,m}(u)) - G_{\theta}^{-1}(u) \right| du d\mu_N(\theta). \quad (61)$$

It will suffice to prove that $\mathcal{I} \lesssim \psi_{\varepsilon,nm}$ and $\mathcal{J} \lesssim \varphi_{\varepsilon,nm}$, up to terms depending only on N . We consider the cases $\text{SJ}_{r,\delta/2}(P) \vee \text{SJ}_{r,\delta/2}(Q) = \infty$ and $\text{SJ}_{r,\delta/2}(P) \vee \text{SJ}_{r,\delta/2}(Q) < \infty$ separately.

Step 2: Bounding $\mathcal{I}$ and $\mathcal{J}$ when $\text{SJ}_{r,\delta/2}(P) \vee \text{SJ}_{r,\delta/2}(Q) = \infty$ We have,

$$\begin{aligned} \mathcal{I} & \lesssim \int_{\mathbb{S}^{d-1}} \left(\sup_{\delta \leq u \leq 1-\delta} \left| \tilde{F}_{\theta,n}^{-1}(u) - \tilde{G}_{\theta,m}^{-1}(u) \right|^{r-1} \right) \times \\ & \quad \left(\int_{\delta}^{1-\delta} \left| F_{\theta,n}^{-1}(\gamma_{a,n}(u)) - F_{\theta}^{-1}(u) \right| du \right) d\mu_N(\theta). \end{aligned}$$

We will bound each factor in the above integral, beginning with the second. Using inequality (57), since $e \geq \varepsilon$, we have for all $u \in [\delta, 1 - \delta]$ and $\theta \in \{\theta_1, \dots, \theta_N\}$,

$$\begin{aligned} & |F_{\theta,n}^{-1}(\gamma_{a,n}(u)) - F_{\theta}^{-1}(u)| \\ & \leq [F_{\theta}^{-1}(\gamma_{\varepsilon,n}^{-1}(\gamma_{a,n}(u))) - F_{\theta}^{-1}(u)] + [F_{\theta}^{-1}(u) - F_{\theta}^{-1}(\eta_{\varepsilon,n}^{-1}(\gamma_{a,n}(u)))]. \end{aligned}$$

Now, write $x_n = \gamma_{\varepsilon,n}^{-1}(\gamma_{a,n}(\delta))$ and $y_n = \eta_{\varepsilon,n}^{-1}(\gamma_{a,n}(1 - \delta))$. By condition **B2** and the definition of $\kappa_{\varepsilon,n}$, we have

$$|x_n - \delta| \vee |y_n - 1 - \delta| \leq \kappa_{\varepsilon,n}, \quad (62)$$

which, combined with the assumption that $\kappa_{\varepsilon,n} \leq \frac{\delta}{2} \wedge (1 - 2\delta)$, also implies

$$\delta \leq x_n \leq 1 - \delta \leq y_n \leq 1 - \delta/2.$$

Thus, for all $\theta \in \{\theta_1, \dots, \theta_N\}$,

$$\begin{aligned} & \int_{\delta}^{1-\delta} [F_{\theta}^{-1}(\gamma_{\varepsilon,n}^{-1}(\gamma_{a,n}(u))) - F_{\theta}^{-1}(u)] du \\ & = \int_{x_n}^{y_n} F_{\theta}^{-1}(u) \left(\frac{\partial \gamma_{a,n}^{-1}(\gamma_{\varepsilon,n}(u))}{\partial u} \right) du - \int_{\delta}^{1-\delta} F_{\theta}^{-1}(u) du \\ & \leq \int_{x_n}^{y_n} F_{\theta}^{-1}(u) du + K_1 \kappa_{\varepsilon,n} \int_{x_n}^{y_n} |F_{\theta}^{-1}(u)| du - \int_{\delta}^{1-\delta} F_{\theta}^{-1}(u) du, \quad (\text{By } \mathbf{B2}) \\ & \leq \int_{1-\delta}^{y_n} F_{\theta}^{-1}(u) du - \int_{\delta}^{x_n} F_{\theta}^{-1}(u) du + K_1 \kappa_{\varepsilon,n} |F_{\theta}^{-1}(y_n)| \\ & \lesssim (y_n - 1 + \delta) [|F_{\theta}^{-1}(y_n)| \vee |F_{\theta}^{-1}(1 - \delta)|] \\ & \quad + (x_n - \delta) [|F_{\theta}^{-1}(\delta)| \vee |F_{\theta}^{-1}(x_n)|] + \kappa_{\varepsilon,n} |F_{\theta}^{-1}(y_n)| \\ & \lesssim (y_n - 1 + \delta + \kappa_{\varepsilon,n}) |F_{\theta}^{-1}(1 - \delta/2)| + (x_n - \delta) |F_{\theta}^{-1}(\delta)|. \end{aligned}$$

Since $P \in \bar{\mathcal{K}}_2(b)$, the quantiles in the above display are bounded above by a universal multiple of $(b/\delta)^{1/2}$, by Lemma 3. Thus, together with equation (62), we arrive at

$$\int_{\delta}^{1-\delta} [F_{\theta}^{-1}(\gamma_{\varepsilon,n}^{-1}(\gamma_{a,n}(u))) - F_{\theta}^{-1}(u)] du \lesssim \kappa_{\varepsilon,n} (b/\delta)^{1/2},$$

We similarly have,

$$\int_{\delta}^{1-\delta} [F_{\theta}^{-1}(u) - F_{\theta}^{-1}(\eta_{\varepsilon,n}^{-1}(\gamma_{a,n}(u)))] du \lesssim \kappa_{\varepsilon,n} (b/\delta)^{1/2}.$$

Combining these facts, we obtain

$$\mathcal{I} \lesssim \kappa_{\varepsilon,n} (b/\delta)^{1/2} \int_{\mathbb{S}^{d-1}} \sup_{\delta \leq u \leq 1-\delta} |\tilde{F}_{\theta,n}^{-1}(u) - \tilde{G}_{\theta,n}^{-1}(u)|^{r-1} d\mu_N(\theta). \quad (63)$$

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

$$\left| \tilde{F}_{\theta,n}^{-1}(u) - \tilde{G}_{\theta,m}^{-1}(u) \right| \quad (64)$$

$$\begin{aligned} &\leq \left| G_{\theta}^{-1}(u) - F_{\theta}^{-1}(u) \right| + \left| F_{\theta}^{-1}(u) - \tilde{F}_{\theta,n}^{-1}(u) \right| + \left| \tilde{G}_{\theta,m}^{-1}(u) - G_{\theta}^{-1}(u) \right| \\ &\leq \left| G_{\theta}^{-1}(u) - F_{\theta}^{-1}(u) \right| \\ &\quad + \left| F_{\theta}^{-1}(u) - F_{\theta}^{-1}(\eta_{\varepsilon,n}^{-1}(\gamma_{a,n}(u))) \right| + \left| G_{\theta}^{-1}(\gamma_{\varepsilon,m}^{-1}(\eta_{a,m}(u))) - G_{\theta}^{-1}(u) \right|. \end{aligned} \quad (65)$$

It follows that

$$\begin{aligned} &\int_{\mathbb{S}^{d-1}} \sup_{\delta \leq u \leq 1-\delta} \left| \tilde{F}_{\theta,n}^{-1}(u) - \tilde{G}_{\theta,m}^{-1}(u) \right|^{r-1} d\mu_N(\theta) \\ &\lesssim \left\{ \text{SW}_{\infty,\delta}^{(r-1)}(P, Q) + U_{\varepsilon,n}(P) + U_{\varepsilon,m}(Q) \right\} + \varkappa_N. \end{aligned}$$

We conclude this section of the proof by combining the above display with equation (63). We then have

$$\mathcal{I} \lesssim \kappa_{\varepsilon,n}(b/\delta)^{1/2} \left\{ \text{SW}_{\infty,\delta}^{(r-1)}(P, Q) + U_{\varepsilon,n}(P) + U_{\varepsilon,m}(Q) + \varkappa_N \right\} \lesssim \psi_{\varepsilon,nm} + \varkappa_N,$$

and by a symmetric argument,

$$\mathcal{J} \lesssim \varphi_{\varepsilon,nm} + \varkappa_N.$$

Step 3: Bounding $\mathcal{I}$ and $\mathcal{J}$ when $\text{SJ}_{r,\delta/2}(P) \vee \text{SJ}_{r,\delta/2}(Q) < \infty$ By means of Hölder's inequality, we have

$$\begin{aligned} \mathcal{I} &\leq \int_{\mathbb{S}^{d-1}} \int_{\delta}^{1-\delta} \left| \tilde{F}_{\theta,n}^{-1}(u) - \tilde{G}_{\theta,m}^{-1}(u) \right|^{r-1} \left| F_{\theta,n}^{-1}(\gamma_{a,n}(u)) - F_{\theta}^{-1}(u) \right| du d\mu_N(\theta) \\ &\leq \left(\int_{\mathbb{S}^{d-1}} \int_{\delta}^{1-\delta} \left| \tilde{F}_{\theta,n}^{-1}(u) - \tilde{G}_{\theta,m}^{-1}(u) \right|^r du d\mu_N(\theta) \right)^{\frac{r-1}{r}} \times \\ &\quad \left(\int_{\mathbb{S}^{d-1}} \int_{\delta}^{1-\delta} \left| F_{\theta,n}^{-1}(\gamma_{a,n}(u)) - F_{\theta}^{-1}(u) \right|^r du d\mu_N(\theta) \right)^{\frac{1}{r}} \\ &\lesssim \left(\int_{\mathbb{S}^{d-1}} \int_{\delta}^{1-\delta} \left| \tilde{F}_{\theta,n}^{-1}(u) - \tilde{G}_{\theta,m}^{-1}(u) \right|^r du d\mu_N(\theta) \right)^{\frac{r-1}{r}} \times \\ &\quad \left(\int_{\mathbb{S}^{d-1}} \int_{\delta}^{1-\delta} \left[F_{\theta}^{-1}(\gamma_{\varepsilon,n}^{-1}(\gamma_{a,n}(u))) - F_{\theta}^{-1}(\eta_{\varepsilon,n}^{-1}(\gamma_{a,n}(u))) \right]^r du d\mu_N(\theta) \right)^{\frac{1}{r}} \\ &= \left(\int_{\mathbb{S}^{d-1}} \int_{\delta}^{1-\delta} \left| \tilde{F}_{\theta,n}^{-1}(u) - \tilde{G}_{\theta,m}^{-1}(u) \right|^r du d\mu_N(\theta) \right)^{\frac{r-1}{r}} \times \end{aligned}$$

$$\begin{aligned}
& \left(\int_{\mathbb{S}^{d-1}} \int_{\gamma_{a,n}(\delta)}^{\gamma_{a,n}(1-\delta)} [F_{\theta}^{-1}(\gamma_{\varepsilon,n}^{-1}(u)) - F_{\theta}^{-1}(\eta_{\varepsilon,n}^{-1}(u))]^r \left(\frac{\partial \gamma_{a,n}^{-1}(u)}{\partial u} \right) du d\mu_N(\theta) \right)^{\frac{1}{r}} \\
& \leq \left(\int_{\mathbb{S}^{d-1}} \int_{\delta}^{1-\delta} |\tilde{F}_{\theta,n}^{-1}(u) - \tilde{G}_{\theta,m}^{-1}(u)|^r du d\mu_N(\theta) \right)^{\frac{r-1}{r}} \times \\
& \left(\int_{\mathbb{S}^{d-1}} \int_{\gamma_{a,n}(\delta)}^{\gamma_{a,n}(1-\delta)} [F_{\theta}^{-1}(\gamma_{\varepsilon,n}^{-1}(u)) - F_{\theta}^{-1}(\eta_{\varepsilon,n}^{-1}(u))]^r (1 + K_1 \kappa_{\varepsilon,n}) du d\mu_N(\theta) \right)^{\frac{1}{r}},
\end{aligned}$$

where we used condition **B2** on the final line. We now reason similarly as in Appendix B.2. Since $\text{SJ}_{r,\delta/2}(P) < \infty$, it follows from Lemma 2 that F_{θ}^{-1} is absolutely continuous over $[\delta/2, 1 - \delta/2]$ for μ -almost every $\theta \in \mathbb{S}^{d-1}$. We thus have for almost surely that for each $\theta \in \{\theta_1, \dots, \theta_N\}$,

$$\begin{aligned}
& \int_{\gamma_{a,n}(\delta)}^{\gamma_{a,n}(1-\delta)} [F_{\theta}^{-1}(\gamma_{\varepsilon,n}^{-1}(u)) - F_{\theta}^{-1}(\eta_{\varepsilon,n}^{-1}(u))]^r du \\
& = \int_{\gamma_{a,n}(\delta)}^{\gamma_{a,n}(1-\delta)} \left(\int_{\eta_{\varepsilon,n}^{-1}(u)}^{\gamma_{\varepsilon,n}^{-1}(u)} \frac{dt}{p_{\theta}(F_{\theta}^{-1}(t))} \right)^r du \\
& \leq \int_{\gamma_{a,n}(\delta)}^{\gamma_{a,n}(1-\delta)} (\gamma_{\varepsilon,n}^{-1}(u) - \eta_{\varepsilon,n}^{-1}(u))^{r-1} \int_{\eta_{\varepsilon,n}^{-1}(u)}^{\gamma_{\varepsilon,n}^{-1}(u)} \left(\frac{1}{p_{\theta}(F_{\theta}^{-1}(t))} \right)^r dt du \\
& \leq \int_{\delta/2}^{1-\delta/2} \left(\int_{\gamma_{\varepsilon,n}(t)}^{\eta_{\varepsilon,n}(t)} (\gamma_{\varepsilon,n}^{-1}(u) - \eta_{\varepsilon,n}^{-1}(u))^{r-1} du \right) \left(\frac{1}{p_{\theta}(F_{\theta}^{-1}(t))} \right)^r dt \\
& \leq \int_{\delta/2}^{1-\delta/2} \left(\sup_{\gamma_{\varepsilon,n}(t) \leq u \leq \eta_{\varepsilon,n}(t)} (\gamma_{\varepsilon,n}^{-1}(u) - \eta_{\varepsilon,n}^{-1}(u))^{r-1} \right) \frac{\eta_{\varepsilon,n}(t) - \gamma_{\varepsilon,n}(t)}{[p_{\theta}(F_{\theta}^{-1}(t))]^r} dt \\
& \lesssim \int_{\delta/2}^{1-\delta/2} \left(\frac{\eta_{\varepsilon,n}(t) - \gamma_{\varepsilon,n}(t)}{p_{\theta}(F_{\theta}^{-1}(t))} \right)^r dt \lesssim \int_{\delta/2}^{1-\delta/2} \left(\frac{\tilde{\kappa}_{\varepsilon,n}(t)}{p_{\theta}(F_{\theta}^{-1}(t))} \right)^r dt,
\end{aligned}$$

where we repeatedly used assumption **B3** on the final line. We thus arrive at,

$$\mathcal{I} \lesssim \left(\int_{\mathbb{S}^{d-1}} \int_{\delta}^{1-\delta} |\tilde{F}_{\theta,n}^{-1}(u) - \tilde{G}_{\theta,m}^{-1}(u)|^r du d\mu_N(\theta) \right)^{\frac{r-1}{r}} [V_{\varepsilon,n}(P)]^{\frac{1}{r}} + \varkappa_N. \quad (66)$$

By using similar calculations as in equations (64) and (66) to bound the first factor in the above display, we have

$$\begin{aligned}
& \int_{\mathbb{S}^{d-1}} \int_{\delta}^{1-\delta} \left(\tilde{F}_{\theta,n}^{-1}(u) - \tilde{G}_{\theta,m}^{-1}(u) \right)^r du d\mu_N(\theta) \\
& \lesssim \text{SW}_{r,\delta}^r(P, Q) + \int_{\mathbb{S}^{d-1}} \int_{\delta}^{1-\delta} |F_{\theta}^{-1}(u) - F_{\theta}^{-1}(\eta_{\varepsilon,n}^{-1}(\gamma_{a,n}(u)))|^r du d\mu_N(\theta) \\
& \quad + \int_{\mathbb{S}^{d-1}} \int_{\delta}^{1-\delta} |G_{\theta}^{-1}(\gamma_{\varepsilon,n}^{-1}(\gamma_{a,n}(u))) - G_{\theta}^{-1}(u)|^r du d\mu_N(\theta)
\end{aligned}$$

$$\lesssim \text{SW}_{r,\delta}^r(P, Q) + V_{\varepsilon,n}(P) + V_{\varepsilon,m}(Q) + \varkappa_N,$$

Putting these facts together with equation (66), we arrive at

$$\mathcal{I} \lesssim (\text{SW}_{r,\delta}^r(P, Q) + V_{\varepsilon,n}(P) + V_{\varepsilon,m}(Q))^{\frac{r-1}{r}} [V_{\varepsilon,n}(P)]^{\frac{1}{r}} + \varkappa_N = \psi_{\varepsilon,nm} + \varkappa_N.$$

Finally, by a symmetric argument, we also have $\mathcal{J} \lesssim \varphi_{\varepsilon,nm} + \varkappa_N$.

Step 4: Conclusion Returning to equation (59), we have shown

$$\int_{\mathbb{S}^{d-1}} \int_{\delta}^{1-\delta} |\xi_{nm}(u)| du d\mu_N(\theta) \lesssim \psi_{\varepsilon,nm} + \varphi_{\varepsilon,nm} + \varkappa_N.$$

By the same arguments, we may obtain the same upper bound, up to universal constant factors, on the integral $\int_{\mathbb{S}^{d-1}} \int_{\delta}^{1-\delta} |\zeta_{nm}(u)| du d\mu_N(\theta)$ in equation (59). We deduce that, for some $c_1 > 0$ (possibly depending on r), we have

$$\begin{aligned} & \left(\frac{1}{1-2\delta} \int_{\mathbb{S}^{d-1}} \int_{\delta}^{1-\delta} A_{nm}^r(u) du d\mu_N(\theta) \right)^{\frac{1}{r}} \\ & \geq \left\{ [\text{SW}_{r,\delta}^{(N)}(P, Q)]^r - c_1(\psi_{\varepsilon,nm} + \varphi_{\varepsilon,nm} + \varkappa_N) \right\}^{\frac{1}{r}} \\ & \geq \left\{ \text{SW}_{r,\delta}^r(P, Q) - c_1(\psi_{\varepsilon,nm} + \varphi_{\varepsilon,nm} + \varkappa_N) \right\}^{\frac{1}{r}}. \end{aligned}$$

By the same arguments, there exists a constant $c_2 > 0$ such that

$$\begin{aligned} & \left(\frac{1}{1-2\delta} \int_{\mathbb{S}^{d-1}} \int_{\delta}^{1-\delta} B_{nm}^r(u) du d\mu_N(\theta) \right)^{\frac{1}{r}} \\ & \leq \left\{ \text{SW}_{r,\delta}^r(P, Q) + c_2(\psi_{\varepsilon,nm} + \varphi_{\varepsilon,nm} + \varkappa_N) \right\}^{\frac{1}{r}}. \end{aligned}$$

The claim follows by choosing $c = c_1 \vee c_2$. $\square$

Appendix F: Proof of Theorem 4

The proof of this result has two main components. In Lemma 13 we show that the Sliced Wasserstein distance is Hadamard differentiable under certain conditions. Theorem 4 then follows via an application of the functional delta method.

Hadamard differentiability of the Sliced Wasserstein distance Throughout this subsection, for a metric space (T, ρ) , $C[T]$ denotes the set of real-valued continuous functions defined on T , endowed with the supremum norm, and $\ell^\infty(T) = \{f : T \rightarrow \mathbb{R} : \sup_{t \in T} |f(t)| < \infty\}$. Let $D[I]$ denote the 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 and 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 $\mathbb{D}_\psi$ 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

$$\psi : \mathbb{D}_\psi \subseteq D[I] \rightarrow \mathbb{R}, \quad \psi : A \mapsto A^{-1}(u) = \inf\{x \in I : A(x) \geq u\}. \quad (67)$$

Lemma 12 (van der Vaart and 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 ξ_u .*

Then, ψ is Hadamard-differentiable at A tangentially to the set of functions $H \in D[I]$ which are continuous at ξ_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 $\mathbb{D} = \ell^\infty(\mathcal{H})$, and let $\mathbb{D}_0$ denote the subspace of $\mathbb{D}$ consisting of maps $F : \mathcal{H} \rightarrow \mathbb{R}$ such that $F(\cdot, \theta) \in C[\mathbb{R}]$ for all $\theta \in \mathbb{S}^{d-1}$. Furthermore, define $\mathbb{D}_\phi$ as the subset of maps $F : \mathcal{H} \rightarrow \mathbb{R}$ such that $F(\cdot, \theta) \in D[\mathbb{R}]$ is a CDF for all $\theta \in \mathbb{S}^{d-1}$. Define the map

$$\phi : \mathbb{D}_\phi^2 \rightarrow \mathbb{R}_+, \quad \phi : (F, G) \mapsto \frac{1}{1-2\delta} \int_{\mathbb{S}^{d-1}} \int_\delta^{1-\delta} |F^{-1}(u, \theta) - G^{-1}(u, \theta)|^r d\mu(\theta),$$

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

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

Lemma 13. *Assume the same conditions as in Theorem 4. Then, the map ϕ is Hadamard differentiable at (F, G) , tangentially to $\mathbb{D}_0$, with Hadamard derivative given by*

$$\begin{aligned} \phi' : \mathbb{D}_0^2 &\rightarrow \mathbb{R}, \\ \phi'(H_1, H_2) &= \frac{r}{1-2\delta} \int_{\mathbb{S}^{d-1}} \int_\delta^{1-\delta} \operatorname{sgn}(F^{-1}(u, \theta) - G^{-1}(u, \theta)) \times \\ &\quad |F^{-1}(u, \theta) - G^{-1}(u, \theta)|^{r-1} \left(\frac{H_2(G^{-1}(u, \theta), \theta)}{q_\theta(G^{-1}(u, \theta))} - \frac{H_1(F^{-1}(u, \theta), \theta)}{p_\theta(F^{-1}(u, \theta))} \right) d\mu(\theta). \end{aligned}$$

Proof of Lemma 13 Let $(H_{1k})_{k=1}^\infty, (H_{2k})_{k=1}^\infty \subseteq \mathbb{D}$ be sequences satisfying $F + t_k H_{1k}, G + t_k H_{2k} \in \mathbb{D}_\phi$ for all $k \geq 1$, and such that H_{jk} converges uniformly to $H_j \in \mathbb{D}_0$, $j = 1, 2$. Let $t_k \downarrow 0$ as $k \rightarrow \infty$, and define for all $k \geq 1$,

$$\begin{aligned} \Delta_k &= \frac{1}{t_k(1-2\delta)} \int_{\mathbb{S}^{d-1}} \int_\delta^{1-\delta} \left\{ |(F + t_k H_{1k})^{-1}(u, \theta) - (G + t_k H_{2k})^{-1}(u, \theta)|^r \right. \\ &\quad \left. - |F^{-1}(u, \theta) - G^{-1}(u, \theta)|^r \right\} d\mu(\theta). \end{aligned}$$

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

$$\begin{aligned} \Delta_k &= \frac{r}{t_k(1-2\delta)} \int_{\mathbb{S}^{d-1}} \int_{\delta}^{1-\delta} \varphi(\tilde{F}_k^{-1}(u, \theta); \tilde{G}_k^{-1}(u, \theta)) \\ &\quad \times \left\{ [(F + t_k H_{1k})^{-1}(u, \theta) - F^{-1}(u, \theta)] \right. \\ &\quad \left. - [(G + t_k H_{2k})^{-1}(u, \theta) - G^{-1}(u, \theta)] \right\} d\mu(\theta), \end{aligned} \quad (68)$$

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

$$\begin{aligned} B_1(u, \theta) &= \lim_{k \rightarrow \infty} B_{1k}(u, \theta), \quad B_{1k}(u, \theta) = \varphi(\tilde{F}_k^{-1}(u, \theta); \tilde{G}_k^{-1}(u, \theta)), \\ B_2(u, \theta) &= \lim_{k \rightarrow \infty} B_{2k}(u, \theta), \quad B_{2k}(u, \theta) = \frac{(F + t_k H_{1k})^{-1}(u, \theta) - F^{-1}(u, \theta)}{t_k}, \\ B_3(u, \theta) &= \lim_{k \rightarrow \infty} B_{3k}(u, \theta), \quad B_{3k}(u, \theta) = \frac{(G + t_k H_{2k})^{-1}(u, \theta) - G^{-1}(u, \theta)}{t_k}, \end{aligned}$$

exist for $(\lambda \otimes \mu)$ -almost every $(u, \theta) \in [\delta, 1 - \delta] \times \mathbb{S}^{d-1}$. Throughout the sequel, we write $I_\theta = [F^{-1}(\delta/2, \theta), F^{-1}(1 - \delta/2, \theta)]$ for all $\theta \in \mathbb{S}^{d-1}$. By Lemma 3, there exist finite constants $a_1 < a_2$, depending on δ and the finite second moments of P and Q , such that $\bigcup_{\theta \in \mathbb{S}^{d-1}} I_\theta \subseteq [a_1, a_2]$. We fix $I = [a_1, a_2]$ in what follows.

Regarding the limit B_1 , we make use of the following observation, which we prove below in Section F.1. The conclusion of this assertion is stronger than necessary, but will be needed again in the sequel.

Lemma 14. *Assume the same conditions as Theorem 4. Then, for all $u \in [\delta, 1 - \delta]$, we have as $k \rightarrow \infty$,*

$$\begin{aligned} \sup_{\theta \in \mathbb{S}^{d-1}} |(F + t_k H_{1k})^{-1}(u, \theta) - F^{-1}(u, \theta)| &\rightarrow 0, \\ \sup_{\theta \in \mathbb{S}^{d-1}} |(G + t_k H_{2k})^{-1}(u, \theta) - G^{-1}(u, \theta)| &\rightarrow 0. \end{aligned}$$

Notice further that the map φ is continuous in both of its arguments, therefore we obtain from Lemma 14 that

$$B_1(u, \theta) = \varphi(F^{-1}(u, \theta); G^{-1}(u, \theta)),$$

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

We now turn to the limit B_2 . Recall that $A := F(\cdot, \theta)$ is absolutely continuous for any given $\theta \in \mathbb{S}^{d-1}$. 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 ψ in equation (67) at A , tangentially to $\mathbb{D}_0$, sufficient conditions for which are given by conditions

(i) and (ii) of Lemma 12. Condition (i) is immediately satisfied for almost all $u \in [\delta, 1 - \delta]$ due to the absolute continuity of A . Furthermore, the assumption $J_{\infty, \delta/2}(P) < \infty$ implies that p_θ is nonzero at $A^{-1}(u)$ for almost every $u \in [\delta, 1 - \delta]$, implying that condition (ii) of Lemma 12 is satisfied for all such u . We deduce from Lemma 12 the limit

$$B_2(u, \theta) = -\frac{H_1(A^{-1}(u), \theta)}{p_\theta(A^{-1}(u))} = -\frac{H_1(F^{-1}(u, \theta), \theta)}{p_\theta(F^{-1}(u, \theta))},$$

for $(\lambda \otimes \mu)$ -almost every (u, θ) . We similarly obtain, almost everywhere,

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

With these facts in place, the claim will follow from the Dominated Convergence Theorem if we are able to interchange the limit as $k \rightarrow \infty$ with the integrations in equation (68). To this end, it will suffice to show that there exists $K \geq 1$ such that

$$\operatorname{esssup}_{\theta \in \mathbb{S}^{d-1}} \operatorname{esssup}_{\delta \leq u \leq 1-\delta} \sup_{k \geq K} |B_{2k}(u, \theta)| < \infty. \quad (69)$$

A similar argument may then be used to obtain the same conclusion with B_{2k} replaced by B_{3k} . These facts will imply, in particular, that the maps $(F + t_k H_{1k})^{-1}$ and $(G + t_k H_{2k})^{-1}$ are uniformly bounded over $[\delta, 1 - \delta] \times \mathbb{S}^{d-1}$, which, together with the continuity of φ , then also implies that the above display holds with B_{2k} replaced by B_{1k} .

It thus remains to prove equation (69). We shall make use of the following properties.

Lemma 15. *Under the assumptions of Theorem 4, the following assertions hold.*

(i) *The family $\{F(\cdot, \theta)\}_{\theta \in \mathbb{S}^{d-1}}$ is uniformly absolutely continuous over I , in the sense that for all $t > 0$, there exists $\epsilon(t) > 0$ such that for any $[\alpha, \beta] \subseteq I$, we have*

$$(\beta - \alpha) \leq \epsilon(t) \implies \sup_{\theta \in \mathbb{S}^{d-1}} |F(\alpha, \theta) - F(\beta, \theta)| \leq t.$$

(ii) *We have,*

$$\gamma := \inf_{\theta \in \mathbb{S}^{d-1}} \inf_{\delta/2 \leq u \leq 1-3\delta/4} [F^{-1}(u + \delta/4, \theta) - F^{-1}(u, \theta)] > 0.$$

(iii) *There exists a constant $K \geq 1$ such that for all $k \geq K$, $u \in [\delta, 1 - \delta]$, and $\theta \in \mathbb{S}^{d-1}$,*

$$F^{-1}(3\delta/4, \theta) \leq (F + t_k H_{1k})^{-1}(u, \theta) \leq F^{-1}(1 - \delta/4, \theta).$$

(iv) *In particular, $(F + t_k H_{1k})^{-1}(u, \theta) - \gamma \in I_\theta$ for all $\theta \in \mathbb{S}^{d-1}$.*

Now, for all $u \in [\delta, 1 - \delta]$ and $\theta \in \mathbb{S}^{d-1}$, write

$$\xi_{\theta,u} = F^{-1}(u, \theta), \quad \xi_{\theta,uk} = (F + t_k H_{1k})^{-1}(u, \theta). \quad (70)$$

For all $k \geq 1$, let $\epsilon_k := \epsilon(t_k)$ be the constant corresponding to the choice $t = t_k$ in the statement of Lemma 15(i). This defines a sequence $(\epsilon_k)_{k \geq 1}$, which we may assume is nonincreasing, and satisfies $\epsilon_k < \gamma$ for all $k \geq K$, without loss of generality. By definition of the quantile function, we have

$$(F + t_k H_{1k})(\xi_{\theta,uk} - \epsilon_k, \theta) \leq u \leq (F + t_k H_{1k})(\xi_{\theta,uk}, \theta).$$

We use these inequalities to bound $\xi_{\theta,uk} - \xi_{\theta,u}$. Assume first that $\xi_{\theta,uk} < \xi_{\theta,u}$. Then, by absolute continuity of $F(\cdot, \theta)$ for all $\theta \in \mathbb{S}^{d-1}$, we have

$$0 \leq (F + t_k H_{1k})(\xi_{\theta,uk}, \theta) - F(\xi_{\theta,u}, \theta) = t_k H_{1k}(\xi_{\theta,uk}, \theta) - \int_{\xi_{\theta,uk}}^{\xi_{\theta,u}} p_\theta(x) dx.$$

By Lemma 15(iii), we have $[\xi_{\theta,uk}, \xi_{\theta,u}] \subseteq I_\theta$ for all $k \geq K$, $u \in [\delta, 1 - \delta]$, and $\theta \in \mathbb{S}^{d-1}$. Therefore, we obtain

$$0 \leq t_k H_{1k}(\xi_{\theta,uk}, \theta) - (\xi_{\theta,u} - \xi_{\theta,uk}) J_{\infty, \delta/2}^{-1}(P_\theta).$$

On the other hand, if $\xi_{\theta,uk} \geq \xi_{\theta,u}$, we have

$$\begin{aligned} 0 &\geq (F + t_k H_{1k})(\xi_{\theta,uk} - \epsilon_k) - F(\xi_{\theta,u}) \\ &= \left[F(\xi_{\theta,uk} - \epsilon_k) - F(\xi_{\theta,uk}) \right] + \left[F(\xi_{\theta,uk}) - F(\xi_{\theta,u}) \right] + t_k H_{1k}(\xi_{\theta,uk} - \epsilon_k) \\ &= - \int_{\xi_{\theta,uk} - \epsilon_k}^{\xi_{\theta,uk}} p_\theta(x) dx + \int_{\xi_{\theta,u}}^{\xi_{\theta,uk}} p_\theta(x) dx + t_k H_{1k}(\xi_{\theta,uk} - \epsilon_k) \\ &\geq -t_k + (\xi_{\theta,uk} - \xi_{\theta,u}) J_{\infty, \delta/2}^{-1}(P_\theta) + t_k H_{1k}(\xi_{\theta,uk} - \epsilon_k), \end{aligned}$$

where on the final line, we lower bounded the first term as follows. Since $\epsilon_k < \gamma$, we have $\xi_{\theta,uk} - \epsilon_k \in I_\theta \subseteq I$ by Lemma 15(iv). Again, we also have $\xi_{\theta,uk} \in I$ by Lemma 15(iii). Therefore, by the definition of $\epsilon_k = \epsilon(t_k)$, Lemma 15(i) can be applied to obtain the stated lower bound. Combine the preceding two displays to deduce,

$$\begin{aligned} |B_{2k}(u, \theta)| &= \left| \frac{\xi_{\theta,uk} - \xi_{\theta,u}}{t_k} \right| \leq (1 + |H_{1k}(\xi_{\theta,uk})| + |H_{1k}(\xi_{\theta,uk} - \epsilon_k)|) J_{\infty, \delta/2}(P_\theta) \\ &\leq (1 + \|H_{1k}\|_\infty + \|H_{2k}\|_\infty) \sup_{\theta \in \mathbb{S}^{d-1}} J_{\infty, \delta/2}(P_\theta). \end{aligned}$$

Now, recall that for $j = 1, 2$, H_{jk} converges uniformly to $H_j \in \mathbb{D}_0 \subseteq \ell^\infty(\mathcal{H})$. It must also follow that, up to modifying the value of $K \geq 1$, the function H_{jk} is bounded, uniformly in $k \geq K$. This fact combined with our assumption on P implies that the right-hand side of the above display is bounded above by a finite constant not depending on k, u, θ . Equation (69) readily follows, leading to the claim. $\square$

Proof of Theorem 4 The claim consists of two statements, to be proven in parallel. Since the set of half-spaces $\mathcal{H}$ forms a separable Vapnik-Chervonenkis class, it is Donsker, implying that the empirical process $\mathbb{G}_{nm} = \sqrt{\frac{nm}{n+m}}(P_n - P, Q_m - Q)$ converges weakly in $\mathbb{D}^2 = \ell^\infty(\mathcal{H}) \times \ell^\infty(\mathcal{H})$,

$$\sup_{h \in \text{BL}_1(\mathbb{D}^2)} \left| \mathbb{E}[h(\mathbb{G}_{nm})] - \mathbb{E}[h(\mathbb{G}_{(P,Q)})] \right| \longrightarrow 0, \quad (71)$$

to a process $\mathbb{G}_{(P,Q)} := (\sqrt{a}\mathbb{G}_P, \sqrt{1-a}\mathbb{G}_Q)$, where $\mathbb{G}_P$ and $\mathbb{G}_Q$ denote independent 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 $\mathbb{G}_P$ takes the form $\mathbb{G}_P(x, \theta) = \mathbb{G} \circ F(x, \theta)$ for a standard Brownian Bridge $\mathbb{G}$, for all $(x, \theta) \in \mathcal{H}$. By assumption, for all $\theta \in \mathbb{S}^{d-1}$, $F(\cdot, \theta)$ is continuous, and since almost all sample paths of $\mathbb{G}$ are continuous, we deduce that almost every sample path of $\mathbb{G}_P(\cdot, \theta)$ is also continuous. We deduce that $\mathbb{G}_P$ takes values in $\mathbb{D}_0$ almost surely, and similarly for $\mathbb{G}_Q = \bar{\mathbb{G}} \circ G$, where $\bar{\mathbb{G}}$ is a standard Brownian Bridge independent of $\mathbb{G}$.

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

$$\begin{aligned} \sup_{h \in \text{BL}_1(\mathbb{D}^2)} \left| \mathbb{E}[h(\mathbb{G}_{nm}^*) | X_1, \dots, X_n, Y_1, \dots, Y_m] - \mathbb{E}[h(\mathbb{G}_{(P,Q)})] \right| &\rightarrow 0, \\ \mathbb{E}[h(\mathbb{G}_{nm}^*) | X_1, \dots, X_n, Y_1, \dots, Y_m]^* - \mathbb{E}[h(\mathbb{G}_{nm}^*) | X_1, \dots, X_n, Y_1, \dots, Y_m]_* &\rightarrow 0, \end{aligned} \quad (72)$$

in outer probability, where h ranges over $\text{BL}_1(\mathbb{D}^2)$. Now, the Hadamard differentiability of ϕ (Lemma 13) together with equation (71) and the functional delta method (see for instance Theorems 3.9.4 of van der Vaart and Wellner (1996)) implies

$$\sup_{h \in \text{BL}_1(\mathbb{R})} \left| \mathbb{E} \left[h \left(\sqrt{\frac{nm}{n+m}} (\phi(P_n, Q_m) - \phi(P, Q)) \right) \right] - \mathbb{E} \left[h \left(\phi'(\mathbb{G}_{(P,Q)}) \right) \right] \right| \rightarrow 0. \quad (73)$$

Likewise, the delta method for the bootstrap (Theorem 3.9.11 of van der Vaart and Wellner (1996)) and equations (71) and (72) imply

$$\begin{aligned} \sup_{h \in \text{BL}_1(\mathbb{R})} \left| \mathbb{E} \left[h \left(\sqrt{\frac{nm}{n+m}} (\phi(P_n^*, Q_m^*) - \phi(P_n, Q_m)) \right) \right] \middle| X_1, \dots, X_n, Y_1, \dots, Y_m \right] \\ - \mathbb{E} \left[h \left(\phi'(\mathbb{G}_{(P,Q)}) \right) \right] \right| \longrightarrow 0, \end{aligned} \quad (74)$$

in outer probability. A combination of equations (73) and (74) readily leads to part (ii) of the claim. In view of equation (73), part (i) of the claim will follow

upon showing that the random variable $\phi'(\mathbb{G}_{(P,Q)})$ is equal in distribution to $N(0, a\sigma_P^2 + (1-a)\sigma_Q^2)$. In the sequel, write for all $u \in [\delta, 1-\delta]$,

$$w_P(u) = \int_{\mathbb{S}^{d-1}} \frac{w(u, \theta)}{p_\theta(F_\theta^{-1}(u))} d\mu(\theta), \quad w_Q(u) = \int_{\mathbb{S}^{d-1}} \frac{w(u, \theta)}{q_\theta(G_\theta^{-1}(u))} d\mu(\theta).$$

Notice that

$$\begin{aligned} \phi'(\mathbb{G}_{(P,Q)}) &= \int_{\mathbb{S}^{d-1}} \int_{\delta}^{1-\delta} w(u, \theta) \left(\frac{\sqrt{1-a}\mathbb{G}_Q(G_\theta^{-1}(u), \theta)}{q_\theta(G_\theta^{-1}(u))} - \frac{\sqrt{a}\mathbb{G}_P(F_\theta^{-1}(u), \theta)}{p_\theta(F_\theta^{-1}(u))} \right) dud\mu(\theta) \\ &= \int_{\mathbb{S}^{d-1}} \int_{\delta}^{1-\delta} w(u, \theta) \left(\frac{\sqrt{1-a}\overline{\mathbb{G}}(u)}{q_\theta(G_\theta^{-1}(u))} - \frac{\sqrt{a}\mathbb{G}(u)}{p_\theta(F_\theta^{-1}(u))} \right) dud\mu(\theta) \\ &= \sqrt{1-a} \int_{\delta}^{1-\delta} w_Q(u) \overline{\mathbb{G}}(u) du - \sqrt{a} \int_{\delta}^{1-\delta} w_P(u) \mathbb{G}(u) du, \end{aligned}$$

where, on the final line, the interchange of order of integration is valid $\mathbb{P}$ -almost surely. Indeed, the sample paths of $\mathbb{G}$ and $\overline{\mathbb{G}}$ are almost surely continuous, whence bounded over $[\delta, 1-\delta]$, which in turn implies that the functions $w(u, \theta)\mathbb{G}(u)/p_\theta(F_\theta^{-1}(u))$ and $w(u, \theta)\overline{\mathbb{G}}(u)/q_\theta(G_\theta^{-1}(u))$ are almost surely bounded, due to the assumptions placed on P and Q . We now make use of the following fact, which is proven below for completeness.

Lemma 16. *Let $f : [\delta, 1-\delta] \rightarrow \mathbb{R}$ be a Lebesgue-measurable and bounded function. Then, the random variable $\int_{\delta}^{1-\delta} f(u)\mathbb{G}(u)du$ has Gaussian distribution with mean zero and variance*

$$\text{Var} \left[\int_{\delta}^{1-\delta} f(u)\mathbb{G}(u)du \right] = \int_0^{1-\delta} \left(\int_{\delta \vee t}^{1-\delta} f(u)du \right)^2 dt - \left(\int_0^{1-\delta} \int_{\delta \vee t}^{1-\delta} f(u)dudt \right)^2.$$

By Lemma 16 and the independence of $\mathbb{G}$ and $\overline{\mathbb{G}}$, we obtain that $\phi'(\mathbb{G}_{(P,Q)})$ has Gaussian distribution with mean zero and variance

$$\begin{aligned} \text{Var} [\phi'(\mathbb{G}_{(P,Q)})] &= a \left[\int_0^{1-\delta} \left(\int_{\delta \vee t}^{1-\delta} w_P(u)du \right)^2 dt - \left(\int_0^{1-\delta} \int_{\delta \vee t}^{1-\delta} w_P(u)dudt \right)^2 \right] \\ &\quad + (1-a) \left[\int_0^{1-\delta} \left(\int_{\delta \vee t}^{1-\delta} w_Q(u)du \right)^2 dt - \left(\int_0^{1-\delta} \int_{\delta \vee t}^{1-\delta} w_Q(u)dudt \right)^2 \right]. \end{aligned}$$

Finally, notice that

$$\int_{\delta \vee t}^{1-\delta} w_P(u)du = \int_{\delta \vee t}^{1-\delta} \int_{\mathbb{S}^{d-1}} \frac{w(u, \theta)}{p_\theta(F_\theta^{-1}(u))} d\mu(\theta) du$$

$$\begin{aligned}
&= \int_{\mathbb{S}^{d-1}} \int_{\delta \vee t}^{1-\delta} \frac{w(u, \theta)}{p_\theta(F_\theta^{-1}(u))} du d\mu(\theta) \\
&= \int_{\mathbb{S}^{d-1}} \int_{F_\theta^{-1}(\delta \vee t)}^{F_\theta^{-1}(1-\delta)} w(F_\theta(x), \theta) dx d\mu(\theta),
\end{aligned}$$

where, once again, the interchange of the order of integration is valid due to the uniform boundedness of the integrands almost everywhere. A similar computation holds with w_P replaced by w_Q , thus $\text{Var}[\phi'(\mathbb{G}_{(P,Q)})] = a\sigma_P^2 + (1-a)\sigma_Q^2$, and the claim follows. $\square$

It remains to prove Lemmas 14–16.

F.1. Proof of Lemma 14

Let $u \in [\delta, 1 - \delta]$. We prove the claim for F , and an identical argument may then be used for G . We use the abbreviations in equation (70). Let $\epsilon > 0$ be an arbitrary real number satisfying

$$\epsilon < \inf_{\theta \in \mathbb{S}^{d-1}} [F^{-1}(1 - \delta/2, \theta) - F^{-1}(1 - \delta, \theta)].$$

The infimum on the right-hand side is strictly positive by Lemma 15(ii), whose proof below is a consequence of the uniform integrability of $\{p_\theta\}_{\theta \in \mathbb{S}^{d-1}}$, and does not require the present result. By definition of ϵ , we have $\xi_{\theta,u}, \xi_{\theta,u} + \epsilon \in I_\theta$, thus, by absolute continuity of $F(\cdot, \theta)$,

$$\begin{aligned}
\inf_{\theta \in \mathbb{S}^{d-1}} [F(\xi_{\theta,u} + \epsilon, \theta) - u] &= \inf_{\theta \in \mathbb{S}^{d-1}} \int_{\xi_{\theta,u}}^{\xi_{\theta,u} + \epsilon} p_\theta(x) dx \\
&\geq \epsilon \inf_{\theta \in \mathbb{S}^{d-1}} \text{essinf}_{\xi_{\theta,u} \leq x \leq \xi_{\theta,u} + \epsilon} p_\theta(x) \\
&\geq \epsilon \inf_{\theta \in \mathbb{S}^{d-1}} \text{essinf}_{x \in I_\theta} p_\theta(x) > 0,
\end{aligned}$$

where the strict inequality follows from the fact that $\sup_\theta J_{\infty, \delta/2}(P_\theta) < \infty$. After repeating a symmetric argument, we deduce

$$\sup_{\theta \in \mathbb{S}^{d-1}} F(\xi_{\theta,u} - \epsilon, \theta) < u < \inf_{\theta \in \mathbb{S}^{d-1}} F(\xi_{\theta,u} + \epsilon, \theta). \quad (75)$$

On the other hand, by definition of quantile, we have for all $\epsilon > 0$ that,

$$\sup_{\theta \in \mathbb{S}^{d-1}} (F + t_k H_{1k})(\xi_{\theta,uk} - \epsilon, \theta) \leq u \leq \inf_{\theta \in \mathbb{S}^{d-1}} (F + t_k H_{1k})(\xi_{\theta,uk}, \theta).$$

Recall that H_{1k} converges in $\ell^\infty(\mathcal{H})$ to $H_1 \in \ell^\infty(\mathcal{H})$, thus there exists $C > 0$ such that

$$\sup_{\theta \in \mathbb{S}^{d-1}} F(\xi_{\theta,uk} - \epsilon, \theta) - Ct_k \leq u \leq \inf_{\theta \in \mathbb{S}^{d-1}} F(\xi_{\theta,uk}, \theta) + Ct_k.$$

Since $t_k \downarrow 0$, the above display contradicts equation (75) for all large enough k if $\xi_{\theta,uk} < \xi_{\theta,u} - \epsilon$, or if $\xi_{\theta,uk} > \xi_{\theta,u} + \epsilon$. We have therefore shown that for all $\epsilon > 0$ small enough, there exists a sufficiently large $K \geq 1$ such that for all $k \geq K$, $\sup_{\theta} |\xi_{\theta,uk} - \xi_{\theta,u}| \leq \epsilon$, thus leading to the claim. $\square$

F.2. Proof of Lemma 15

Consider the family $\mathcal{F} = \{p_{\theta}\}_{\theta \in \mathbb{S}^{d-1}}$. $\mathcal{F}$ is assumed to be uniformly integrable with respect to the Lebesgue measure over $\mathbb{R}$, and hence is also uniformly integrable with respect to the finite measure $\nu = \lambda|_I$ (that is, the restriction of the Lebesgue measure to the bounded interval I). Since ν does not possess any atoms, uniform integrability of $\mathcal{F}$ is equivalent to $\mathcal{F}$ having uniformly absolutely continuous integrals, by Proposition 4.5.3 of Bogachev (2007). Thus, for any $t > 0$, there exists $\epsilon(t) > 0$ such that for any interval $[\alpha, \beta] \subseteq I$ for which $\beta - \alpha \leq \epsilon(t)$, we have

$$\sup_{\theta \in \mathbb{S}^{d-1}} |F(\alpha, \theta) - F(\beta, \theta)| \leq t.$$

This proves property (i). For part (ii), choose $t < \delta/4$ and fix the corresponding value of $\epsilon(t)$. Let $u \in [\delta/2, 1 - \delta/4]$, and choose a sequence $\theta_j \in \mathbb{S}^{d-1}$ such that for all $j \geq 1$,

$$|F^{-1}(u + \delta/4, \theta_j) - F^{-1}(u, \theta_j)| \leq \frac{1}{j} + \inf_{\theta \in \mathbb{S}^{d-1}} |F^{-1}(u + \delta/4, \theta) - F^{-1}(u, \theta)|. \quad (76)$$

Let $\alpha_j = F^{-1}(u, \theta_j)$ and $\beta_j = F^{-1}(u + \delta/4, \theta_j)$. Clearly, $F(\beta_j, \theta_j) - F(\alpha_j, \theta_j) = \delta/4 > t$, thus from the uniform absolute continuity of $\{F_{\theta}\}$ in part (i), it must hold that

$$\beta_j - \alpha_j = F^{-1}(u + \delta/4, \theta_j) - F^{-1}(u, \theta_j) > \epsilon(t).$$

Since this property holds for all $j \geq 1$, we deduce from equation (76) that

$$\inf_{\theta \in \mathbb{S}^{d-1}} |F^{-1}(u + \delta/4, \theta) - F^{-1}(u, \theta)| \geq \epsilon(t)/2.$$

Since $\epsilon(t)$ did not depend on u , we finally arrive at

$$\gamma = \inf_{\delta/2 \leq u \leq 1 - \delta/4} \inf_{\theta \in \mathbb{S}^{d-1}} |F^{-1}(u + \delta/4, \theta) - F^{-1}(u, \theta)| \geq \epsilon(t)/2 > 0,$$

which proves the second claim. To prove the third and fourth, notice simply that

$$\begin{aligned} & (F + t_k H_{1k})^{-1}(\delta, \theta) \\ & \geq F^{-1}(\delta, \theta) - \sup_{\theta \in \mathbb{S}^{d-1}} |F^{-1}(\delta, \theta) - (F + t_k H_{1k})^{-1}(\delta, \theta)| \\ & \geq F^{-1}(3\delta/4, \theta) + \gamma - \sup_{\theta \in \mathbb{S}^{d-1}} |F^{-1}(\delta, \theta) - (F + t_k H_{1k})^{-1}(\delta, \theta)|. \end{aligned}$$

By Lemma 14, recall that there exists $K \geq 1$ such that $\sup_{\theta \in \mathbb{S}^{d-1}} |F^{-1}(\delta, \theta) - (F + t_k H_{1k})^{-1}(\delta, \theta)| \leq \gamma$, thus for all such k , we have

$$(F + t_k H_{1k})^{-1}(\delta, \theta) \geq F^{-1}(3\delta/4, \theta).$$

Similarly, up to modifying the value of K , we have for all $k \geq K$,

$$(F + t_k H_{1k})^{-1}(1 - \delta, \theta) \leq F^{-1}(1 - \delta/4, \theta). \quad (77)$$

The claim follows from here. $\square$

F.3. Proof of Lemma 16

We first prove the claim for step functions f . Let $M \geq 1$ be an integer and define a partition of $[\delta, 1 - \delta]$ via $\delta = s_0 < \dots < s_{M+1} = 1 - \delta$. Let $\alpha_0, \dots, \alpha_M \in \mathbb{R}$ and set $f = \sum_{i=0}^M \alpha_i I_{[s_i, s_{i+1})}$. Clearly, we may always rewrite f in terms of any refinement of the partition $s_0, \dots, s_{M+1}$. Indeed, for any $K \geq 1$ and any set of real numbers $0 = t_0 < \dots < t_{K+1} = 1 - \delta$ containing $\{s_0, \dots, s_{M+1}\}$, we may find real numbers $a_0, \dots, a_K$ contained in $\{\alpha_0, \dots, \alpha_M\}$ such that $f = \sum_{k=0}^K a_k I_{[t_k, t_{k+1})}$. We must have $a_0 = 0$ when $t_0 = 0 < \delta$. Since $\mathbb{G}$ is almost surely continuous over $[\delta, 1 - \delta]$, and the function f is piecewise continuous, $f\mathbb{G}$ is almost surely Riemann integrable. Therefore, for any choice of the partition $\{t_0, \dots, t_{K+1}\}$ with vanishing mesh as $K \rightarrow \infty$, we have

$$\int_{\delta}^{1-\delta} f(u) \mathbb{G}(u) du = \lim_{K \rightarrow \infty} \sum_{k=1}^{K+1} \mathbb{G}(t_{k-1})(t_k - t_{k-1}) a_{k-1}. \quad (78)$$

Notice that,

$$\begin{aligned} & \sum_{k=1}^{K+1} \mathbb{G}(t_{k-1})(t_k - t_{k-1}) a_{k-1} \\ &= \sum_{k=1}^{K+1} \mathbb{G}(t_{k-1}) \left[\sum_{j=k-1}^K a_j(t_{j+1} - t_j) - \sum_{j=k}^K a_j(t_{j+1} - t_j) \right] \\ &= \sum_{k=0}^K \mathbb{G}(t_k) \sum_{j=k}^K a_j(t_{j+1} - t_j) - \sum_{k=1}^{K+1} \mathbb{G}(t_{k-1}) \sum_{j=k}^K a_j(t_{j+1} - t_j) \\ &= \sum_{k=1}^K \mathbb{G}(t_k) \sum_{j=k}^K a_j(t_{j+1} - t_j) - \sum_{k=1}^K \mathbb{G}(t_{k-1}) \sum_{j=k}^K a_j(t_{j+1} - t_j) \\ &= \sum_{k=1}^K (\mathbb{G}(t_k) - \mathbb{G}(t_{k-1})) \sum_{j=k}^K a_j(t_{j+1} - t_j) \\ &= \sum_{k=1}^K (\mathbb{G}(t_k) - \mathbb{G}(t_{k-1})) \int_{t_k}^{1-\delta} f(x) dx. \end{aligned}$$

Now, since f is bounded, the function $t \in [0, 1 - \delta] \mapsto \int_{\delta \vee t}^{1-\delta} f(x) dx$ is continuous and bounded, thus for any partition as in equation (78), we obtain

$$\begin{aligned} \int_{\delta}^{1-\delta} f(u) \mathbb{G}(u) du \\ = \lim_{K \rightarrow \infty} \sum_{k=1}^{K+1} (\mathbb{G}(t_k) - \mathbb{G}(t_{k-1})) \int_{t_k}^{1-\delta} f(x) dx = \int_0^{1-\delta} \int_{\delta \vee t}^{1-\delta} f(x) dx d\mathbb{G}(t), \end{aligned}$$

where convergence in the final equality is understood to hold in probability (see, for instance, Proposition 2.13 of Revuz and Yor (2013)). It now follows from Proposition 2.2.1 of Denker (1985) that the random variable on the right-hand side of the above display has mean-zero Gaussian distribution, with variance

$$\int_0^{1-\delta} \left(\int_{\delta \vee t}^{1-\delta} f(u) du \right)^2 dt - \left(\int_0^{1-\delta} \int_{\delta \vee t}^{1-\delta} f(u) du dt \right)^2,$$

which leads to the claim when f is a step function.

Assume now that f is a Lebesgue measurable bounded function. By Theorem 4.3 of Stein and Shakarchi (2009), there exists a sequence of step functions f_n converging pointwise to f , Lebesgue-almost everywhere on $[\delta, 1 - \delta]$. In view of the preceding result, we have in probability,

$$\begin{aligned} \left| \int_{\delta}^{1-\delta} f(u) \mathbb{G}(u) du - \int_0^{1-\delta} \int_{t \vee \delta}^{1-\delta} f(u) du d\mathbb{G}(t) \right| \\ \leq \int_{\delta}^{1-\delta} |f_n - f| |\mathbb{G}| + \int_0^{1-\delta} \left(\int_{\delta \vee t}^{1-\delta} |f_n - f| \right) d\mathbb{G}(t). \end{aligned}$$

Since f is bounded, we may clearly take the functions f_n to be uniformly bounded. Furthermore, since $\mathbb{G}$ is $\mathbb{P}$ -almost surely bounded on the compact set $[\delta, 1 - \delta]$, the first term on the right-hand side of the above display vanishes by the Dominated Convergence Theorem, while the second vanishes in probability by Theorem 2.12 of Revuz and Yor (2013). Deduce that the identity $\int f(u) \mathbb{G}(u) du = \int_0^{1-\delta} \int_{t \vee \delta}^{1-\delta} f(u) du d\mathbb{G}(t)$ holds, with convergence holding in probability. The claim then follows as before, by Denker (1985). $\square$

Appendix G: Proofs of additional results

G.1. Proof of Proposition 6

Given $\theta \sim \mu$, we have

$$\begin{aligned} W_r(P_\theta, Q_\theta) &= \frac{1}{1 - 2\delta} \int_{\delta}^{1-\delta} |F_\theta^{-1}(u) - G_\theta^{-1}(u)|^r du \\ &\lesssim \max_{a \in \{\delta, 1-\delta\}} |F_\theta^{-1}(a)|^r + \max_{a \in \{\delta, 1-\delta\}} |G_\theta^{-1}(a)|^r. \end{aligned}$$

Therefore, it follows from Lemma 3 that

$$\sup_{P, Q \in \mathcal{K}_{2r}(b)} \text{Var}_\mu [W_{r,\delta}^r(P_\theta, Q_\theta)] \leq \sup_{P, Q \in \mathcal{K}_{2r}(b)} \int_{\mathbb{S}^{d-1}} W_r^{2r}(P_\theta, Q_\theta) d\mu(\theta) \lesssim b/\delta^r.$$

Thus, denoting $S_N = \text{SW}_{r,\delta}^{(N)}(P, Q)$, $S = \text{SW}_{r,\delta}(P, Q)$ and $\Delta_N = M_N/\sqrt{N}$, we obtain

$$\begin{aligned} & \mathbb{P}(S \notin \bar{C}_{nm}^{(N)}) \\ &= \mathbb{P}(S \notin \bar{C}_{nm}^{(N)}, |S^r - S_N^r| > \Delta_N) + \mathbb{P}(S \notin \bar{C}_{nm}^{(N)}, |S^r - S_N^r| \leq \Delta_N) \\ &\leq \mathbb{P}(|S^r - S_N^r| > \Delta_N) \\ &+ \mathbb{P}((\{S^r \leq L_{N,nm} - \Delta_N\} \cup \{U_{N,nm} + \Delta_N \leq S^r\}) \cap \{|S^r - S_N^r| \leq \Delta_N\}) \\ &\leq \mathbb{P}(|S^r - S_N^r| > \Delta_N) + \mathbb{P}(S_N^r \notin C_{nm}^{(N)}) \\ &= \mathbb{P}(|S^r - S_N^r| > \Delta_N) + \mathbb{E} \left[\mathbb{P}(S_N \notin C_{nm}^{(N)} \mid \theta_1, \dots, \theta_N) \right] \\ &\leq \frac{\text{Var}_{\theta \sim \mu} [W_{r,\delta}^r(P_\theta, Q_\theta)]}{N\Delta_N^2} + \alpha \leq \frac{bc/\delta^r}{M_N^2} + \alpha, \end{aligned}$$

for a constant $c > 0$ depending only on r , as claimed. $\square$

G.2. Proof of Corollary 1

Under the stated conditions on ϵ, δ , it can be seen by direct verification that conditions **B1**–**B3** and the remaining conditions of Theorem 3 hold for the confidence bands of Examples 1–2, for constants K_1, K_2 possibly depending on α . In what follows, the symbol “ $\lesssim$ ” is used to hide constants possibly depending on α, b, δ_0 and r .

Suppose first that $\text{SJ}_{r,\frac{\delta}{2}}(P) \vee \text{SJ}_{r,\frac{\delta}{2}}(Q) = \infty$. Then, we use the bound

$$\begin{aligned} \psi_{\varepsilon,nm} &\lesssim (\text{SW}_{\infty,\delta}^{(r-1)}(P, Q) + U_{\varepsilon,n}(P) + U_{\varepsilon,m}(Q)) \frac{\kappa_{\varepsilon,n}}{\sqrt{\delta}} \\ &\lesssim \left(\max_{a \in \{\delta/2, 1-\delta/2\}} \sup_{\mathbb{S}^{d-1}} |F_\theta^{-1}(a)|^{r-1} \right) \frac{\kappa_{\varepsilon,n}}{\sqrt{\delta}} \lesssim \frac{\kappa_{\varepsilon,n}}{\delta^{r/2}}, \end{aligned}$$

by Lemma 3, under the assumption $P, Q \in \bar{\mathcal{K}}_2(b)$. A similar argument can be used to bound $\varphi_{\varepsilon,nm}$, leading to

$$\begin{aligned} \lambda(C_{nm}^{(N)}) &\leq \left\{ \text{SW}_{r,\delta}^r(P, Q) + c(\psi_{\varepsilon,nm} + \varphi_{\varepsilon,nm} + \varkappa_N) \right\}^{1/r} - \text{SW}_{r,\delta}(P, Q) \\ &\leq c^{1/r} (\psi_{\varepsilon,nm} + \varphi_{\varepsilon,nm} + \varkappa_N)^{1/r} \\ &\lesssim \varkappa_N^{1/r} + \frac{1}{\sqrt{\delta}} (\kappa_{\varepsilon,n}^{1/r} + \kappa_{\varepsilon,m}^{1/r}), \end{aligned}$$

with probability at least $1 - \epsilon$. Parts (i) and (ii) now follow from Lemma 1 in the case $\text{SJ}_{r, \frac{\delta}{2}}(P) \vee \text{SJ}_{r, \frac{\delta}{2}}(Q) = \infty$.

Suppose now that $\text{SJ}_{r, \frac{\delta}{2}}(P) \vee \text{SJ}_{r, \frac{\delta}{2}}(Q) < \infty$. Using the shorthand notations

$$\Delta = V_{\varepsilon, n}(P) + V_{\varepsilon, m}(Q), \quad S = \text{SW}_{r, \delta}(P, Q),$$

we have the bounds $\psi_{\varepsilon, nm}, \varphi_{\varepsilon, nm} \lesssim (S^r + \Delta)^{\frac{r-1}{r}} \Delta^{\frac{1}{r}}$, whence

$$\lambda(C_{nm}^{(N)}) \lesssim \varkappa_N^{\frac{1}{r}} + \left\{ S^r + (S^r + \Delta)^{\frac{r-1}{r}} \Delta^{\frac{1}{r}} \right\}^{\frac{1}{r}} - S,$$

with probability at least $1 - \epsilon$. If $S^r \lesssim \Delta$, the right-hand side of the above display is clearly of order $\varkappa_N^{1/r} + \Delta^{1/r}$. Likewise, if $S^r \gtrsim \Delta$, we obtain similarly as in equation (43),

$$\begin{aligned} \lambda(C_{nm}^{(N)}) &\lesssim \varkappa_N^{\frac{1}{r}} + S \left[\left\{ 1 + (1 + (\Delta/S^r))^{\frac{r-1}{r}} (\Delta/S^r)^{\frac{1}{r}} \right\}^{\frac{1}{r}} - 1 \right] \\ &\leq \varkappa_N^{\frac{1}{r}} + S(1 + (\Delta/S^r))^{\frac{r-1}{r}} (\Delta/S^r)^{\frac{1}{r}} \\ &\lesssim \varkappa_N^{\frac{1}{r}} + \Delta^{\frac{1}{r}}, \end{aligned}$$

with probability at least $1 - \epsilon$. The conclusion of the above display thus holds irrespective of S and Δ . The claim now follows by substituting the expressions for $V_{\varepsilon, n}(P)$ stated in Lemma 1 for each of parts (i) and (ii). $\square$

G.3. Proof of Corollary 2

We reason similarly as in the proof of Corollary 1. By Theorem 3, we have with probability at least $1 - \epsilon$,

$$\begin{aligned} \lambda(C_{nm}) &\leq \left\{ \text{SW}_{r, \delta}^r(P, Q) + c(\psi_{\varepsilon, nm} + \varphi_{\varepsilon, nm} + \varkappa_N) \right\}^{1/r} - \text{SW}_{r, \delta}(P, Q) \\ &\lesssim \text{SW}_{r, \delta}^{1-r}(P, Q)(\psi_{\varepsilon, nm} + \varphi_{\varepsilon, nm} + \varkappa_N) \\ &\leq \Gamma^{1-r}(\psi_{\varepsilon, nm} + \varphi_{\varepsilon, nm} + \varkappa_N), \end{aligned}$$

since $\text{SW}_{r, \delta}(P, Q) \geq \Gamma$. The claim now follows by invoking similar bounds on $\psi_{\varepsilon, nm}$ and $\varphi_{\varepsilon, nm}$ as in the proof of Corollary 1. $\square$

G.4. Proof of Proposition 7

Notice that $\sigma_P, \sigma_Q > 0$ whenever $\text{SW}_{r, \delta}(P, Q) > 0$. In view of Theorem 4, it is then a standard result that the percentile bootstrap interval C_{nm}^* satisfies

$$\liminf_{n, m \rightarrow \infty} \mathbb{P}(\text{SW}_{r, \delta}(P, Q) \in C_{nm}^*) \geq 1 - \alpha/2, \quad (79)$$

under the assumptions of Theorem 4 and the assumption $\text{SW}_{r,\delta}(P, Q) > 0$ (see, for instance, Lemma 23.3 of van der Vaart (1998)). Therefore, when these assumptions hold, we have

$$\begin{aligned} & \mathbb{P}(\text{SW}_{r,\delta}(P, Q) \notin C_{nm}) \\ &= \mathbb{P}(\text{SW}_{r,\delta}(P, Q) \notin C_{nm}^*, 0 \notin C_{nm}^\dagger) + \mathbb{P}(\text{SW}_{r,\delta}(P, Q) \notin C_{nm}^\dagger, 0 \in C_{nm}^\dagger) \\ &\leq \mathbb{P}(\text{SW}_{r,\delta}(P, Q) \notin C_{nm}^*) + \mathbb{P}(\text{SW}_{r,\delta}(P, Q) \notin C_{nm}^\dagger) \leq \alpha + o(1), \end{aligned}$$

where on the final line, we use equation (79) and Proposition 6. Notice that the assumptions of Proposition 6 are satisfied, since $\bar{\mathcal{K}}_2 \subseteq \mathcal{K}_{2r}$. On the other hand, when $\text{SW}_{r,\delta}(P, Q) = 0$, we obtain

$$\begin{aligned} & \mathbb{P}(\text{SW}_{r,\delta}(P, Q) \notin C_{nm}) \\ &= \mathbb{P}(\text{SW}_{r,\delta}(P, Q) \notin C_{nm}^*, 0 \notin C_{nm}^\dagger) + \mathbb{P}(\text{SW}_{r,\delta}(P, Q) \notin C_{nm}^\dagger, 0 \in C_{nm}^\dagger) \\ &\leq 2\mathbb{P}(0 \notin C_{nm}^\dagger) \leq \alpha + o(1). \end{aligned}$$

This proves the stated asymptotic coverage property of the confidence interval C_{nm} . In order to bound its length, note that it is a direct consequence of Theorem 4 that the bootstrap quantiles $F_{nm}^*(\alpha/2)$ and $F_{nm}^*(1-\alpha/2)$ are of the order $O_p(n^{-1/2})$ as $n/(n+m) \rightarrow a \in (0, 1)$. Thus,

$$b_{nm}^* - a_{nm}^* = O_p(n^{-1/2}),$$

where we write $C_{nm}^* = [(a_{nm}^*)^{1/r}, (b_{nm}^*)^{1/r}]$. Furthermore, under the conditions of Theorem 4, and for $N \asymp n^{r^2}$ and $M_N \asymp \log N$, it can be deduced from Theorem 3 and Corollary 1 that when $\text{SW}_{r,\delta}(P, Q) = 0$, we have

$$b_{nm}^\dagger - a_{nm}^\dagger = O_p\left(\left(\frac{\log n}{n}\right)^{\frac{r}{2}}\right)$$

where we write $C_{nm}^\dagger = [(a_{nm}^\dagger)^{1/r}, (b_{nm}^\dagger)^{1/r}]$. Finally, as in the above proof of coverage of C_{nm} , when $\text{SW}_{r,\delta}(P, Q) > 0$ we have $C_{nm} = C_{nm}^*$ with probability at least $1 - \alpha/2 - o(1)$, while when $\text{SW}_{r,\delta}(P, Q) = 0$ we have $C_{nm} = C_{nm}^\dagger$ with probability at least $1 - \alpha/2 - o(1)$. Combine these facts to deduce that for any $\epsilon > 0$, there exist constants $C, n_0 > 0$ such that for all $n \geq n_0$,

$$b_{nm} - a_{nm} \leq C \left(\left(\frac{\log n}{n} \right)^{\frac{r}{2}} + \frac{\text{SW}_{r,\delta}(P, Q)}{\sqrt{n}} \right)$$

with probability at least $1 - \alpha/2 - \epsilon$. Choosing $\epsilon = \alpha/2$ leads to the claim. $\square$

G.5. Proof of Proposition 8

The proof is straightforward. We have,

$$\mathbb{P}(\eta_0 \notin \bar{C}_{nm}^{(N)}) \leq \mathbb{P}(\bar{\ell}_{nm}^{(N)}(\eta_0) > \epsilon)$$

$$\leq \mathbb{P}\left(\bar{\ell}_{nm}^{(N)}(\eta_0) > \epsilon_0\right) = \mathbb{P}\left(\bar{\ell}_{nm}^{(N)}(\eta_0) > \text{SW}_{r,\delta}(P, P_{\eta_0})\right).$$

The claim now follows from Proposition 6. $\square$

G.6. Proof of Example 2

We begin by proving the validity of the inequality in equation (24). Let $\mathcal{A}$ be a collection of sets, and let $\mathcal{S}_{\mathcal{A}}(n)$ denote the shattering number (Vapnik, 2013) of $\mathcal{A}$. The relative VC inequality is then given by

$$\mathbb{P}\left(\sup_{A \in \mathcal{A}} \frac{|P_n(A) - P(A)|}{\sqrt{P_n(A)}} \geq t\right) \leq 4\mathcal{S}_{\mathcal{A}}(2n)e^{-nt^2/4}, \quad t > 0.$$

Letting $\mathcal{A} = \{(-\infty, x] : x \in \mathbb{R}\}$ and $\mathcal{A} = \{[x, \infty) : x \in \mathbb{R}\}$ respectively, we obtain

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

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

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

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

$$|F_n(x) - F(x)| \leq \nu_{\alpha,n} \sqrt{F_n(x)(1 - F_n(x))}, \quad \forall x \in \mathbb{R}. \quad (80)$$

This proves the validity of equation (24).

We now invert equation (80) 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,

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

with empirical analogues given by

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

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

$$F_n(x) \geq u \implies x \geq F_n^{-1}(u), \quad (81)$$

$$\bar{F}_n(x) \leq u \implies x \leq \bar{F}_n^{-1}(u), \quad (82)$$

$$F(x) \geq u \iff x \geq F^{-1}(u), \quad (83)$$

$$\bar{F}(x) \leq u \iff x \leq \bar{F}^{-1}(u). \quad (84)$$

We now turn to the proof. The calculations which follow are elementary, but tedious. Let $v = F(x)$. By equation (80), we have with probability at least $1 - \alpha/2$ that for all $x \in \mathbb{R}$,

$$\begin{aligned} F_n(x) + \nu_{\alpha,n} \sqrt{F_n(x)(1 - F_n(x))} &\geq v \geq F_n(x) - \nu_{\alpha,n} \sqrt{F_n(x)(1 - F_n(x))} \\ \implies F_n(x)(1 - F_n(x))\nu_{\alpha,n}^2 &\geq (v - F_n(x))^2 \\ \implies (F_n(x) - F_n(x)^2)\nu_{\alpha,n}^2 &\geq v^2 - 2vF_n(x) + F_n(x)^2 \\ \implies F_n(x)^2(1 + \nu_{\alpha,n}^2) - F_n(x)(2v + \nu_{\alpha,n}^2) + v^2 &\leq 0 \\ \implies F_n(x) &\geq \frac{2v + \nu_{\alpha,n}^2}{2(1 + \nu_{\alpha,n}^2)} - \frac{\sqrt{[2v + \nu_{\alpha,n}^2]^2 - 4(1 + \nu_{\alpha,n}^2)v^2}}{2(1 + \nu_{\alpha,n}^2)} \\ \implies F_n(x) &\geq \frac{2v + \nu_{\alpha,n}^2}{2(1 + \nu_{\alpha,n}^2)} - \frac{\nu_{\alpha,n} \sqrt{\nu_{\alpha,n}^2 + 4v(1 - v)}}{2(1 + \nu_{\alpha,n}^2)} = \gamma_{\alpha,n}(v) \\ \implies x &\geq F_n^{-1}(\gamma_{\alpha,n}(v)), \quad (\text{By (81)}) \\ \implies x &\geq F_n^{-1}(\gamma_{\alpha,n}(F(x))). \end{aligned}$$

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 (83), 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 (80), we obtain

$$\bar{F}_n(x) - \nu_{\alpha,n} \sqrt{\bar{F}_n(x)(1 - \bar{F}_n(x))} \leq \bar{F}(x) \leq \bar{F}_n(x) + \nu_{\alpha,n} \sqrt{\bar{F}_n(x)(1 - \bar{F}_n(x))}$$

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

$$\begin{aligned} \nu_{\alpha,n}^2 \bar{F}_n(x)(1 - \bar{F}_n(x)) &\geq (\bar{F}_n(x) - v)^2 \\ \implies \nu_{\alpha,n}^2 \bar{F}_n(x) - \nu_{\alpha,n}^2 \bar{F}_n(x)^2 &\geq \bar{F}_n(x)^2 - 2v\bar{F}_n(x) + v^2 \\ \implies \bar{F}_n(x)^2(1 + \nu_{\alpha,n}^2) - (\nu_{\alpha,n}^2 + 2v)\bar{F}_n(x) + v^2 &\leq 0 \end{aligned}$$

$$\begin{aligned}
&\Rightarrow \bar{F}_n(x) \leq \frac{\nu_{\alpha,n}^2 + 2v}{2(1 + \nu_{\alpha,n}^2)} + \frac{\sqrt{[2v + \nu_{\alpha,n}^2]^2 - 4(1 + \nu_{\alpha,n}^2)v^2}}{2(1 + \nu_{\alpha,n}^2)} \\
&\Rightarrow \bar{F}_n(x) \leq \frac{\nu_{\alpha,n}^2 + 2v + \nu_{\alpha,n}\sqrt{\nu_{\alpha,n}^2 + 4v(1 - v)}}{2(1 + \nu_{\alpha,n}^2)} = \eta_{\alpha,n}(v) \\
&\Rightarrow x \leq \bar{F}_n^{-1}(\eta_{\alpha,n}(v)), \quad (\text{By (82)}) \\
&\Rightarrow x \leq \bar{F}_n^{-1}(\eta_{\alpha,n}(\bar{F}(x))).
\end{aligned}$$

Therefore, setting $x = \bar{F}^{-1}(u)$ for $u \in (0, 1)$, and using the fact that $\bar{F} \circ \bar{F}^{-1}(u) \leq u$ by equation (84), we obtain

$$\bar{F}^{-1}(u) \leq \bar{F}_n^{-1}(\eta_{\alpha,n}(\bar{F}(\bar{F}^{-1}(u)))) \leq \bar{F}_n^{-1}(\eta_{\alpha,n}(u)).$$

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 (25) follows. $\square$

Appendix H: Further interpretation of Theorem 3

We now briefly discuss Theorem 3 in the situation where $\text{SJ}_{r,\delta/2}(P) < \infty$ but $\text{SJ}_{r,\delta/2}(Q) = \infty$. 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 $\text{SJ}_{r,\delta/2}(P) < \infty, \text{SJ}_{r,\delta/2}(Q) = \infty$ need not be violated so long as $P \neq Q$. In this specialized situation, Corollary 1 merely implies the nonparametric rate $n^{-\frac{1}{2r}} + m^{-\frac{1}{2r}}$, though one may expect a better dependence on n as a result of the condition $\text{SJ}_{r,\delta/2}(P) < \infty$, as well as a better dependence on m by virtue of Q being in proximity to P .

We focus on the one-dimensional case for simplicity, and we omit superscripts depending on $N = 1$ below. In this setting, upper bounds on the ∞ -Wasserstein distance in terms of r -Wasserstein distances have been established by Bouchitté, Jimenez and Rajesh (2007): If P is absolutely continuous with respect to the Lebesgue measure, with strictly positive density p over a compact set $\mathcal{X} \subseteq \mathbb{R}$, then for all $r \geq 1$,

$$W_{\infty}^{r+1}(P, Q) \leq C_r(\mathcal{X}) W_r^r(P, Q) \sup_{x \in \mathcal{X}} \left| \frac{1}{p(x)} \right|, \quad (85)$$

where $C_r(\mathcal{X}) > 0$ is a constant depending only on $r, \mathcal{X}$. Equipped with this result, we arrive at the following Corollary to Theorem 3.

Corollary 3. Let $P, Q \in \mathcal{P}(\mathbb{R})$ be absolutely continuous with respect to the Lebesgue measure, and assume the same conditions and notation as Theorem 3. Assume there exists a fixed constant $s > 0$ such that $J_{\infty, \delta/2}(P) \leq s$, and let δ be fixed. Assume further that $W_{r, \delta}(P, Q) \lesssim \kappa_{\epsilon, n \wedge m}$. Then, there exists a constant $C > 0$ depending on b, r, δ, s such that with probability at least $1 - \epsilon$, we have

$$\lambda(C_{nm}^{(N)}) \leq C [\kappa_{\epsilon, n \wedge m}]^{\frac{r^2+1}{r(r+1)}}.$$

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 rate $n^{-\frac{5}{12}} + m^{-\frac{5}{12}}$, which is strictly improves the $n^{-\frac{1}{4}} + m^{-\frac{1}{4}}$ 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^δ, Q^δ denote the δ -trimmings of P and Q , in the sense of equation (5). The quantile functions of P^δ, Q^δ 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^δ, Q_h^δ 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 $h \leq \delta/2$ by assumption. Then, by definition of $U_{\epsilon, m}$,

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

Furthermore, the finiteness of $J_{\infty, \delta/2}(P)$ implies $W_\infty(P^\delta, P_h^\delta) \lesssim h$, and we have,

$$\begin{aligned} W_r(P_h^\delta, Q_h^\delta) &= W_r(P^\delta, Q^\delta) + \int_{\delta+h}^{1-\delta+h} |F^{-1} - G^{-1}|^r - \int_\delta^{1-\delta} |F^{-1} - G^{-1}|^r \\ &\lesssim W_r(P^\delta, Q^\delta) + h. \end{aligned}$$

Combining these facts, together with equation (85) and the bound $J_{\infty, \delta/2}(P) \leq s$, we arrive at

$$\begin{aligned} U_{\epsilon, m}^{\frac{1}{r-1}}(Q) &\leq W_\infty(Q^\delta, Q_h^\delta) \\ &\leq W_\infty(Q^\delta, P^\delta) + W_\infty(P^\delta, P_h^\delta) + W_\infty(P_h^\delta, Q_h^\delta) \\ &\lesssim W_r^{\frac{r}{r+1}}(P^\delta, Q^\delta) + h + W_r^{\frac{r}{r+1}}(P_h^\delta, Q_h^\delta) \\ &\lesssim W_{r, \delta}^{\frac{r}{r+1}}(P, Q) + h^{\frac{r}{r+1}}. \end{aligned}$$

Thus, by Theorem 3, under the assumption $W_{r, \delta}(P, Q) \lesssim h$, we have with probability at least $1 - \epsilon$,

$$\begin{aligned} \lambda(C_{nm}) &\lesssim \left\{ W_{r, \delta}^r(P, Q) + h \left(W_{\infty, \delta}^{r-1}(P, Q) + U_{\epsilon, n}(P) + U_{\epsilon, m}(Q) \right) \right\}^{1/r} - W_{r, \delta}(P, Q) \end{aligned}$$

$$\begin{aligned}
&\lesssim W_{r,\delta}(P, Q) \left\{ \left[1 + \frac{h}{W_{r,\delta}^r(P, Q)} \left(W_{r,\delta}^{\frac{r(r-1)}{r+1}}(P, Q) + h^{r-1} + h^{\frac{r(r-1)}{r+1}} \right) \right]^{\frac{1}{r}} - 1 \right\} \\
&\lesssim W_{r,\delta}(P, Q) \left(\frac{h^{\frac{r^2+1}{r+1}}}{W_{r,\delta}^r(P, Q)} \right)^{1/r} \asymp h^{\frac{r^2+1}{r(r+1)}}.
\end{aligned}$$

The claim follows. $\square$

Acknowledgments

The authors would like to thank an anonymous referee for thoughtful comments which significantly helped improve the content and presentation of this paper. Tudor Manole would like to thank Niccolò Dalmaso for conversations which inspired the likelihood-free inference application in Section 7, and Arun Kumar Kuchibhotla for suggesting references on concentration inequalities for self-normalized empirical processes.

References

- ÁLVAREZ-ESTEBAN, P. C., DEL BARRIO, E., CUESTA-ALBERTOS, J. A. and MATRAN, C. (2008). Trimmed Comparison of Distributions. *Journal of the American Statistical Association* **103** 697–704. [MR2435470](#)
- ARJOVSKY, M., CHINTALA, S. and BOTTOU, L. (2017). Wasserstein Generative Adversarial Networks. In *Proceedings of the 34th International Conference on Machine Learning* 214–223.
- BERNTON, E., JACOB, P. E., GERBER, M. and ROBERT, C. P. (2019a). Approximate Bayesian Computation with the Wasserstein Distance. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **81** 235–269. [MR3928142](#)
- BERNTON, E., JACOB, P. E., GERBER, M. and ROBERT, C. P. (2019b). On Parameter Estimation with the Wasserstein Distance. *Information and Inference: A Journal of the IMA* **8** 657–676. [MR4049226](#)
- BERTHET, P., FORT, J.-C. and KLEIN, T. (2020). A Central Limit Theorem for Wasserstein Type Distances between Two Distinct Univariate Distributions. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques* **56** 954–982. [MR4076772](#)
- BOBKOV, S. and LEDOUX, M. (2019). One-Dimensional Empirical Measures, Order Statistics, and Kantorovich Transport Distances. *Memoirs of the American Mathematical Society* **261**. [MR4028181](#)
- BOGACHEV, V. I. (2007). *Measure Theory* **1**. Springer-Verlag, Berlin, Germany. [MR2267655](#)
- BOISSARD, E. and LE GOUIC, T. (2014). On the Mean Speed of Convergence of Empirical and Occupation Measures in Wasserstein Distance. In *Annales de l'Institut Henri Poincaré Probabilités et Statistiques* **50** 539–563. [MR3189084](#)

- BONASSI, F. V. and WEST, M. (2015). Sequential Monte Carlo with Adaptive Weights for Approximate Bayesian Computation. *Bayesian Analysis* **10** 171–187. [MR3420901](#)
- BONASSI, F. V., YOU, L. and WEST, M. (2011). Bayesian Learning from Marginal Data in Bionetwork Models. *Statistical Applications in Genetics and Molecular Biology* **10**. [MR2851291](#)
- BONNEEL, N., RABIN, J., PEYRÉ, G. and PFISTER, H. (2015). Sliced and Radon Wasserstein Barycenters of Measures. *Journal of Mathematical Imaging and Vision* **51** 22–45. [MR3300482](#)
- BONNOTTE, N. (2013). Unidimensional and Evolution Methods for Optimal Transportation, PhD thesis, Paris 11.
- BOUCHITTÉ, G., JIMENEZ, C. and RAJESH, M. (2007). A New L^∞ Estimate in Optimal Mass Transport. *Proceedings of the American Mathematical Society* **135** 3525–3535. [MR2336567](#)
- BOUSQUET, O., BOUCHERON, S. and LUGOSI, G. (2003). Introduction to Statistical Learning Theory. In *Summer School on Machine Learning* 169–207. Springer. [MR2001870](#)
- BREHMER, J., KLING, F., ESPEJO, I. and CRANMER, K. (2020). MadMiner: Machine Learning-Based Inference for Particle Physics. *Computing and Software for Big Science* **4** 1–25.
- CHEN, Q. and FANG, Z. (2019). Inference on Functionals under First Order Degeneracy. *Journal of Econometrics* **210** 459–481. [MR3958414](#)
- COURTY, N., FLAMARY, R., TUIA, D. and RAKOTOMAMONJY, A. (2016). Optimal Transport for Domain Adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **39** 1853–1865.
- CSORGO, M. and REVESZ, P. (1978). Strong Approximations of the Quantile Process. *The Annals of Statistics* **6** 882–894. [MR0501290](#)
- DALMASSO, N., IZBICKI, R. and LEE, A. B. (2020). Confidence Sets and Hypothesis Testing in a Likelihood-Free Inference Setting. In *Proceedings of the 37th International Conference on Machine Learning* **119** 2323–2334. PMLR.
- DALMASSO, N., LEE, A., IZBICKI, R., POSPISIL, T., KIM, I. and LIN, C.-A. (2020). Validation of Approximate Likelihood and Emulator Models for Computationally Intensive Simulations. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics* **108** 3349–3361. PMLR.
- DEL BARRIO, E., GINÉ, E. and 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. [MR2121458](#)
- DEL BARRIO, E., GONZÁLEZ-SANZ, A. and LOUBES, J.-M. (2021). Central Limit Theorems for General Transportation Costs. *arXiv preprint arXiv:2102.06379*.
- DEL BARRIO, E., GORDALIZA, P. and LOUBES, J.-M. (2019). A Central Limit Theorem for L_p Transportation Cost on the Real Line with Application to Fairness Assessment in Machine Learning. *Information and Inference: A Journal of the IMA* **8** 817–849. [MR4045479](#)
- DEL BARRIO, E. and LOUBES, J.-M. (2019). Central Limit Theorems for Empirical Transportation Cost in General Dimension. *The Annals of Probability*

- 47 926–951. [MR3916938](#)
- DEL BARRIO, E., CUESTA-ALBERTOS, J. A., MATRÁN, C. and RODRÍGUEZ-RODRÍGUEZ, J. M. (1999). Tests of Goodness of Fit Based on the L_2 -Wasserstein Distance. *The Annals of Statistics* **27** 1230–1239. [MR1740113](#)
- DENKER, M. (1985). *Asymptotic Distribution Theory in Nonparametric Statistics*. Braunschweig-Wiesbaden: Vieweg. [MR0889896](#)
- DESHPANDE, I., HU, Y.-T., SUN, R., PYRROS, A., SIDDIQUI, N., KOYEJO, S., ZHAO, Z., FORSYTH, D. and 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* 10648–10656.
- DIACONIS, P., HOLMES, S. and SHAHSHAHANI, M. (2013). Sampling from a Manifold. In *Advances in Modern Statistical Theory and Applications: A Festschrift in Honor of Morris L. Eaton* 102–125. Institute of Mathematical Statistics. [MR3586941](#)
- DUDLEY, R. M. (1969). The Speed of Mean Glivenko-Cantelli Convergence. *The Annals of Mathematical Statistics* **40** 40–50. [MR0236977](#)
- DVORETZKY, A., KIEFER, J. and 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. [MR0083864](#)
- EBERT, A., DUTTA, R., MENGENSEN, K., MIRA, A., RUGGERI, F. and WU, P. (2021). Likelihood-Free Parameter Estimation for Dynamic Queueing Networks: Case Study of Passenger Flow in an International Airport Terminal. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **70** 770–792. [MR4275846](#)
- EFRON, B. and TIBSHIRANI, R. J. (1994). *An Introduction to the Bootstrap*. CRC press. [MR1270903](#)
- FOURNIER, N. and GUILLIN, A. (2015). On the Rate of Convergence in Wasserstein Distance of the Empirical Measure. *Probability Theory and Related Fields* **162** 707–738. [MR3383341](#)
- FREITAG, G., CZADO, C. and 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. [MR2301710](#)
- FREITAG, G., MUNK, A. and VOGT, M. (2003). Assessing Structural Relationships between Distributions—a Quantile Process Approach Based on Mallows Distance. In *Recent Advances and Trends in Nonparametric Statistics* 123–137. Elsevier. [MR2498237](#)
- FREITAG, G. and 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. [MR2161214](#)
- GANGBO, W. and MCCANN, R. J. (1996). The Geometry of Optimal Transportation. *Acta Mathematica* **177** 113–161. [MR1440931](#)
- GINÉ, E. and KOLTCHINSKII, V. (2006). Concentration Inequalities and Asymptotic Results for Ratio Type Empirical Processes. *The Annals of Probability*

- 34** 1143–1216. [MR2243881](#)
- GUTMANN, M. U., DUTTA, R., KASKI, S. and CORANDER, J. (2018). Likelihood-Free Inference via Classification. *Statistics and Computing* **28** 411–425. [MR3747571](#)
- HO, N., YANG, C.-Y. and JORDAN, M. I. (2019). Convergence Rates for Gaussian Mixtures of Experts. *arXiv preprint arXiv:1907.04377*.
- HO, N., NGUYEN, X., YUROCHKIN, M., BUI, H. H., HUYNH, V. and PHUNG, D. (2017). Multilevel Clustering via Wasserstein Means. In *Proceedings of the 34th International Conference on Machine Learning* 1501–1509.
- HUNDRIESER, S., KLATT, M., STAUDT, T. and MUNK, A. (2022). A Unifying Approach to Distributional Limits for Empirical Optimal Transport. *arXiv preprint arXiv:2202.12790*.
- IMAIZUMI, M., OTA, H. and HAMAGUCHI, T. (2019). Hypothesis Test and Confidence Analysis with Wasserstein Distance on General Dimension. *arXiv preprint arXiv:1910.07773*.
- 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* **84** 1711–1721. PMLR.
- KIM, I., BALAKRISHNAN, S. and WASSERMAN, L. (2020). Robust Multivariate Nonparametric Tests via Projection Averaging. *The Annals of Statistics* **48** 3417–3441. [MR4185814](#)
- KLATT, M., MUNK, A. and ZEMEL, Y. (2020). Limit Laws for Empirical Optimal Solutions in Stochastic Linear Programs. *arXiv preprint arXiv:2007.13473*.
- KLATT, M., TAMELING, C. and MUNK, A. (2020). Empirical Regularized Optimal Transport: Statistical Theory and Applications. *SIAM Journal on Mathematics of Data Science* **2** 419–443. [MR4105566](#)
- KOLOURI, S., PARK, S. R., THORPE, M., SLEPCEV, D. and ROHDE, G. K. (2017). Optimal Mass Transport: Signal Processing and Machine-Learning Applications. *IEEE Signal Processing Magazine* **34** 43–59.
- KOLOURI, S., NADJAH, K., SIMSEKLI, U., BADEAU, R. and ROHDE, G. (2019). Generalized Sliced Wasserstein Distances. In *Advances in Neural Information Processing Systems* **32** 261–272.
- LE, T., YAMADA, M., FUKUMIZU, K. and CUTURI, M. (2019). Tree-Sliced Variants of Wasserstein Distances. In *Advances in Neural Information Processing Systems* **32** 12283–12294.
- LEI, J. (2020). Convergence and Concentration of Empirical Measures under Wasserstein Distance in Unbounded Functional Spaces. *Bernoulli* **26** 767–798. [MR4036051](#)
- LIANG, T. (2019). On the Minimax Optimality of Estimating the Wasserstein Metric. *arXiv preprint arXiv:1908.10324*.
- LOTKA, A. J. (1920a). Undamped Oscillations Derived from the Law of Mass Action. *Journal of the American Chemical Society* **42** 1595–1599.
- LOTKA, A. J. (1920b). Analytical Note on Certain Rhythmic Relations in Organic Systems. *Proceedings of the National Academy of Sciences* **6** 410–415.
- MASSART, P. (1990). The Tight Constant in the Dvoretzky-Kiefer-Wolfowitz

- Inequality. *The Annals of Probability* 1269–1283. [MR1062069](#)
- MUNK, A. and 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. [MR1625620](#)
- NADJAH, K., DE BORTOLI, V., DURMUS, A., BADEAU, R. and ŞİMŞEKLI, U. (2020). Approximate Bayesian Computation with the Sliced-Wasserstein Distance. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* 5470–5474.
- NGUYEN, X. (2013). Convergence of Latent Mixing Measures in Finite and Infinite Mixture Models. *The Annals of Statistics* **41** 370–400. [MR3059422](#)
- NGUYEN, K., HO, N., PHAM, T. and BUI, H. (2020). Distributional Sliced-Wasserstein and Applications to Generative Modeling. In *International Conference on Learning Representations*.
- NILES-WEED, J. and RIGOLLET, P. (2022). Estimation of Wasserstein Distances in the Spiked Transport Model. *To appear, Bernoulli*. *arXiv preprint arXiv:1909.07513*.
- PANARETOS, V. M. and ZEMEL, Y. (2019a). *An Invitation to Statistics in Wasserstein Space*. Springer Nature. [MR4350694](#)
- PANARETOS, V. M. and ZEMEL, Y. (2019b). Statistical Aspects of Wasserstein Distances. *Annual Review of Statistics and Its Application* **6** 405–431. [MR3939527](#)
- PARK, M., JITKRITTUM, W. and SEJDINOVIC, D. (2016). K2-ABC: Approximate Bayesian Computation with Kernel Embeddings. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics* 398–407.
- PATY, F.-P. and CUTURI, M. (2019). Subspace Robust Wasserstein Distances. In *Proceedings of the 36th International Conference on Machine Learning* 5072–5081.
- PEYRÉ, G. and CUTURI, M. (2019). Computational Optimal Transport. *Foundations and Trends® in Machine Learning* **11** 355–607.
- POLLARD, D. (2002). *A User's Guide to Measure Theoretic Probability* **8**. Cambridge University Press. [MR1873379](#)
- RABIN, J., PEYRÉ, G., DELON, J. and BERNOT, M. (2011). Wasserstein Barycenter and its Application to Texture Mixing. In *International Conference on Scale Space and Variational Methods in Computer Vision* 435–446. Springer.
- RAMDAS, A., TRILLOS, N. and CUTURI, M. (2017). On Wasserstein Two-Sample Testing and Related Families of Nonparametric Tests. *Entropy* **19** 47. [MR3608466](#)
- REVUZ, D. and YOR, M. (2013). *Continuous Martingales and Brownian Motion* **293**. Springer Science & Business Media. [MR3184878](#)
- RIPPL, T., MUNK, A. and STURM, A. (2016). Limit Laws of the Empirical Wasserstein Distance: Gaussian Distributions. *Journal of Multivariate Analysis* **151** 90–109. [MR3545279](#)
- SHAO, J. and TU, D. (2012). *The Jackknife and Bootstrap*. Springer Science & Business Media. [MR1351010](#)

- SHORACK, G. R. and WELLNER, J. A. (2009). *Empirical Processes with Applications to Statistics*. SIAM. [MR3396731](#)
- SINGH, S. and PÓCZOS, B. (2019). Minimax Distribution Estimation in Wasserstein Distance. *arXiv preprint arXiv:1802.08855*.
- SISSON, S. A., FAN, Y. and BEAUMONT, M. (2018). *Handbook of Approximate Bayesian Computation*. Chapman and Hall/CRC. [MR3889281](#)
- SOMMERFELD, M. and MUNK, A. (2018). Inference for Empirical Wasserstein Distances on Finite Spaces. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **80** 219–238. [MR3744719](#)
- STEIN, E. M. and SHAKARCHI, R. (2009). *Real Analysis: Measure Theory, Integration, and Hilbert Spaces*. Princeton University Press. [MR2129625](#)
- TAMELING, C., SOMMERFELD, M. and MUNK, A. (2019). Empirical Optimal Transport on Countable Metric Spaces: Distributional Limits and Statistical Applications. *The Annals of Applied Probability* **29** 2744–2781. [MR4019874](#)
- TSYBAKOV, A. B. (2008). *Introduction to Nonparametric Estimation*. Springer Science & Business Media. [MR2724359](#)
- VAN DER VAART, A. (1998). *Asymptotic Statistics. Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, UK; New York, NY, USA. [MR1652247](#)
- VAN DER VAART, A. W. and WELLNER, J. A. (1996). *Weak Convergence and Empirical Processes*. Springer. [MR1385671](#)
- VAPNIK, V. (2013). *The Nature of Statistical Learning Theory*. Springer Science & Business Media. [MR1367965](#)
- VERDINELLI, I. and WASSERMAN, L. (2019). Hybrid Wasserstein Distance and Fast Distribution Clustering. *Electronic Journal of Statistics* **13** 5088–5119. [MR4041703](#)
- VERDINELLI, I. and WASSERMAN, L. (2021). Decorrelated Variable Importance. *arXiv preprint arXiv:2111.10853*.
- VILLANI, C. (2003). *Topics in Optimal Transportation*. American Mathematical Soc. [MR1964483](#)
- WEED, J. and BACH, F. (2019). Sharp Asymptotic and Finite-Sample Rates of Convergence of Empirical Measures in Wasserstein Distance. *Bernoulli* **25** 2620–2648. [MR4003560](#)