SLICED ORLICZ-WASSERSTEIN

Anonymous authors
Paper under double-blind review

ABSTRACT

We propose sliced Orlicz-Wasserstein (SOW) distance which is a generalization of sliced Wasserstein (SW) distance. SOW replaces the L^p norm in SW with a Luxemburg norm cost induced by an Orlicz function ϕ. First, we prove that SOW distance is a metric on the space of measures with finite Orlicz norm, and show that it recovers the SW distance when the Orlicz function is $\phi(x) = x^p$. Next, we derive the topological properties of the SOW distance. In particular, we show that convergence under SOW implies weak convergence, and the converse is true under the compact support condition. We then present the theoretical results for estimating the SOW distance. We derive sample complexity for both the distance itself and the powered functional of the distance, and prove their minimax optimality. In addition, we discuss the computational algorithm for approximating the SOW distance by Monte-Carlo estimation and bisection search, as well as the associated approximation error and computational complexity analysis. Our experimental results reveal the superior computational efficiency of SOW compared with Orlicz-Wasserstein (OW) distance. Also, in the experiments, we demonstrate the favorable flexibility of SOW distance over SW in detecting differences between distributions by comparing their performance in two-sample tests and evaluating generative models on image datasets.

1 INTRODUCTION

Comparing probability measures is the most fundamental problem in statistics, machine learning, and data sciences. Among statistical distances, the Wasserstein distance (Villani, 2003; 2008) has recently attracted a lot of attention due to its rich geometrical structure from optimal transport (OT). Within generative modeling, it has served to improve generative adversarial networks (Arjovsky et al., 2017), flow-based approaches (Lipman et al., 2023), and drifting models (He et al., 2026). It has likewise been employed for matching source and target distributions in domain adaptation (Courty et al., 2017). It also plays a significant role in computational biology (Schiebinger et al., 2019), image processing (Feydy et al., 2017), signal processing (Kolouri et al., 2017), computer graphics (Solomon et al., 2016; 2015), statistical inference (Bernton et al., 2019a;b), and the measurement of dependence (Catalano et al., 2021; 2024), among numerous other domains.

Wasserstein distance of order p aggregates transport cost by averaging the p-th power of the ground distance. Therefore, a single exponent controls how much weight is placed on small versus large displacements. Replacing this power function by an arbitrary Orlicz function (see Section 2) yields Orlicz-Wasserstein (OW) distance (Sturm, 2011; Kell, 2017), which measures the cost of a coupling through a Luxemburg-type norm. Since the power function is itself an Orlicz function, this family strictly generalizes Wasserstein distance. The choice of the Orlicz function controls how transported mass is weighted through the associated Luxemburg norm. In particular, as illustrated by equation 5, different growth regimes of ϕ induce different sensitivities to discrepancies carried by a small fraction of mass. This flexibility has already led to improved theoretical guarantees for parameter estimation in mixture models, where the mass a posteriori assigned to atoms far from the true support vanishes only logarithmically under Wasserstein distance, but at an almost polynomial rate under a suitable OW distance (Guha et al., 2023). Moreover, the work of Altschuler & Chewi (2024) leverages OW to establish fast convergence of hypocoercive differential equations.

In this work, we focus on discussing the computational aspect of OW which is an underexplored area. As discussed in Guha et al. (2023), computing OW requires using a bisection search where each iteration evaluates a Wasserstein distance at least once. Therefore, OW becomes even more

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

computationally expensive than Wasserstein, which is already computationally demanding. For better context, the worst-case time complexity of Wasserstein distance is $\mathcal{O}(n^3 \log n)$ (Peyré et al., 2019) in the discrete case with at most n atoms. Guha et al. (2023) mitigate the problem by adapting entropic regularization (Cuturi, 2013) to the case of OW, which reduces the computational complexity for each iteration to $\mathcal{O}(n^2)$. However, this quadratic complexity is not scalable enough for a very large n. As a result, it is desirable to have a faster alternative.

We propose sliced Orlicz-Wasserstein (SOW) distance which achieves a quasi-linear time complexity of $\mathcal{O}(n \log n)$ in the discrete setting. SOW is a sliced version of OW analogous distance to sliced Wasserstein (Rabin et al., 2012) (SW) distance. Similar to SW, SOW leverages the closed-form of one-dimensional OT (Cambanis et al., 1976) which accelerates computation of one-dimensional OW. In addition to the appealing computational complexity, SOW also enjoys a dimension-independent sample complexity as SW under the compact support assumption, in contrast to the curse of dimensionality problem of Wasserstein distance (Fournier & Guillin, 2015). With the ability to sharpen control over tail behavior, SOW can be a suitable replacement for SW in settings where large displacements should be penalized more aggressively than the polynomial cost of SW allows, e.g., when the two measures differ mainly in their tails or in low-mass modes that SW is known to under-weight. More broadly, the choice of the Orlicz function makes SOW a family rather than a single metric. In particular, SOW recovers SW_p when $\phi(t) = t^p$, while more general choices of ϕ allow the sensitivity to small amounts of transported mass to depart from the fixed power-law behavior imposed by Wasserstein distances. SOW extends the family of sliced optimal transport metrics (Nguyen, 2025) and the family of sliced probability metrics (Nadjahi et al., 2020). With existing broad applications of sliced optimal transport metrics, e.g., generative modeling (Deshpande et al., 2018), clustering (Kolouri et al., 2018b), representation learning (Kolouri et al., 2018a), and many others, SOW can serve as a new choice in these applications. In summary, our contributions are threefold:

1. We establish a quantile representation for the one-dimensional OW distance and use it to introduce SOW, obtained by aggregating projected one-dimensional OW distances. The proposed distance recovers SW as a special case.

2. Beyond basic metric properties, we develop the topological and statistical theory of SOW. We show that SOW convergence implies weak convergence, with equivalence under compact support. Under the compact-support assumption, we derive dimension-independent sample complexities for both SOW and its powered functional, and prove that the resulting rates are minimax optimal. The powered-functional case requires additional perturbation and minimax lower-bound arguments beyond those used for standard SW. We provide a proof roadmap in Appendix I, highlighting where the classical SW arguments break down and the additional ingredients needed in the general Orlicz setting. Besides, we also take the first step beyond compact support by identifying a necessary tail integrability condition for empirical SOW convergence.

3. We provide a practical approximation scheme based on Monte Carlo projections and bisection, together with approximation-error and computational-complexity guarantees. Experiments demonstrate substantially improved scalability over OW and show that suitable choices of the Orlicz function can increase sensitivity to low-mass or large-displacement discrepancies in two-sample testing and generative model evaluation.

Organization. The remainder of the paper is organized as follows. We first review basic definitions of Wasserstein distance, SW distance, Orlicz space and OW distance, and fundamental functions in Section 2. We then discuss SOW, including its definitions, theoretical properties, and computations in Section 3. Experiments and corresponding discussions are given in Section 4. We conclude the paper and discuss some future works in Section 5. Technical proofs are provided in the Appendices.

Notation. Throughout, $\|\cdot\|$ denotes the Euclidean norm. For $d \geq 1$, we write $\mathbb{S}^{d-1} := \left\{\theta \in \mathbb{R}^d : \|\theta\| = 1\right\}$ for the unit sphere in $\mathbb{R}^d$ and let σ denote the uniform probability measure on $\mathbb{S}^{d-1}$. Given $\theta \in \mathbb{S}^{d-1}$, let $\Pi^\theta : \mathbb{R}^d \to \mathbb{R}$ denote the projection $\Pi^\theta(x) = \theta^\top x$. For a measurable map f and a measure μ, the pushforward of μ through f is denoted by $f_\# \mu := \mu \circ f^{-1}$. For a Borel set A in a Euclidean space, let $\mathcal{P}(A)$ denote the set of Borel probability measures on A. For $\mu \in \mathcal{P}(\mathbb{R})$, let F_μ denote its cumulative distribution function and for $u \in (0, 1)$, define its quantile function by $F_\mu^{-1}(u) := \inf\{x \in \mathbb{R} : F_\mu(x) \geq u\}$. Weak convergence of μ_n to μ is denoted by $\mu_n \rightharpoonup \mu$, and δ_x denotes the Dirac measure at x. For probability measures μ and ν on the same

Euclidean space, $\Pi(\mu,\nu)$ denotes the set of their couplings. For two positive sequences $(a_n)_{n\geq 1}$ and $(b_n)_{n\geq 1}$, we write $a_n = \mathcal{O}(b_n)$, or equivalently $a_n \lesssim b_n$, if there exists a constant $C > 0$, independent of n, such that $a_n \leq Cb_n$ for every $n \geq 1$. We write $a_n \asymp b_n$ if both $a_n \lesssim b_n$ and $b_n \lesssim a_n$. Subscripts indicate the parameters on which the implicit constants may depend. Increasing and decreasing are understood in the weak sense unless explicitly qualified as strict.

2 BACKGROUND

We first review the Wasserstein and SW distances, and then recall the basic notions of Orlicz spaces and the OW distance used throughout the paper.

Wasserstein distance. For $p \geq 1$, let $\mathcal{P}_p(\mathbb{R}^d)$ denote the set of probability measures on $\mathbb{R}^d$ with finite p-th moment. The p-Wasserstein distance (Villani, 2008) between $\mu,\nu \in \mathcal{P}_p(\mathbb{R}^d)$ is defined as follows:

$$\mathrm{W}_p^p(\mu,\nu) \coloneqq \inf_{\pi\in\Pi(\mu,\nu)} \int_{\mathbb{R}^d\times\mathbb{R}^d} \|x-y\|^p \mathrm{d}\pi(x,y). \tag{1}$$

SW distance. SW leverages the fact that the Wasserstein distance in one dimension admits an explicit quantile representation. In particular, for $\mu,\nu \in \mathcal{P}_p(\mathbb{R})$, the p-Wasserstein distance admits the quantile representation $\mathrm{W}_p^p(\mu,\nu) = \int_0^1 \left|F_\mu^{-1}(u) - F_\nu^{-1}(u)\right|^p \mathrm{d}u$ (Bobkov & Ledoux, 2019, Theorem 2.10). For $\mu,\nu \in \mathcal{P}_p(\mathbb{R}^d)$, the p-sliced Wasserstein distance (Rabin et al., 2012; Bonneel et al., 2015; Nguyen, 2025) is defined by

$$\mathrm{SW}_p^p(\mu,\nu) \coloneqq \int_{\mathbb{S}^{d-1}} \mathrm{W}_p^p\left(\Pi_\#^\theta\mu, \Pi_\#^\theta\nu\right) \mathrm{d}\sigma(\theta). \tag{2}$$

Thus, SW compares high-dimensional probability measures through their one-dimensional linear projections, allowing each projected Wasserstein distance to be computed using the explicit one-dimensional quantile formula. In practice, the spherical integral is approximated using L directions sampled independently from σ which is the uniform distribution over $\mathbb{S}^{d-1}$.

Orlicz spaces and OW distance. We first recall Orlicz spaces and the Luxemburg norm. For further details on Orlicz spaces, we refer the reader to Rao & Ren (1991); Rubshtein et al. (2016). We consider an Orlicz function[1] $\phi : [0,+\infty) \to [0,+\infty)$ that is strictly increasing and convex, with $\phi(0) = 0$. Given a measure space (Ω, m) and an Orlicz function ϕ, the Orlicz space $L^\phi(\Omega, m)$ consists of measurable functions f such that $\int_\Omega \phi(|f|/\lambda)\mathrm{d}m < \infty$ for some $\lambda > 0$. It is equipped with the Luxemburg norm (Luxemburg et al., 1955) defined by $\|f\|_\phi \coloneqq \inf\left\{\lambda > 0 : \int_\Omega \phi\left(\frac{|f|}{\lambda}\right)\mathrm{d}m \leq 1\right\}$, where the underlying measure m is omitted from the notation whenever it is clear from context. When $\phi(t) = t^p$ with $p \geq 1$, one has $L^\phi(\Omega,m) = L^p(\Omega,m)$, and the Luxemburg norm coincides with the usual L^p norm (Rubshtein et al., 2016, Section 13.4). To define the corresponding transport distance, let

$$\mathcal{P}_\phi(\mathbb{R}^d) \coloneqq \left\{P \in \mathcal{P}(\mathbb{R}^d) : \int_{\mathbb{R}^d} \phi\left(\frac{\|x\|}{\lambda}\right)\mathrm{d}P(x) < \infty \text{ for some } \lambda > 0\right\}.$$

For $\mu,\nu \in \mathcal{P}_\phi\left(\mathbb{R}^d\right)$ and $\pi \in \Pi(\mu,\nu)$, the transport displacement is the function $(x,y) \mapsto \|x-y\|$ on the measure space $(\mathbb{R}^d\times\mathbb{R}^d, \pi)$. The OW distance (Sturm, 2011; Kell, 2017; Guha et al., 2023) is obtained by minimizing the Luxemburg norm of this displacement function over all couplings:

$$\mathrm{W}_\phi(\mu,\nu) \coloneqq \inf_{\pi\in\Pi(\mu,\nu)} \inf\left\{\lambda > 0 : \int_{\mathbb{R}^d\times\mathbb{R}^d} \phi\left(\frac{\|x-y\|}{\lambda}\right)\mathrm{d}\pi(x,y) \leq 1\right\}. \tag{3}$$

When $\phi(t) = t^p$ for $p \geq 1$, one has $\mathcal{P}_\phi(\mathbb{R}^d) = \mathcal{P}_p(\mathbb{R}^d)$ and $\mathrm{W}_\phi = \mathrm{W}_p$.

Fundamental function. By Rubshtein et al. (2016, Proposition 13.3.1), when m is a probability measure, the fundamental function associated with the Luxemburg norm is given by

$$\omega_\phi(u) \coloneqq \frac{1}{\phi^{-1}(1/u)}, \qquad u \in (0,1], \qquad \omega_\phi(0) \coloneqq 0. \tag{4}$$

[1] Definitions of Orlicz functions vary in the literature. For instance, Rubshtein et al. (2016, Definition 13.1.1) allows Orlicz functions to take the value $+\infty$ and does not require strict monotonicity. Following the OW framework of Sturm (2011), we restrict attention to finite-valued, strictly increasing Orlicz functions.

In the one-dimensional OW setting, this fundamental function admits a direct transport interpretation. Indeed, for any $\Delta > 0$ and $u \in [0, 1]$, a direct calculation gives

$$\mathrm{W}_\phi\left(\delta_0, (1-u)\delta_0 + u\delta_\Delta\right) = \Delta\, \omega_\phi(u). \tag{5}$$

This identity separates mass and distance: $\omega_\phi(u)$ captures the effect of the transported mass u, while Δ represents the transport displacement.

3 Sliced Orlicz-Wasserstein (SOW)

We now introduce SOW and establish its topological, statistical, and computational properties.

3.1 Definition

We begin by discussing the one-dimensional OW distance, which serves as the basic building block of the later sliced construction.

One-dimensional OW distance. Let $\mu, \nu \in \mathcal{P}_\phi(\mathbb{R})$. Since ϕ is strictly increasing and convex, for each $\lambda > 0$, the cost function $(x, y) \mapsto \phi(|x - y|/\lambda)$ satisfies the conditions of Cambanis et al. (1976, Theorem 2). Thus,

$$\inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathbb{R}\times\mathbb{R}} \phi\left(\frac{|x-y|}{\lambda}\right) \mathrm{d}\pi(x, y) = \int_0^1 \phi\left(\frac{|F_\mu^{-1}(u) - F_\nu^{-1}(u)|}{\lambda}\right) \mathrm{d}u. \tag{6}$$

Therefore,

$$\mathrm{W}_\phi(\mu, \nu) = \inf\left\{\lambda > 0 : \int_0^1 \phi\left(\frac{|F_\mu^{-1}(u) - F_\nu^{-1}(u)|}{\lambda}\right) \mathrm{d}u \le 1\right\}. \tag{7}$$

Building on the one-dimensional quantile representation above, we introduce the SOW distance.

Definition 1. *For $p \ge 1$ and $\mu, \nu \in \mathcal{P}_\phi(\mathbb{R}^d)$, we define the* SOW *distance by*

$$\mathrm{SOW}^p_{\phi,p}(\mu, \nu) \coloneqq \int_{\mathbb{S}^{d-1}} \mathrm{W}^p_\phi\left(\Pi^\theta_\# \mu, \Pi^\theta_\# \nu\right) \mathrm{d}\sigma(\theta). \tag{8}$$

Since $\Pi^\theta_\# \mu$ and $\Pi^\theta_\# \nu$ belong to $\mathcal{P}_\phi(\mathbb{R})$ for every $\theta \in \mathbb{S}^{d-1}$, equation 7 applies directly to each projected pair. The proposed distance recovers the classical SW_p as a special case. Indeed, when $\phi(t) = t^p$ with $p \ge 1$, we have $\mathrm{W}_\phi = \mathrm{W}_p$, and therefore $\mathrm{SOW}_{\phi,p} = \mathrm{SW}_p$.

3.2 Topological Properties

We first establish the metric and comparison properties of the SOW distance.

Theorem 1. *Let ϕ be an Orlicz function and let $p \ge 1$. Then $\mathrm{SOW}_{\phi,p}$ defines a metric on $\mathcal{P}_\phi(\mathbb{R}^d)$.*

Proposition 1. *Let ϕ be an Orlicz function and let $p \ge 1$. Then*

$$\mathrm{SOW}_{\phi,p}(\mu, \nu) \le \mathrm{W}_\phi(\mu, \nu) \quad \textit{for all } \mu, \nu \in \mathcal{P}_\phi(\mathbb{R}^d)$$

Proposition 2. *Let ϕ_1 and ϕ_2 be Orlicz functions satisfying $\phi_1(t) \le \phi_2(t)$ for every $t \ge 0$. Then, for every $p \ge 1$ and every $\mu, \nu \in \mathcal{P}_{\phi_2}(\mathbb{R}^d)$, $\mathrm{SOW}_{\phi_1,p}(\mu, \nu) \le \mathrm{SOW}_{\phi_2,p}(\mu, \nu)$.*

The proofs build on the arguments in Guha et al. (2023, Lemmas 3.3–3.4) and are given in Appendix B. We next study the convergence induced by $\mathrm{SOW}_{\phi,p}$. We first show that convergence in $\mathrm{SOW}_{\phi,p}$ is equivalent to uniform convergence of the projected OW distances.

Proposition 3. *Let $p \in [1, \infty)$ and let ϕ be an Orlicz function. Let $(\mu_n)_{n\ge1} \subset \mathcal{P}_\phi(\mathbb{R}^d)$, and let $\mu \in \mathcal{P}_\phi(\mathbb{R}^d)$. Then* $\mathrm{SOW}_{\phi,p}(\mu_n, \mu) \to 0 \iff \sup_{\theta\in\mathbb{S}^{d-1}} \mathrm{W}_\phi\left(\Pi^\theta_\# \mu_n, \Pi^\theta_\# \mu\right) \to 0.$

The key point is that the projected OW distances vary continuously with the projection direction (see Lemma 5). Moreover, SOW convergence yields a uniform bound on the Orlicz moments defined in equation 15. Together, these properties prevent the projected distances from remaining large on arbitrarily small sets of directions. The complete proof is given in Appendix C.1. As a direct consequence, convergence in $\mathrm{SOW}_{\phi,p}$ implies weak convergence.

Corollary 1. *Under the assumptions of Proposition 3,* $\mathrm{SOW}_{\phi,p}(\mu_n, \mu) \to 0 \quad \Longrightarrow \quad \mu_n \rightharpoonup \mu$.

Remark 1. *The converse implication does not hold in general, even in dimension one. Indeed, one can construct a sequence that converges weakly while its* OW *distance from the limit remains bounded away from zero. A concrete counterexample is given in Appendix C.2. However, the converse becomes valid under a common compact-support assumption.*

Proposition 4. *Let* $p \in [1, \infty)$*, let* ϕ *be an Orlicz function, and let* $K \subset \mathbb{R}^d$ *be compact. Suppose that* $(\mu_n)_{n\geq 1}$ *and* μ *belong to* $\mathcal{P}(K)$*. Then* $\mu_n \rightharpoonup \mu \quad \Longleftrightarrow \quad \mathrm{SOW}_{\phi,p}(\mu_n, \mu) \to 0$.

The proof of this property is deferred to Appendix C.1.

3.3 Statistical Properties

We now study plug-in estimation of $\mathrm{SOW}_{\phi,p}$. We first establish a one-sample empirical convergence bound. We then study two related two-sample estimation problems: estimation of the distance $\mathrm{SOW}_{\phi,p}$ and estimation of the powered functional $\mathrm{SOW}^p_{\phi,p}$. Hereafter, ϕ denotes a fixed Orlicz function introduced in Section 2. For the statistical and minimax results, we also fix a compact set $K \subset \mathbb{R}^d$ with $D \coloneqq \mathrm{diam}(K) > 0$. For $\mu, \nu \in \mathcal{P}(K)$, let $\widehat{\mu}_n$ and $\widehat{\nu}_m$ denote the empirical measures based on independent samples of size n and m from μ and ν, respectively.

Theorem 2. *Let* $p \in [1, \infty)$ *and* $\mu \in \mathcal{P}(K)$*. Then*

$$\left[\mathbb{E}\,\mathrm{SOW}^p_{\phi,p}\left(\widehat{\mu}_n, \mu\right)\right]^{1/p} \lesssim_p D\,\omega_\phi\left(n^{-1/2}\right).$$

The rate is governed by the fundamental function ω_ϕ defined in equation 4 and has no explicit dependence on the ambient dimension d. The complete proof is deferred to Appendix D.

We next consider the two-sample setting. The metric property of $\mathrm{SOW}_{\phi,p}$ and Theorem 2 yield the following bound. The complete proof of Corollary 2 is given in Appendix D.

Corollary 2. *Let* $p \in [1, \infty)$ *and let* $\mu, \nu \in \mathcal{P}(K)$*. Then*

$$\mathbb{E}\left|\mathrm{SOW}_{\phi,p}\left(\widehat{\mu}_n, \widehat{\nu}_m\right) - \mathrm{SOW}_{\phi,p}(\mu, \nu)\right| \lesssim_p D\left[\omega_\phi\left(n^{-1/2}\right) + \omega_\phi\left(m^{-1/2}\right)\right].$$

We next turn to the powered functional $\mathrm{SOW}^p_{\phi,p}$. Unlike the preceding distance bound, its analysis requires controlling the variation of the powered Luxemburg norm under L^1 perturbations. To this end, we define

$$\psi_{\phi,p}(u) \coloneqq \omega_\phi(u)^p = \frac{1}{\phi^{-1}(1/u)^p}, \qquad u \in (0, 1], \tag{9}$$

with $\psi_{\phi,p}(0) \coloneqq 0$. A direct calculation gives

$$\psi_{\phi,p}(u) = \mathrm{W}^p_\phi\left(\delta_0, (1-u)\delta_0 + u\delta_1\right), \quad u \in [0, 1].$$

Thus, $\psi_{\phi,p}(u)$ measures the powered OW response to transporting a mass fraction u over a unit distance. Let $\psi^{\mathrm{cav}}_{\phi,p}$ denote the least concave majorant of $\psi_{\phi,p}$ on $[0, 1]$, i.e., the smallest concave function on $[0, 1]$ that dominates $\psi_{\phi,p}$. For the powered-functional estimation problem, we need to control how the powered Luxemburg norm changes when a quantile-gap profile is perturbed in L^1. To capture the worst-case variation, define

$$\Omega_{\phi,p}(u) \coloneqq \sup_{\substack{0\leq f,g\leq 1 \\ \|f-g\|_{L^1(0,1)}\leq u}} \left|\|f\|^p_\phi - \|g\|^p_\phi\right|, \quad u \in [0, 1].$$

Proposition 6 shows that $\Omega_{\phi,p}(u) \asymp_p \psi^{\mathrm{cav}}_{\phi,p}(u)$. Thus, $\psi^{\mathrm{cav}}_{\phi,p}$ gives the sharp perturbation scale underlying the following two-sample bound. The proof is provided in Appendix E.

Theorem 3. *Let* $p \in [1, \infty)$ *and let* $\mu, \nu \in \mathcal{P}(K)$*. Then*

$$\mathbb{E}\left|\mathrm{SOW}^p_{\phi,p}\left(\widehat{\mu}_n, \widehat{\nu}_m\right) - \mathrm{SOW}^p_{\phi,p}(\mu, \nu)\right| \leq p 2^p D^p \psi^{\mathrm{cav}}_{\phi,p}\left(\frac{1}{2}n^{-1/2} + \frac{1}{2}m^{-1/2}\right).$$

Remark 2 (Recovery of SW rates). *When $\phi(t) = t^p$ with $p \geq 1$, a direct computation yields*

$$\omega_\phi(u) = u^{1/p}, \qquad \psi_{\phi,p}(u) = \psi^{\mathrm{cav}}_{\phi,p}(u) = u, \qquad \mathrm{SOW}_{\phi,p} = \mathrm{SW}_p\,.$$

Assuming $n \leq m$, Theorem 2 and Corollary 2 give the rate $n^{-1/(2p)}$ for one-sample empirical convergence and two-sample estimation of SW_p, respectively. These rates are consistent with those established for sliced Wasserstein distances in Nadjahi et al. (2020). Moreover, Theorem 3 gives the rate $n^{-1/2}$ for the powered functional SW_p^p, matching the parametric rate obtained under compact-support assumptions in Nietert et al. (2022).

3.4 Minimax Optimality

The preceding results provide upper bounds for the empirical plug-in estimators but do not establish their statistical optimality. We therefore study the corresponding minimax risks over $\mathcal{P}(K)$, taking the infimum over all estimators based on the observed samples, following the standard minimax framework of (Tsybakov, 2009, Chapter 2). Matching lower bounds then show that the empirical plug-in estimators are minimax-rate optimal up to constant factors.

Theorem 4 (Minimax optimality). *Let $p \in [1, \infty)$. For one-sample distribution estimation,*

$$\inf_{\widetilde{\mu}_n} \sup_{\mu \in \mathcal{P}(K)} \left[\mathbb{E}\,\mathrm{SOW}^p_{\phi,p}\left(\widetilde{\mu}_n, \mu\right)\right]^{1/p} \asymp_{d,p} D\omega_\phi\left(n^{-1/2}\right).$$

For the two-sample problems, we have

$$\inf_{\widehat{T}} \sup_{\mu,\nu \in \mathcal{P}(K)} \mathbb{E}\left|\widehat{T} - \mathrm{SOW}_{\phi,p}(\mu,\nu)\right| \asymp_{d,p} D\left[\omega_\phi\left(n^{-1/2}\right) + \omega_\phi\left(m^{-1/2}\right)\right],$$

and

$$\inf_{\widehat{T}} \sup_{\mu,\nu \in \mathcal{P}(K)} \mathbb{E}\left|\widehat{T} - \mathrm{SOW}^p_{\phi,p}(\mu,\nu)\right| \asymp_{d,p} D^p \psi^{\mathrm{cav}}_{\phi,p}\left(\frac{1}{2}n^{-1/2} + \frac{1}{2}m^{-1/2}\right).$$

Here, the first infimum is taken over all estimators based on n observations, while the last two are taken over all estimators based on the two independent samples of sizes n and m.

The proof of Theorem 4, based on Le Cam's two-point method (LeCam, 1973; Le Cam, 2012), is given in Appendix F. Consequently, the empirical plug-in estimators studied in the preceding subsection are minimax-rate optimal over $\mathcal{P}(K)$. For fixed d, p, ϕ, and K, the sample-size dependence of these rates is tight up to constant factors.

3.5 Computational Properties

In this subsection, we focus on the SOW between the empirical measures $\widehat{\mu}_n$ and $\widehat{\nu}_m$. The observed samples are held fixed throughout. We approximate the spherical integral by Monte Carlo sampling and compute the projected OW distances by bisection when no closed-form expression is available[2].

Monte Carlo approximation. For $\theta_1, \ldots, \theta_L \overset{\text{i.i.d.}}{\sim} \sigma$, we first approximate $\mathrm{SOW}^p_{\phi,p}\left(\widehat{\mu}_n, \widehat{\nu}_m\right)$ by

$$\widehat{\mathrm{SOW}}^p_{\phi,p}(\widehat{\mu}_n, \widehat{\nu}_m, L) \coloneqq \frac{1}{L}\sum_{l=1}^{L} \mathrm{W}^p_\phi\left(\Pi^{\theta_l}_{\#}\widehat{\mu}_n, \Pi^{\theta_l}_{\#}\widehat{\nu}_m\right). \tag{10}$$

Bisection for the one-dimensional distances. Fix $\theta \in \mathbb{S}^{d-1}$ and write

$$\Delta_\theta(u) \coloneqq F^{-1}_{\Pi^\theta_{\#}\widehat{\mu}_n}(u) - F^{-1}_{\Pi^\theta_{\#}\widehat{\nu}_m}(u), \qquad \Psi_\theta(\lambda) \coloneqq \int_0^1 \phi\left(\frac{|\Delta_\theta(u)|}{\lambda}\right)\mathrm{d}u, \quad \lambda > 0. \tag{11}$$

By the quantile representation equation 7, $\mathrm{W}_\phi(\Pi^\theta_{\#}\widehat{\mu}_n, \Pi^\theta_{\#}\widehat{\nu}_m) = \inf\{\lambda > 0 : \Psi_\theta(\lambda) \leq 1\}$. Since ϕ is increasing, Ψ_θ is decreasing in λ, so the sublevel set $\{\Psi_\theta \leq 1\}$ is a half-line and the value of

[2] See Appendix G.1.

W_ϕ can be located by bisection. By Jensen's inequality and the monotonicity of ϕ, we obtain the following initial bracket:

$$L_\theta \coloneqq \omega_\phi(1)\|\Delta_\theta\|_{L^1(0,1)} \le \mathrm{W}_\phi\left(\Pi^\theta_\# \widehat{\mu}_n, \Pi^\theta_\# \widehat{\nu}_m\right) \le \omega_\phi(1) \max_{u\in(0,1)} |\Delta_\theta(u)| \coloneqq U_\theta. \tag{12}$$

Let $\widehat{\mathrm{W}}_\phi(\Pi^\theta_\# \widehat{\mu}_n, \Pi^\theta_\# \widehat{\nu}_m, T)$ denote the upper endpoint of the bracket produced by T bisection steps started from $[L_\theta, U_\theta]$ (see Algorithm 1). The resulting estimator of $\mathrm{SOW}^p_{\phi,p}$ is

$$\widehat{\mathrm{SOW}}^p_{\phi,p}(\widehat{\mu}_n, \widehat{\nu}_m, L, T) \coloneqq \frac{1}{L}\sum_{l=1}^{L} \widehat{\mathrm{W}}^p_\phi\left(\Pi^{\theta_l}_\# \widehat{\mu}_n, \Pi^{\theta_l}_\# \widehat{\nu}_m, T\right). \tag{13}$$

The following proposition bounds the approximation error of Algorithm 1. Its proof is given in Appendix G.2.

Proposition 5. *The approximation returned by Algorithm 1 satisfies*

$$\begin{aligned}&\mathbb{E}_{\theta_1,\dots,\theta_L}\left|\widehat{\mathrm{SOW}}^p_{\phi,p}(\widehat{\mu}_n, \widehat{\nu}_m, L, T) - \mathrm{SOW}^p_{\phi,p}(\widehat{\mu}_n, \widehat{\nu}_m)\right| \\ &\le \frac{p}{2^T}\int_{\mathbb{S}^{d-1}} U^p_\theta \,\mathrm{d}\sigma(\theta) + \frac{1}{\sqrt{L}}\,\mathrm{Var}_{\theta\sim\sigma}\left[\mathrm{W}^p_\phi\left(\Pi^\theta_\# \widehat{\mu}_n, \Pi^\theta_\# \widehat{\nu}_m\right)\right]^{1/2}.\end{aligned}$$

Computational complexity. For empirical measures with n atoms, Δ_θ is a step function with n pieces, so each evaluation of Ψ_θ costs $\mathcal{O}(n)$ operations. Algorithm 1 therefore runs in $\mathcal{O}\left(L\left(dn + n\log n + Tn\right)\right)$ time and $\mathcal{O}(L(d+n))$ memory, the three terms accounting for projection, sorting, and bisection respectively.

4 Experiments

4.1 Synthetic Studies

We create data from $\mu = \mathcal{N}(0, I_d)$ and $\nu = 0.95\mathcal{N}(0, I_d) + 0.025\mathcal{N}(2.5\,e_1, I_d) + 0.025\mathcal{N}(-2.5\,e_1, I_d)$, with $d = 2$ and $e_1 = (1, 0)$. We report the computational speed of SOW ($L = 100$, $p = 2$) and OW using entropic regularization (entropic coefficient 5, up to 150 Sinkhorn iterations with 10^{-7} marginal error) (Guha et al., 2023) in Figure 1. Here, we consider $\phi_1(x) = \frac{x^2+x^4}{2}$, and $\phi_2(x) = \frac{e^{x^2}-1}{e-1}$ as two Orlicz functions. ϕ_1 additionally admits an exact closed-form per-projection solve for SOW (no bisection at all). SOW with ϕ_2 and OW (for both ϕ_1 and ϕ_2) are solved by bisection. The results are averaged over 5 runs. The left figure shows that SOW is much more scalable than OW. Here, we choose the number of bisection iterations T for OW and for SOW's ϕ_2 such that we have a guaranteed error of 10^{-9} or reaching 100 iterations (the tolerance is often reached after $\sim$ 33–34 iterations). In the figure, the computational speed of SOW with $n = 10^4$ is even lower than OW with $n = 100$. Moreover, the scaling rate is much lower for SOW. The middle figure confirms that the scaling in the number of projections L for SOW is linear. The right figure presents the computational time of OW and SOW in terms of the number of bisection iterations T, forced directly up to $T = 100$ for SOW's ϕ_2 and OW's bisection (rather than derived from the tolerance above). We also observe that SOW's ϕ_2 has a better scaling rate than OW in terms of T, due to a faster computational time per iteration. SOW's ϕ_1 curve is flat in T by construction, since its closed-form solve does not depend on T at all.

We further investigate the empirical sample-size, projection, and bisection errors in Figure 2. For the left panel, we first generate and fix two empirical reference distributions, each consisting of 10^5 observations. For each replication, we draw an i.i.d. sequence of $50,000$ observations with replacement from each fixed empirical reference distribution and use its prefixes to form nested empirical samples with sizes n ranging from 100 to $50,000$. We fix $L = 100$ and use the same projection directions to evaluate each sample pair and the corresponding reference pair. The reported errors are averaged over 100 replications. For the middle panel, we fix $n = 1000$ and vary L from 5 to 5000, using the value at $L = 10^4$ on the same data and with the same projection sequence as the numerical reference. For the right panel, we fix $n = 1000$ and $L = 100$ and vary T from 1 to 100 for SOW with ϕ_2, with the reference computed using a tolerance of 10^{-15}. The left panel shows

Algorithm 1 Computation of $\mathrm{SOW}_{\phi,p}^{p}\left(\widehat{\mu}_n, \widehat{\nu}_m\right)$

Input: $\widehat{\mu}_n, \widehat{\nu}_m$, Orlicz function ϕ, $p \geq 1$, number of projections L, number of bisection steps T.
Sample $\theta_1, \ldots, \theta_L \overset{\text{i.i.d.}}{\sim} \sigma$.
Project and sort to obtain the quantile gaps $\Delta_{\theta_l}, l = 1, \ldots, L$.
Initialize $\ell_l \leftarrow L_{\theta_l}$ and $u_l \leftarrow U_{\theta_l}$ for $l = 1, \ldots, L$ by using equation 12.
Keep u_l unchanged for directions with $\ell_l = u_l$.
Perform the following steps only for directions with $\ell_l < u_l$ at initialization.
for $t = 1$ to T **do**
  $v_l \leftarrow (\ell_l + u_l)/2$.
  Evaluate $\Psi_{\theta_l}(v_l)$ as in equation 11.
  $u_l \leftarrow v_l$ if $\Psi_{\theta_l}(v_l) \leq 1$; $\ell_l \leftarrow v_l$ otherwise.
end for
Return: $\frac{1}{L}\sum_{l=1}^{L} u_l^p$.

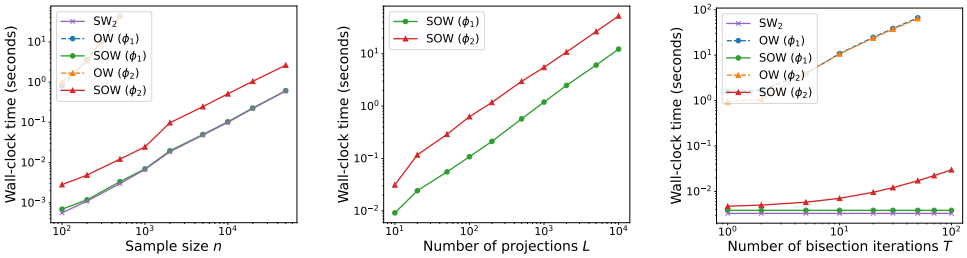


Figure 1: Computational speed of SOW and OW.

decreasing empirical errors for all three distances, with SOW using ϕ_2 exhibiting the slowest decay, consistent with the dependence on the Orlicz function through ω_ϕ in Section 3.3. The middle panel shows decreasing Monte Carlo projection error as L increases, while the right panel shows that the bisection error decreases rapidly and becomes very small after about 50 iterations.

4.2 Two-sample Testing

We evaluate SOW for two-sample testing using $\mathrm{SOW}_{\phi,2}\left(\widehat{\mu}_n, \widehat{\nu}_n\right)$ as the test statistic, with ϕ_1 and ϕ_2 from Section 4.1. We compare against SW_2 (Tran & Schreuder, 2026) and MMD with RBF and Laplace kernels (Gretton et al., 2012; Schrab, 2025), using the median heuristic for bandwidth selection. For SW and SOW, we use $L = n/2$ shared projection directions within each trial. Empirical power is estimated over 200 independent trials using permutation test with 199 permutations at level $\alpha = 0.05$, thereby controlling the finite-sample Type I error under the null exchangeability assumption (Hemerik & Goeman, 2018). In Figure 3, we consider three settings. For the synthetic experiments, $\mu = \mathcal{N}(0, I_{15})$ and $\nu = 0.95\mu + 0.025\mathcal{N}(2.5e_1, I_{15}) + 0.025\mathcal{N}(-2.5e_1, I_{15})$. For MNIST (LeCun et al., 1998), following Wang et al. (2022), we preprocess the dataset by applying a sigmoid transformation to each image so that all pixel values lie within $[0, 1]$. We then compare the processed digit-6 distribution μ_6 with $0.9\mu_6 + 0.1\mu_9$. For CIFAR, following Biggs et al. (2023), we compare the uniform distributions over CIFAR-10 (Krizhevsky et al., 2009) and CIFAR-10.1 (Recht et al., 2019). Figure 3 shows that suitable choices of the Orlicz function can yield higher empirical test power than SW_2 at the same sample size.

4.3 Evaluating Generative Models

Evaluating a generative model is to measure a statistical distance between the model distribution and data distribution. The standard FID (Heusel et al., 2017) uses Wasserstein-2 between fitted Gaussians on Inception features and compares only their first two moments, which requires estimating a $d \times d$ covariance (hence 50k samples for Inception features). Authors in Berthet et al. (2026) propose to use SW directly on the set of samples and demonstrate that SW can avoid the weaknesses of FID which is failing to separate distributions that match in mean and covariance but differ elsewhere. Since SOW is a family indexed by ϕ, evaluating a model at several Orlicz functions of increasing growth yields a profile rather than a single score. Fast-growing ϕ inflates the contribution of the quantiles at which the two projected measures are far apart, i.e., the tails and the rare or missing

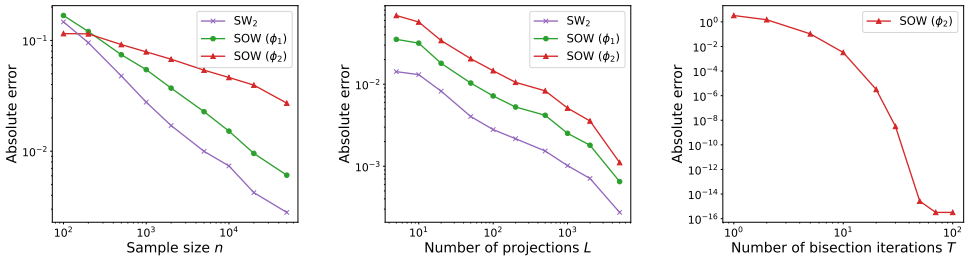


Figure 2: Sample and Monte Carlo projection errors of SW and SOW, and bisection error of SOW.

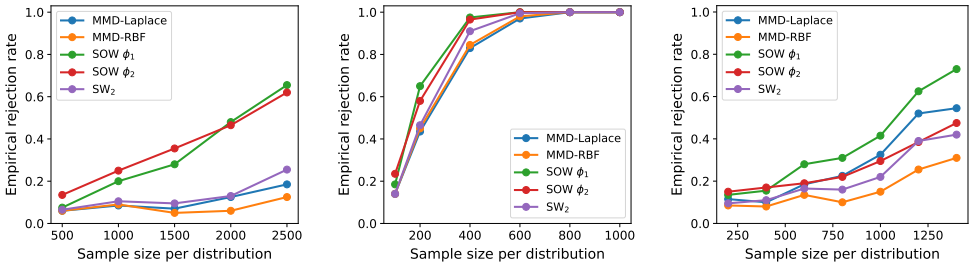


Figure 3: Power versus sample-size: synthetic data, MNIST and CIFAR-10 versus CIFAR-10.1.

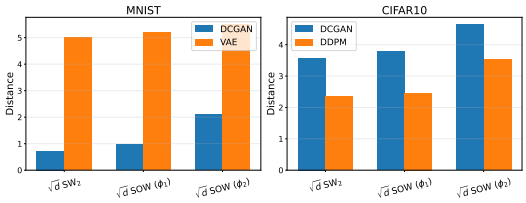


Figure 4: Evaluating generative models with SW and SOW.

modes. How fast the score grows along the profile therefore says where the mismatch sits, which a single SW value cannot reveal.

In Figure 4 we compare a DCGAN (Radford et al., 2015) against a VAE (Kingma & Welling, 2014) on MNIST and against a DDPM (Ho et al., 2020) on CIFAR10, reporting $\sqrt{d}\,\mathrm{SW}_2$, $\sqrt{d}\,\mathrm{SOW}(\phi_1)$ and $\sqrt{d}\,\mathrm{SOW}(\phi_2)$ with $\phi_1(x) = \frac{x^2+x^4}{2}$ and $\phi_2(x) = \frac{e^{x^2}-1}{e-1}$. On MNIST the DCGAN score rises considerably while VAE score increases slightly using $10,000$ samples and $L = 10,000$ projections. This demonstrates that VAE's error is a large but uniformly spread blur that every cost charges equally, whereas the DCGAN concentrates its error on a small fraction of mass transported far, exactly what ϕ_1 and ϕ_2 penalize. We can then conclude that DCGAN is better mainly where most of the mass lies and much less so once far transport is charged heavily. On CIFAR-10, the gaps between DCGAN and DDPM stay approximately constant. This says both models miss the data in the tails by a similar amount, and the DDPM is simply better by the same margin everywhere.

5 Conclusion

We introduced SOW, a generalization of SW based on Orlicz geometry. We established its metric, topological, statistical, and computational properties, including minimax-optimal sample complexities under compact support and developed an approximation algorithm. Experiments show that SOW scales substantially better than OW and provide initial evidence that suitable choices of ϕ can increase sensitivity to discrepancies involving small amounts of mass transported over large distances. Our goal here is to demonstrate this flexibility rather than optimize ϕ for individual tasks. Principled and data-adaptive selection of ϕ is an important direction for future work. Another direction is to use SOW directly as a risk for model estimation, which requires differentiating through the scalar Luxemburg-norm solve and controlling its numerical error. Further extensions include non-uniform slicing distributions (Nguyen et al., 2021; Nguyen & Ho, 2023) and non-linear projections (Kolouri et al., 2019; Bonet et al., 2025). Extending the statistical guarantees beyond compact support is also an important direction (see Appendix H for our first step).

AI USE STATEMENT

In this work, we used generative AI tools for editing the paper, for coding, and for brainstorming proof strategies. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We provide the code reproduce the results in the paper.

REFERENCES

Jason M Altschuler and Sinho Chewi. Faster high-accuracy log-concave sampling via algorithmic warm starts. *Journal of the ACM*, 71(3):1–55, 2024.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International Conference on Machine Learning*, pp. 214–223, 2017.

Robert G Bartle and Donald R Sherbert. *Introduction to real analysis*, volume 2. Wiley New York, 2000.

Colin Bennett, Robert C Sharpley, et al. *Interpolation of operators*, volume 129. Academic press Boston, 1988.

Espen Bernton, Pierre E Jacob, Mathieu Gerber, and Christian P Robert. Approximate Bayesian computation with the Wasserstein distance. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 81(2):235–269, 2019a.

Espen Bernton, Pierre E Jacob, Mathieu Gerber, and Christian P Robert. On parameter estimation with the Wasserstein distance. *Information and Inference: A Journal of the IMA*, 8(4):657–676, 2019b.

Quentin Berthet, Yu-Han Wu, Clement Crepy, Romuald Elie, Klaus Greff, and Michael Eli Sander. MIND: Monge inception distance for generative models evaluation. *arXiv preprint arXiv:2605.06797*, 2026.

Felix Biggs, Antonin Schrab, and Arthur Gretton. Mmd-fuse: Learning and combining kernels for two-sample testing without data splitting. *Advances in Neural Information Processing Systems*, 36:75151–75188, 2023.

Sergey Bobkov and Michel Ledoux. *One-dimensional empirical measures, order statistics, and Kantorovich transport distances*, volume 261. American Mathematical Society, 2019.

Clément Bonet, Lucas Drumetz, and Nicolas Courty. Sliced-Wasserstein distances and flows on Cartan-Hadamard manifolds. *Journal of Machine Learning Research*, 26(32):1–76, 2025.

Nicolas Bonneel, Julien Rabin, Gabriel Peyré, and Hanspeter Pfister. Sliced and Radon Wasserstein barycenters of measures. *Journal of Mathematical Imaging and Vision*, 1(51):22–45, 2015.

Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, Cambridge, UK, 2004. ISBN 978-0-521-83378-3.

Richard L Burden. *Numerical analysis*. Brooks/Cole Cengage Learning, 2011.

Stamatis Cambanis, Gordon Simons, and William Stout. Inequalities for e k (x, y) when the marginals are fixed. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 36(4): 285–294, 1976.

Marta Catalano, Antonio Lijoi, and Igor Prünster. Measuring dependence in the Wasserstein distance for Bayesian nonparametric models. *The Annals of Statistics*, 49(5):2916–2947, 2021.

Marta Catalano, Hugo Lavenant, Antonio Lijoi, and Igor Prünster. A Wasserstein index of dependence for random measures. *Journal of the American Statistical Association*, 119(547):2396–2406, 2024.

Nicolas Courty, Rémi Flamary, Amaury Habrard, and Alain Rakotomamonjy. Joint distribution optimal transportation for domain adaptation. In *Advances in Neural Information Processing Systems*, pp. 3730–3739, 2017.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In C.J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger (eds.), *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013. URL https://proceedings.neurips.cc/paper_files/paper/2013/file/af21d0c97db2e27e13572cbf59eb343d-Paper.pdf.

Ishan Deshpande, Ziyu Zhang, and Alexander G Schwing. Generative modeling using the sliced Wasserstein distance. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3483–3491, 2018.

Rick Durrett. *Probability: theory and examples*, volume 49. Cambridge university press, 2019.

Jean Feydy, Benjamin Charlier, François-Xavier Vialard, and Gabriel Peyré. Optimal transport for diffeomorphic registration. In *Medical Image Computing and Computer Assisted Intervention-MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, September 11-13, 2017, Proceedings, Part I 20*, pp. 291–299. Springer, 2017.

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

Martin Franců, Ron Kerman, and Gord Sinnamon. A new algorithm for approximating the least concave majorant. *Czechoslovak Mathematical Journal*, 67:1071–1093, 2017. doi: 10.21136/CMJ.2017.0408-16.

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012.

Aritra Guha, Nhat Ho, and XuanLong Nguyen. On excess mass behavior in Gaussian mixture models with Orlicz-Wasserstein distances. In *International Conference on Machine Learning*, pp. 11847–11870. PMLR, 2023.

Ping He, Om Khangaonkar, Hamed Pirsiavash, Yikun Bai, and Soheil Kolouri. Sinkhorn-drifting generative models. *arXiv preprint arXiv:2603.12366*, 2026.

Jesse Hemerik and Jelle Goeman. Exact testing with random permutations. *Test*, 27(4):811–825, 2018.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *Advances in Neural Information Processing Systems*, pp. 6626–6637, 2017.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

Olav Kallenberg. *Foundations of modern probability*, volume 2. Springer, 1997.

Martin Kell. On interpolation and curvature via Wasserstein geodesics. *Advances in Calculus of Variations*, 10(2):125–167, 2017.

Diederik P Kingma and Max Welling. Auto-encoding variational Bayes. *International Conference on Learning Representations*, 2014.

Soheil Kolouri, Se Rim Park, Matthew Thorpe, Dejan Slepcev, and Gustavo K Rohde. Optimal mass transport: Signal processing and machine-learning applications. *IEEE signal processing magazine*, 34(4):43–59, 2017.

Soheil Kolouri, Phillip E Pope, Charles E Martin, and Gustavo K Rohde. Sliced Wasserstein auto-encoders. In *International Conference on Learning Representations*, 2018a.

Soheil Kolouri, Gustavo K Rohde, and Heiko Hoffmann. Sliced Wasserstein distance for learning Gaussian mixture models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3427–3436, 2018b.

Soheil Kolouri, Kimia Nadjahi, Umut Simsekli, Roland Badeau, and Gustavo Rohde. Generalized sliced Wasserstein distances. In *Advances in Neural Information Processing Systems*, pp. 261–272, 2019.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. *Master's thesis, Department of Computer Science, University of Toronto*, 2009.

Lucien Le Cam. *Asymptotic methods in statistical decision theory*. Springer Science & Business Media, 2012.

Lucien LeCam. Convergence of estimates under dimensionality restrictions. *The Annals of Statistics*, pp. 38–53, 1973.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=PqvMRDCJT9t.

Wilhelmus Anthonius Josephus Luxemburg et al. *Banach function spaces*. Number 6. Van Gorcum, 1955.

Pascal Massart. The tight constant in the dvoretzky-kiefer-wolfowitz inequality. *The annals of Probability*, pp. 1269–1283, 1990.

Kimia Nadjahi, Alain Durmus, Lénaïc Chizat, Soheil Kolouri, Shahin Shahrampour, and Umut Simsekli. Statistical and topological properties of sliced probability divergences. *Advances in Neural Information Processing Systems*, 33:20802–20812, 2020.

Khai Nguyen. An introduction to sliced optimal transport: foundations, advances, extensions, and applications. *Foundations and Trends® in Computer Graphics and Vision*, 17(3-4):171–391, 2025.

Khai Nguyen and Nhat Ho. Energy-based sliced Wasserstein distance. *Advances in Neural Information Processing Systems*, 2023.

Khai Nguyen, Nhat Ho, Tung Pham, and Hung Bui. Distributional sliced-Wasserstein and applications to generative modeling. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=QYjO70ACDK.

Sloan Nietert, Ritwik Sadhu, Ziv Goldfeld, and Kengo Kato. Statistical, robustness, and computational guarantees for sliced Wasserstein distances. *Advances in Neural Information Processing Systems*, 2022.

Gabriel Peyré. Optimal transport for machine learners. *arXiv preprint arXiv:2505.06589*, 2025.

Gabriel Peyré, Marco Cuturi, et al. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

Julien Rabin, Gabriel Peyré, Julie Delon, and Marc Bernot. Wasserstein barycenter and its application to texture mixing. In *Scale Space and Variational Methods in Computer Vision: Third International Conference, SSVM 2011, Ein-Gedi, Israel, May 29–June 2, 2011, Revised Selected Papers 3*, pp. 435–446. Springer, One New York Plaza, Suite 4600, New York, NY 10004-1562, 2012.

Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

Malempati Madhusudana Rao and Zhong Dao Ren. *Theory of Orlicz spaces*. 1991.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do ImageNet classifiers generalize to ImageNet? In *International Conference on Machine Learning*, pp. 5389–5400. PMLR, 2019.

Ralph Tyrrell Rockafellar. *Convex Analysis*. Princeton University Press, 1970. ISBN 9780691015866.

Ben-Zion A Rubshtein, Genady Ya Grabarnik, Mustafa A Muratov, and Yulia S Pashkova. *Foundations of symmetric spaces of measurable functions*, volume 45. Springer, 2016.

Geoffrey Schiebinger, Jian Shu, Marcin Tabaka, Brian Cleary, Vidya Subramanian, Aryeh Solomon, Joshua Gould, Siyan Liu, Stacie Lin, Peter Berube, et al. Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming. *Cell*, 176(4):928–943, 2019.

Antonin Schrab. A practical introduction to kernel discrepancies: MMD, HSIC & KSD. *arXiv preprint arXiv:2503.04820*, 2025.

Justin Solomon, Fernando De Goes, Gabriel Peyré, Marco Cuturi, Adrian Butscher, Andy Nguyen, Tao Du, and Leonidas Guibas. Convolutional Wasserstein distances: Efficient optimal transportation on geometric domains. *ACM Transactions on Graphics (ToG)*, 34(4):1–11, 2015.

Justin Solomon, Gabriel Peyré, Vladimir G Kim, and Suvrit Sra. Entropic metric alignment for correspondence problems. *ACM Transactions on Graphics (TOG)*, 35(4):72, 2016.

Karl-Theodor Sturm. Generalized Orlicz spaces and Wasserstein distances for convex–concave scale functions. *Bulletin des sciences mathématiques*, 135(6-7):795–802, 2011.

Binh Thuan Tran and Nicolas Schreuder. Minimax-optimal two-sample test with sliced Wasserstein. In Emtiyaz Khan, Yingzhen Li, Arno Solin, and Aaditya Ramdas (eds.), *Proceedings of The 29th International Conference on Artificial Intelligence and Statistics*, volume 300 of *Proceedings of Machine Learning Research*, pp. 1927–1935. PMLR, 02–05 May 2026. URL `https://proceedings.mlr.press/v300/tran26c.html`.

Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Series in Statistics. Springer, New York, 2009. doi: 10.1007/b13794.

Cédric Villani. *Topics in optimal transportation*. Number 58. American Mathematical Soc., 2003.

Cédric Villani. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.

Jie Wang, Rui Gao, and Yao Xie. Two-sample test with kernel projected Wasserstein distance. In Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera (eds.), *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pp. 8022–8055. PMLR, 28–30 Mar 2022. URL `https://proceedings.mlr.press/v151/wang22f.html`.

A Basic Properties of the Orlicz Quantities

The purpose of this appendix is to collect the elementary properties that will be used later in the proofs. For further background on Orlicz functions and spaces, we refer the reader to Rao & Ren (1991) and Rubshtein et al. (2016, Section 13.1).

A.1 Properties of ϕ and ϕ^{-1}

Lemma 1 (Basic properties of the Orlicz function)**.** *Suppose that ϕ is the Orlicz function introduced in Section 2. Then ϕ is continuous on $[0, +\infty)$ and $\lim_{t\to+\infty} \phi(t) = +\infty$. Consequently, ϕ is a bijection from $[0, +\infty)$ onto $[0, +\infty)$. Its inverse $\phi^{-1} : [0, +\infty) \to [0, +\infty)$ is continuous and strictly increasing, with $\phi^{-1}(0) = 0$ and $\lim_{s\to+\infty} \phi^{-1}(s) = +\infty$.*

Proof. Since ϕ is finite-valued and convex, it is continuous on $(0, +\infty)$; see, for example, Rockafellar (1970, Theorem 10.1). It remains to verify right-continuity at zero. Fix $a > 0$. For every $t \in [0, a]$, convexity and $\phi(0) = 0$ give

$$0 \leq \phi(t) = \phi\left[\frac{t}{a}a + \left(1 - \frac{t}{a}\right)0\right] \leq \frac{t}{a}\phi(a).$$

Therefore, $\phi(t) \to 0 = \phi(0)$ as $t \to 0^+$. Hence, ϕ is continuous on $[0, +\infty)$.

Next, since ϕ is strictly increasing on $[0, +\infty)$ and $\phi(0) = 0$, $\phi(1) > 0$. For every $t \geq 1$, the convexity of ϕ gives

$$\phi(1) = \phi\left[\frac{1}{t}t + \left(1 - \frac{1}{t}\right)0\right] \leq \frac{\phi(t)}{t}.$$

Then, $\phi(t) \geq t\phi(1)$. Therefore, $\phi(t) \geq t\phi(1) \to +\infty$ as $t \to +\infty$. Together with $\phi(0) = 0$ and the continuity of ϕ on $[0, +\infty)$, this implies that $\phi([0, +\infty)) = [0, +\infty)$. Since ϕ is strictly increasing, it follows that ϕ is a bijection from $[0, +\infty)$ onto $[0, +\infty)$.

Since the inverse of a continuous strictly increasing function is continuous and strictly increasing on its image (see, for example, (Bartle & Sherbert, 2000, Theorem 5.6.5)), ϕ^{-1} is continuous and strictly increasing on $[0, +\infty)$. Moreover, $\phi^{-1}(0) = 0$ as $\phi(0) = 0$. Finally, suppose that $\phi^{-1}(s)$ did not tend to infinity as $s \to +\infty$. Then there would exist $M > 0$ and a sequence $(s_n)_{n\geq 1}$ such that $s_n \to +\infty$ and $\phi^{-1}(s_n) \leq M$ for every $n \geq 1$. Since ϕ is increasing,

$$s_n = \phi\left(\phi^{-1}(s_n)\right) \leq \phi(M),$$

which contradicts $\lim_{n\to+\infty} s_n = +\infty$. Therefore, $\lim_{s\to+\infty} \phi^{-1}(s) = +\infty$. □

A.2 Properties of ω_ϕ and $\psi_{\phi,p}$

Given an Orlicz function ϕ and $p \in [1, +\infty)$, we recall that the fundamental function of the Orlicz space L^ϕ, equipped with the Luxemburg norm, is given by

$$\omega_\phi(u) \coloneqq \frac{1}{\phi^{-1}(1/u)}, \quad u \in (0, 1]$$

and $\omega_\phi(0) \coloneqq 0$. Its p-th power is given by

$$\psi_{\phi,p}(u) \coloneqq \omega_\phi(u)^p = \frac{1}{\phi^{-1}(1/u)^p}, \quad u \in (0, 1], \quad \psi_{\phi,p}(0) = 0.$$

Further properties of fundamental functions can be found in Bennett et al. (1988, Section 5) and Rubshtein et al. (2016, Sections 10.1 and 13.3). We state below the properties that will be used later. The following result is closely related to Bennett et al. (1988, Corollary 5.3).

Lemma 2 (Basic properties of ω_ϕ and $\psi_{\phi,p}$)**.** *The functions ω_ϕ and $\psi_{\phi,p}$ are continuous and strictly increasing on $[0, 1]$. Moreover, for every $a \geq 1$ and $u \in (0, 1]$ such that $au \leq 1$,*

$$\omega_\phi(au) \leq a\omega_\phi(u).$$

Proof. On $(0, 1]$, the continuity of ω_ϕ follows directly from Lemma 1. At the origin, Lemma 1 gives

$$\lim_{u\to 0^+} \phi^{-1}(1/u) = +\infty.$$

Therefore,

$$\lim_{u\to 0^+} \omega_\phi(u) = \lim_{u\to 0^+} \frac{1}{\phi^{-1}(1/u)} = 0 = \omega_\phi(0).$$

Thus, ω_ϕ is continuous on $[0, 1]$.

Let $0 < u_1 < u_2 \leq 1$. Then

$$\frac{1}{u_1} > \frac{1}{u_2}.$$

Since ϕ^{-1} is strictly increasing,

$$\phi^{-1}(1/u_1) > \phi^{-1}(1/u_2),$$

and hence

$$\omega_\phi(u_1) < \omega_\phi(u_2).$$

Also, $\omega_\phi(0) = 0 < \omega_\phi(u)$ for every $u > 0$. Hence ω_ϕ is strictly increasing on $[0, 1]$.

Since $p \geq 1$, the map $r \mapsto r^p$ is continuous and strictly increasing on $[0, +\infty)$. Therefore, $\psi_{\phi,p}$ is continuous and strictly increasing on $[0, 1]$.

Finally, by Rubshtein et al. (2016, Proposition 13.3.1), ω_ϕ is the fundamental function of the corresponding Orlicz space. Moreover, by Rubshtein et al. (2016, Theorem 10.4.2), every fundamental function of a symmetric space is quasiconcave; see Rubshtein et al. (2016, Definition 10.1.1) for the definition of a quasiconcave function. Therefore,

$$v \mapsto \frac{\omega_\phi(v)}{v}$$

is decreasing on $(0, 1]$. Now, let $a \geq 1$ and $u \in (0, 1]$ such that $au \leq 1$. Since $u \leq au$ and the map $v \mapsto \omega_\phi(v)/v$ is decreasing on $(0, 1]$,

$$\frac{\omega_\phi(au)}{au} \leq \frac{\omega_\phi(u)}{u}.$$

Hence,

$$\omega_\phi(au) \leq a\omega_\phi(u).$$

□

A.3 The existence and properties of $\psi^{\mathrm{cav}}_{\phi,p}$

Recall that $\psi^{\mathrm{cav}}_{\phi,p}$ denotes the least concave majorant of $\psi_{\phi,p}$ on $[0, 1]$, i.e., the smallest concave function on $[0, 1]$ that dominates $\psi_{\phi,p}$. We summarize below the properties that will be used later. For further properties of $\psi^{\mathrm{cav}}_{\phi,p}$, we refer the reader to Franců et al. (2017).

Lemma 3. *The least concave majorant $\psi^{\mathrm{cav}}_{\phi,p}$ exists and is continuous on $[0, 1]$. Moreover,*

$$\psi^{\mathrm{cav}}_{\phi,p}(0) = 0, \quad \psi^{\mathrm{cav}}_{\phi,p}(1) = \psi_{\phi,p}(1),$$

and $\psi^{\mathrm{cav}}_{\phi,p}$ is strictly increasing on $[0, 1]$.

Proof. Let $\mathcal{G}$ denote the family of concave functions on $[0, 1]$ dominating $\psi_{\phi,p}$. Since $\psi_{\phi,p}$ is strictly increasing on $[0, 1]$, we have

$$\psi_{\phi,p}(u) \leq \psi_{\phi,p}(1), \quad u \in [0, 1].$$

Hence, the constant function $G(u) \equiv \psi_{\phi,p}(1)$ is a concave majorant of $\psi_{\phi,p}$ on $[0, 1]$. Therefore, $\mathcal{G} \neq \varnothing$.

For $u \in [0, 1]$ define

$$H(u) \coloneqq \inf_{G \in \mathcal{G}} G(u).$$

Since every $G \in \mathcal{G}$ dominates $\psi_{\phi,p}$, so does H. Besides, since the pointwise infimum of a family of concave functions is concave (Boyd & Vandenberghe, 2004, Section 3.2.3), H is concave on $[0, 1]$. Thus, H is a concave majorant of $\psi_{\phi,p}$. Furthermore, by construction, $H \leq G$ for every $G \in \mathcal{G}$. Hence, H is the least concave majorant of $\psi_{\phi,p}$.

Applying Franců et al. (2017, Lemma 1) (with $F = \psi_{\phi,p}$ and $I = [0, 1]$) gives

$$\psi^{\mathrm{cav}}_{\phi,p} \text{ is continuous on } [0, 1], \quad \psi^{\mathrm{cav}}_{\phi,p}(0) = \psi_{\phi,p}(0) = 0, \quad \psi^{\mathrm{cav}}_{\phi,p}(1) = \psi_{\phi,p}(1).$$

By Lemma 2, $\psi_{\phi,p}$ is strictly increasing on $[0, 1]$, so it attains its maximum uniquely at 1. Therefore, applying Franců et al. (2017, Lemma 3) yields that $\psi^{\mathrm{cav}}_{\phi,p}$ is strictly increasing on $(0, 1)$. It remains only to include the two endpoints. For $u \in (0, 1]$,

$$\psi^{\mathrm{cav}}_{\phi,p}(u) \geq \psi_{\phi,p}(u) > 0 = \psi^{\mathrm{cav}}_{\phi,p}(0).$$

Thus, $\psi_{\phi,p}^{\mathrm{cav}}$ is strictly increasing on $[0,1)$. Now, fix $u \in [0,1)$. Choose w such that $u < w < 1$. For every $v \in (w,1)$, the strict monotonicity on $[0,1)$ yields

$$\psi_{\phi,p}^{\mathrm{cav}}(u) < \psi_{\phi,p}^{\mathrm{cav}}(w) < \psi_{\phi,p}^{\mathrm{cav}}(v).$$

Letting v tend to 1^- and using continuity of $\psi_{\phi,p}^{\mathrm{cav}}$ at 1, we obtain

$$\psi_{\phi,p}^{\mathrm{cav}}(u) < \psi_{\phi,p}^{\mathrm{cav}}(w) \leq \lim_{v \to 1^-} \psi_{\phi,p}^{\mathrm{cav}}(v) = \psi_{\phi,p}^{\mathrm{cav}}(1).$$

Thus, $\psi_{\phi,p}^{\mathrm{cav}}$ is strictly increasing on $[0,1]$. □

A.4 Elementary properties of the Luxemburg norm

We record below a few short proofs, derived directly from the definition of the Luxemburg norm, of properties that will be used later. For further properties, see, for example, Rubshtein et al. (2016, Chapter 13). Recall that, given a measure space (Ω, m) and an Orlicz function ϕ, the Luxemburg norm (Luxemburg et al., 1955) is defined by

$$\|f\|_\phi := \inf\left\{\lambda > 0 : \int_\Omega \phi\left(\frac{|f|}{\lambda}\right) \mathrm{d}m \leq 1\right\},$$

where the underlying measure m is omitted from the notation whenever it is clear from the context.

Lemma 4. *Let f and g be measurable functions with finite Luxemburg norms.*

1. *If $|f| \leq |g|$ m-almost everywhere, then*
$$\|f\|_\phi \leq \|g\|_\phi.$$
2. *For every scalar $c \in \mathbb{R}$,*
$$\|cf\|_\phi = |c|\,\|f\|_\phi.$$

Proof. Let $\lambda > 0$ be such that

$$\int_\Omega \phi\left(\frac{|g|}{\lambda}\right) \mathrm{d}m \leq 1. \tag{14}$$

Since ϕ is increasing and $|f| \leq |g|$ m-almost everywhere,

$$\phi\left(\frac{|f|}{\lambda}\right) \leq \phi\left(\frac{|g|}{\lambda}\right)$$

m-almost everywhere. Hence, if $\lambda > 0$ satisfies equation 14, then

$$\int_\Omega \phi\left(\frac{|f|}{\lambda}\right) \mathrm{d}m \leq 1.$$

Thus, every admissible λ in the definition of $\|g\|_\phi$ is also admissible in the definition of $\|f\|_\phi$. Taking the infimum yields $\|f\|_\phi \leq \|g\|_\phi$.

For the second assertion, the case $c = 0$ is immediate. Assume $c \neq 0$. Then

$$\begin{aligned}\|cf\|_\phi &= \inf\left\{\lambda > 0 : \int_\Omega \phi\left(\frac{|c|\,|f|}{\lambda}\right) \mathrm{d}m \leq 1\right\} \\ &= |c| \inf\left\{\eta > 0 : \int_\Omega \phi\left(\frac{|f|}{\eta}\right) \mathrm{d}m \leq 1\right\} = |c|\,\|f\|_\phi,\end{aligned}$$

where, in the second equality, we used the substitution $\lambda = |c|\eta$. □

B Proofs for Metric and Comparison Properties

We first establish a simple regularity property of the projected Orlicz–Wasserstein distance that will also be useful in the topological analysis.

Lemma 5. *For $\rho \in \mathcal{P}_\phi(\mathbb{R}^d)$, define*

$$m_\phi(\rho) \coloneqq \inf\left\{\lambda > 0 : \int_{\mathbb{R}^d} \phi\left(\frac{\|x\|}{\lambda}\right) \mathrm{d}\rho(x) \le 1\right\}. \tag{15}$$

Then, for any $\mu, \nu \in \mathcal{P}_\phi(\mathbb{R}^d)$ and $\theta, \eta \in \mathbb{S}^{d-1}$,

$$\left|\mathrm{W}_\phi(\Pi^\theta_\# \mu, \Pi^\theta_\# \nu) - \mathrm{W}_\phi(\Pi^\eta_\# \mu, \Pi^\eta_\# \nu)\right| \le \|\theta - \eta\|\big(m_\phi(\mu) + m_\phi(\nu)\big).$$

Proof. By the triangle inequality for W_ϕ,

$$\left|\mathrm{W}_\phi(\Pi^\theta_\# \mu, \Pi^\theta_\# \nu) - \mathrm{W}_\phi(\Pi^\eta_\# \mu, \Pi^\eta_\# \nu)\right| \le \mathrm{W}_\phi(\Pi^\theta_\# \mu, \Pi^\eta_\# \mu) + \mathrm{W}_\phi(\Pi^\theta_\# \nu, \Pi^\eta_\# \nu).$$

Let $X \sim \mu$. Using the coupling induced by the pair $(\langle\theta, X\rangle, \langle\eta, X\rangle)$ and Lemma 4, we obtain

$$\mathrm{W}_\phi\left(\Pi^\theta_\# \mu, \Pi^\eta_\# \mu\right) \le \|\langle\theta - \eta, X\rangle\|_\phi \le \|\theta - \eta\| \, \|\|X\|\|_\phi = \|\theta - \eta\| m_\phi(\mu).$$

The same argument for ν gives

$$\mathrm{W}_\phi(\Pi^\theta_\# \nu, \Pi^\eta_\# \nu) \le \|\theta - \eta\| m_\phi(\nu),$$

which proves the result. □

Proof of Theorem 1 and Proposition 1. Since W_ϕ is a metric (see, e.g., (Guha et al., 2023)), the metricity assertion in Theorem 1 also follows directly from Nadjahi et al. (2020), which shows that slicing preserves metricity. We nevertheless provide a direct proof here for completeness, together with the comparison inequality in Proposition 1).

We first establish the comparison inequality in Proposition 1, which also proves that $\mathrm{SOW}_{\phi,p}$ is finite on $\mathcal{P}_\phi(\mathbb{R}^d)$.

For every $\theta \in \mathbb{S}^{d-1}$,

$$|\langle\theta, x - y\rangle| \le \|x - y\| \quad \text{for all } x, y \in \mathbb{R}^d.$$

Combining with the monotonicity of ϕ, we obtain

$$\mathrm{W}_\phi(\Pi^\theta_\# \mu, \Pi^\theta_\# \nu) \le \mathrm{W}_\phi(\mu, \nu).$$

Since σ is a probability measure,

$$\mathrm{SOW}_{\phi,p}(\mu, \nu) \le \mathrm{W}_\phi(\mu, \nu) < \infty.$$

Non-negativity and symmetry follow directly from the corresponding properties of W_ϕ.

For the triangle inequality, let $\mu, \nu, \rho \in \mathcal{P}_\phi(\mathbb{R}^d)$. For every $\theta \in \mathbb{S}^{d-1}$,

$$\mathrm{W}_\phi(\Pi^\theta_\# \mu, \Pi^\theta_\# \rho) \le \mathrm{W}_\phi(\Pi^\theta_\# \mu, \Pi^\theta_\# \nu) + \mathrm{W}_\phi(\Pi^\theta_\# \nu, \Pi^\theta_\# \rho).$$

Applying Minkowski's inequality in $L^p(\mathbb{S}^{d-1}, \sigma)$ yields

$$\mathrm{SOW}_{\phi,p}(\mu, \rho) \le \mathrm{SOW}_{\phi,p}(\mu, \nu) + \mathrm{SOW}_{\phi,p}(\nu, \rho).$$

Clearly, $\mu = \nu$ implies $\mathrm{SOW}_{\phi,p}(\mu, \nu) = 0$. It remains to show that $\mathrm{SOW}_{\phi,p}(\mu, \nu) = 0$ implies $\mu = \nu$.

Suppose that $\mathrm{SOW}_{\phi,p}(\mu, \nu) = 0$. Then $\mathrm{W}_\phi(\Pi^\theta_\# \mu, \Pi^\theta_\# \nu) = 0$ for σ-almost every $\theta \in \mathbb{S}^{d-1}$. By Lemma 5, the map $\theta \mapsto \mathrm{W}_\phi(\Pi^\theta_\# \mu, \Pi^\theta_\# \nu)$ is continuous. Since σ has full support on $\mathbb{S}^{d-1}$, it follows that

$$\mathrm{W}_\phi(\Pi^\theta_\# \mu, \Pi^\theta_\# \nu) = 0 \qquad \text{for every } \theta \in \mathbb{S}^{d-1}.$$

Since W_ϕ is a metric (see, e.g., (Guha et al., 2023, Lemma 3.3)),

$$\Pi^\theta_\# \mu = \Pi^\theta_\# \nu \qquad \text{for every } \theta \in \mathbb{S}^{d-1}.$$

The Cramér–Wold theorem then gives $\mu = \nu$. □

Proof of Proposition 2. Since $\phi_1(t) \le \phi_2(t)$ for every $t \ge 0$, Guha et al. (2023, Lemma 3.4) yields

$$\mathrm{W}_{\phi_1}(\alpha, \beta) \le \mathrm{W}_{\phi_2}(\alpha, \beta)$$

for every admissible pair (α, β).

Applying this inequality to each projected pair gives

$$\mathrm{W}_{\phi_1}(\Pi^\theta_\# \mu, \Pi^\theta_\# \nu) \le \mathrm{W}_{\phi_2}(\Pi^\theta_\# \mu, \Pi^\theta_\# \nu)$$

for every $\theta \in \mathbb{S}^{d-1}$. Therefore,

$$\mathrm{SOW}_{\phi_1,p}(\mu, \nu) \le \mathrm{SOW}_{\phi_2,p}(\mu, \nu).$$

□

C Proofs and Counterexample for Topological Properties

C.1 Omitted Proofs for Topological Properties

We first establish a comparison between the radial Orlicz moment and the SOW distance to the Dirac measure at the origin.

Lemma 6. *For $\rho \in \mathcal{P}_\phi(\mathbb{R}^d)$, let $m_\phi(\rho)$ be defined in equation 15. Then there exists a constant $C_{d,p} > 0$, depending only on d and p, such that*

$$\mathrm{SOW}_{\phi,p}(\rho, \delta_0) \le m_\phi(\rho) \le C_{d,p}\, \mathrm{SOW}_{\phi,p}(\rho, \delta_0).$$

Proof. Let $X \sim \rho$. For every $\theta \in \mathbb{S}^{d-1}$, the monotonicity of the Luxemburg norm in Lemma 4 gives

$$\mathrm{W}_\phi(\Pi^\theta_\# \rho, \delta_0) = \|\langle \theta, X \rangle\|_\phi \le \|\|X\|\|_\phi = m_\phi(\rho).$$

Integrating over θ gives

$$\mathrm{SOW}_{\phi,p}(\rho, \delta_0) \le m_\phi(\rho).$$

For the reverse inequality, define

$$M \coloneqq \max_{\theta \in \mathbb{S}^{d-1}} \|\langle \theta, X \rangle\|_\phi.$$

The maximum exists by continuity of the map $\theta \mapsto \|\langle \theta, X \rangle\|_\phi$. Indeed, by Lemma 5, applied with $\mu = \rho$ and $\nu = \delta_0$, the map

$$\theta \mapsto \mathrm{W}_\phi(\Pi^\theta_\# \rho, \delta_0) = \|\langle \theta, X \rangle\|_\phi$$

is continuous on $\mathbb{S}^{d-1}$. Since $\mathbb{S}^{d-1}$ is compact, the maximum is attained.

Let $e_1, \dots, e_d$ be the canonical basis of $\mathbb{R}^d$. Since

$$\|X\| \le \sum_{j=1}^d |X_j|,$$

the triangle inequality for the Luxemburg norm gives

$$m_\phi(\rho) = \|\|X\|\|_\phi \le \sum_{j=1}^d \|X_j\|_\phi \le dM.$$

Choose $\theta_0 \in \mathbb{S}^{d-1}$ such that

$$\|\langle \theta_0, X \rangle\|_\phi = M.$$

For every $\theta \in \mathbb{S}^{d-1}$ satisfying $\|\theta - \theta_0\| \le 1/(2d)$,

$$\begin{aligned} \|\langle \theta, X \rangle\|_\phi &\ge \|\langle \theta_0, X \rangle\|_\phi - \|\langle \theta - \theta_0, X \rangle\|_\phi \\ &\ge M - \|\theta - \theta_0\|\, m_\phi(\rho) \\ &\ge \frac{M}{2}. \end{aligned}$$

Let

$$c_d \coloneqq \sigma\left(\left\{\theta \in \mathbb{S}^{d-1} : \|\theta - \theta_0\| \le \frac{1}{2d}\right\}\right) > 0.$$

By rotation invariance of the uniform measure σ, c_d does not depend on θ_0. Hence,

$$\mathrm{SOW}_{\phi,p}^p(\rho, \delta_0) \ge c_d \left(\frac{M}{2}\right)^p.$$

Therefore,

$$M \le 2c_d^{-1/p}\, \mathrm{SOW}_{\phi,p}(\rho, \delta_0),$$

and consequently,

$$m_\phi(\rho) \le 2dc_d^{-1/p}\, \mathrm{SOW}_{\phi,p}(\rho, \delta_0).$$

This proves the result. □

Proof of Proposition 3. Define $M_n \coloneqq \sup_{\theta \in \mathbb{S}^{d-1}} \mathrm{W}_\phi\left(\Pi_\#^\theta \mu_n, \Pi_\#^\theta \mu\right)$. Since σ is a probability measure,

$$\mathrm{SOW}_{\phi,p}(\mu_n, \mu) \le M_n.$$

Hence,

$$M_n \to 0 \quad \Longrightarrow \quad \mathrm{SOW}_{\phi,p}(\mu_n, \mu) \to 0.$$

We now prove the converse. Suppose that $\mathrm{SOW}_{\phi,p}(\mu_n, \mu) \to 0$.

$$m_\phi(\mu_n) \le C_{d,p}\, \mathrm{SOW}_{\phi,p}(\mu_n, \delta_0) \le C_{d,p}\left[\mathrm{SOW}_{\phi,p}(\mu_n, \mu) + \mathrm{SOW}_{\phi,p}(\mu, \delta_0)\right].$$

Since $\mathrm{SOW}_{\phi,p}(\mu_n, \mu) \to 0$, the sequence $\left(\mathrm{SOW}_{\phi,p}(\mu_n, \mu)\right)_{n\ge1}$ is bounded. Therefore,

$$\sup_{n\ge1} m_\phi(\mu_n) < \infty.$$

Since $m_\phi(\mu) < \infty$, define

$$L \coloneqq 1 + \sup_{n\ge1} m_\phi(\mu_n) + m_\phi(\mu) < \infty.$$

Then Lemma 5 gives, for every $\theta, \eta \in \mathbb{S}^{d-1}$ and every $n \ge 1$,

$$\left|\mathrm{W}_\phi(\Pi_\#^\theta \mu_n, \Pi_\#^\theta \mu) - \mathrm{W}_\phi(\Pi_\#^\eta \mu_n, \Pi_\#^\eta \mu)\right| \le L\|\theta - \eta\|.$$

Thus, the projected distance functions are uniformly Lipschitz with respect to the projection direction, with a constant independent of n.

Suppose, by contradiction, that M_n does not converge to zero. Then there exist $\varepsilon > 0$ and a subsequence $(n_k)_{k\ge1}$ such that

$$M_{n_k} \ge \varepsilon \qquad \text{for all } k \ge 1.$$

For each k, the map $\theta \mapsto \mathrm{W}_\phi(\Pi_\#^\theta \mu_{n_k}, \Pi_\#^\theta \mu)$ is continuous by Lemma 5. Since $\mathbb{S}^{d-1}$ is compact, there exists $\theta_k \in \mathbb{S}^{d-1}$ such that

$$\mathrm{W}_\phi\left(\Pi_\#^{\theta_k} \mu_{n_k}, \Pi_\#^{\theta_k} \mu\right) = M_{n_k}.$$

Set

$$r \coloneqq \min\left\{1, \frac{\varepsilon}{2L}\right\}.$$

For every $\theta \in \mathbb{S}^{d-1}$ satisfying $\|\theta - \theta_k\| \le r$, we have

$$\begin{aligned}
\mathrm{W}_\phi\left(\Pi_\#^\theta \mu_{n_k}, \Pi_\#^\theta \mu\right) &\ge \mathrm{W}_\phi\left(\Pi_\#^{\theta_k} \mu_{n_k}, \Pi_\#^{\theta_k} \mu\right) - L\|\theta - \theta_k\| \\
&\ge \varepsilon - Lr \\
&\ge \frac{\varepsilon}{2}.
\end{aligned}$$

By rotation invariance of σ, the spherical cap

$$\left\{\theta \in \mathbb{S}^{d-1} : \|\theta - \theta_k\| \le r\right\}$$

has measure $c_{d,r} > 0$, where $c_{d,r}$ does not depend on k. Therefore,

$$\begin{aligned}\mathrm{SOW}_{\phi,p}^p(\mu_{n_k}, \mu) &= \int_{\mathbb{S}^{d-1}} \mathrm{W}_\phi^p\left(\Pi_\#^\theta \mu_{n_k}, \Pi_\#^\theta \mu\right) \mathrm{d}\sigma(\theta) \\ &\ge c_{d,r}\left(\frac{\varepsilon}{2}\right)^p.\end{aligned}$$

The right-hand side is strictly positive and independent of k, which contradicts

$$\mathrm{SOW}_{\phi,p}(\mu_{n_k}, \mu) \xrightarrow{k\to+\infty} 0.$$

Hence,

$$M_n \xrightarrow{n\to+\infty} 0,$$

which completes the proof. □

We next prove Corollary 1. The main ingredient is the following comparison between the Orlicz–Wasserstein distance and the 1-Wasserstein distance.

Lemma 7. *For any $\alpha, \beta \in \mathcal{P}_\phi(\mathbb{R}^d)$,*

$$\mathrm{W}_1(\alpha, \beta) \le \phi^{-1}(1)\, \mathrm{W}_\phi(\alpha, \beta).$$

Proof. Let $\pi \in \Pi(\alpha, \beta)$ and let $\lambda > 0$ satisfy

$$\int_{\mathbb{R}^d \times \mathbb{R}^d} \phi\left(\frac{\|x - y\|}{\lambda}\right) \mathrm{d}\pi(x, y) \le 1.$$

Since π is a probability measure, Jensen's inequality gives

$$\phi\left(\frac{1}{\lambda}\int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - y\| \mathrm{d}\pi(x, y)\right) \le 1.$$

Since ϕ is strictly increasing,

$$\int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - y\| \mathrm{d}\pi(x, y) \le \phi^{-1}(1)\lambda.$$

Taking the infimum first over admissible λ and then over $\pi \in \Pi(\alpha, \beta)$ yields

$$\mathrm{W}_1(\alpha, \beta) \le \phi^{-1}(1)\, \mathrm{W}_\phi(\alpha, \beta).$$

□

Proof of Corollary 1. Assume that $\mathrm{SOW}_{\phi,p}(\mu_n, \mu) \to 0$. By Proposition 3,

$$\sup_{\theta \in \mathbb{S}^{d-1}} \mathrm{W}_\phi\left(\Pi_\#^\theta \mu_n, \Pi_\#^\theta \mu\right) \to 0.$$

Hence, for every $\theta \in \mathbb{S}^{d-1}$,

$$\mathrm{W}_\phi\left(\Pi_\#^\theta \mu_n, \Pi_\#^\theta \mu\right) \to 0.$$

By Lemma 7,

$$\mathrm{W}_1\left(\Pi_\#^\theta \mu_n, \Pi_\#^\theta \mu\right) \to 0.$$

Hence, by Villani (2008, Theorem 6.9),

$$\Pi_\#^\theta \mu_n \rightharpoonup \Pi_\#^\theta \mu \qquad \text{for every } \theta \in \mathbb{S}^{d-1}.$$

The Cramér–Wold theorem (Durrett, 2019, Theorem 3.10.6) then implies

$$\mu_n \rightharpoonup \mu.$$

□

We now turn to Proposition 4. The key ingredient is the following comparison between the Orlicz–Wasserstein distance and the 1-Wasserstein distance on bounded supports.

Lemma 8. *Let $D \in (0, \infty)$, and let α and β be probability measures supported on a common interval of length at most D. Then*

$$\mathrm{W}_\phi(\alpha, \beta) \le D\, \omega_\phi \left(\frac{\mathrm{W}_1(\alpha, \beta)}{D} \right),$$

where ω_ϕ is introduced in equation 4.

Proof. If $\mathrm{W}_1(\alpha, \beta) = 0$, then $\alpha = \beta$ and the result is immediate. Hence, assume that $\mathrm{W}_1(\alpha, \beta) > 0$.

Let $\pi^\star \in \Pi(\alpha, \beta)$ be an optimal coupling for $\mathrm{W}_1(\alpha, \beta)$. Since α and β are supported on a common interval of length at most D,

$$|x - y| \le D \qquad \pi^\star\text{-almost surely.}$$

Since $0 \le z \le D$, we have $z/D \in [0, 1]$. By the convexity of ϕ and $\phi(0) = 0$,

$$\phi\left(\frac{z}{t}\right) = \phi\left(\frac{z}{D}\frac{D}{t} + \left(1 - \frac{z}{D}\right) 0\right) \le \frac{z}{D}\phi\left(\frac{D}{t}\right).$$

Therefore,

$$\begin{aligned}\int_{\mathbb{R}\times\mathbb{R}} \phi\left(\frac{|x-y|}{t}\right) \mathrm{d}\pi^\star(x, y) &\le \frac{\phi(D/t)}{D} \int_{\mathbb{R}\times\mathbb{R}} |x - y| \,\mathrm{d}\pi^\star(x, y) \\ &= \frac{\phi(D/t)}{D}\, \mathrm{W}_1(\alpha, \beta).\end{aligned}$$

Choosing

$$t = \frac{D}{\phi^{-1}\left(D/\,\mathrm{W}_1(\alpha, \beta)\right)},$$

we obtain

$$\int_{\mathbb{R}\times\mathbb{R}} \phi\left(\frac{|x-y|}{t}\right) \mathrm{d}\pi^\star(x, y) \le 1.$$

Hence, by the definition of W_ϕ,

$$\mathrm{W}_\phi(\alpha, \beta) \le \frac{D}{\phi^{-1}\left(D/\,\mathrm{W}_1(\alpha, \beta)\right)} = D\omega_\phi\left(\frac{\mathrm{W}_1(\alpha, \beta)}{D}\right).$$

□

Proof of Proposition 4. The implication

$$\mathrm{SOW}_{\phi,p}(\mu_n, \mu) \to 0 \quad \Longrightarrow \quad \mu_n \rightharpoonup \mu$$

follows from Corollary 1. We only need to prove the reverse implication.

Assume that

$$\mu_n \rightharpoonup \mu,$$

where $(\mu_n)_{n\ge1} \subset \mathcal{P}(K)$ and $\mu \in \mathcal{P}(K)$ for some compact set $K \subset \mathbb{R}^d$. Let

$$D \coloneqq \mathrm{diam}(K).$$

If $D = 0$, then K is a singleton, so $\mu_n = \mu$ for every n and the conclusion is immediate. Hence, we assume that $D > 0$. Fix $\theta \in \mathbb{S}^{d-1}$. Since Π^θ is continuous, the continuous mapping theorem (see, e.g., (Kallenberg, 1997, Lemma 4.3)) gives

$$\Pi^\theta_\# \mu_n \rightharpoonup \Pi^\theta_\# \mu.$$

Moreover, all projected measures are supported on a common compact interval of length at most D. Therefore, by Villani (2008, Theorem 6.9),

$$\mathrm{W}_1\left(\Pi^\theta_\# \mu_n, \Pi^\theta_\# \mu\right) \to 0.$$

Applying Lemma 8, we obtain

$$\mathrm{W}_\phi\left(\Pi^\theta_\# \mu_n, \Pi^\theta_\# \mu\right) \le D\omega_\phi\left(\frac{\mathrm{W}_1(\Pi^\theta_\# \mu_n, \Pi^\theta_\# \mu)}{D}\right).$$

By Lemma 2, $\omega_\phi(u) \to 0$ as $u \to 0$. Therefore,

$$\mathrm{W}_\phi\left(\Pi^\theta_\# \mu_n, \Pi^\theta_\# \mu\right) \to 0 \text{ for every } \theta \in \mathbb{S}^{d-1}.$$

Furthermore, since all projected measures are supported on an interval of length at most D,

$$\mathrm{W}_\phi\left(\Pi^\theta_\# \mu_n, \Pi^\theta_\# \mu\right) \le \frac{D}{\phi^{-1}(1)},$$

for every $n \ge 1$ and every $\theta \in \mathbb{S}^{d-1}$. Hence,

$$\mathrm{W}^p_\phi\left(\Pi^\theta_\# \mu_n, \Pi^\theta_\# \mu\right) \le \left(\frac{D}{\phi^{-1}(1)}\right)^p.$$

Since σ is a probability measure, the right-hand side is integrable. Therefore, by the dominated convergence theorem (see, e.g., (Kallenberg, 1997, Theorem 1.21)),

$$\mathrm{SOW}^p_{\phi,p}(\mu_n, \mu) = \int_{\mathbb{S}^{d-1}} \mathrm{W}^p_\phi\left(\Pi^\theta_\# \mu_n, \Pi^\theta_\# \mu\right) \mathrm{d}\sigma(\theta) \to 0.$$

Hence, $\mathrm{SOW}_{\phi,p}(\mu_n, \mu) \to 0$. This completes the proof. □

C.2 Counterexample to the Converse of Corollary 1

We provide a counterexample showing that weak convergence does not, in general, imply convergence in $\mathrm{SOW}_{\phi,p}$, even in dimension one.

Let

$$\mu = \delta_0, \qquad \mu_n = \left(1 - \frac{1}{n}\right)\delta_0 + \frac{1}{n}\delta_{\phi^{-1}(n)}, \qquad n \ge 1.$$

We first show that $\mu_n \rightharpoonup \mu$. Indeed, for every bounded continuous function $f : \mathbb{R} \to \mathbb{R}$,

$$\int_{\mathbb{R}} f \,\mathrm{d}\mu_n = \left(1 - \frac{1}{n}\right) f(0) + \frac{1}{n} f\big(\phi^{-1}(n)\big),$$

while

$$\int_{\mathbb{R}} f \,\mathrm{d}\mu = f(0).$$

Therefore,

$$\begin{aligned}\left|\int_{\mathbb{R}} f \,\mathrm{d}\mu_n - \int_{\mathbb{R}} f \,\mathrm{d}\mu\right| &= \frac{1}{n}\left|f\big(\phi^{-1}(n)\big) - f(0)\right| \\ &\le \frac{2\|f\|_\infty}{n} \longrightarrow 0.\end{aligned}$$

Hence, $\mu_n \rightharpoonup \mu$.

On the other hand,

$$\begin{aligned}\mathrm{W}_\phi(\mu_n, \mu) &= \inf\left\{\lambda > 0 : \frac{1}{n}\phi\left(\frac{\phi^{-1}(n)}{\lambda}\right) \le 1\right\} \\ &= \inf\left\{\lambda > 0 : \phi\left(\frac{\phi^{-1}(n)}{\lambda}\right) \le n\right\}.\end{aligned}$$

Since ϕ is strictly increasing,

$$\phi\left(\frac{\phi^{-1}(n)}{\lambda}\right) \le n = \phi\big(\phi^{-1}(n)\big)$$

if and only if

$$\frac{\phi^{-1}(n)}{\lambda} \le \phi^{-1}(n),$$

or equivalently, $\lambda \ge 1$. Thus,

$$\mathrm{W}_\phi(\mu_n, \mu) = 1 \text{ for every } n \ge 1.$$

Finally, when $d = 1$, $\mathbb{S}^0 = \{-1, +1\}$. For either projection direction $\theta \in \{-1, +1\}$,

$$\mathrm{W}_\phi\left(\Pi^\theta_\# \mu_n, \Pi^\theta_\# \mu\right) = 1.$$

Therefore,

$$\mathrm{SOW}^p_{\phi,p}(\mu_n, \mu) = \frac{1}{2}\left(1^p + 1^p\right) = 1,$$

and hence

$$\mathrm{SOW}_{\phi,p}(\mu_n, \mu) = 1 \text{ for every } n \ge 1.$$

Thus, although $\mu_n \rightharpoonup \mu$, we have $\mathrm{SOW}_{\phi,p}(\mu_n, \mu) \not\to 0$. This shows that the converse of Corollary 1 does not hold in general.

D Proofs for Statistical Properties

We first establish a link between the one-dimensional Orlicz–Wasserstein distance and the uniform distance between the corresponding cumulative distribution functions.

Lemma 9. *Let $D > 0$, and let α and β be probability measures on $\mathbb{R}$ supported on a common interval of length at most D. Let F_α and F_β denote their respective cumulative distribution functions. Then*

$$\mathrm{W}_\phi(\alpha, \beta) \le D\omega_\phi\left(\|F_\alpha - F_\beta\|_\infty\right),$$

where

$$\|F_\alpha - F_\beta\|_\infty := \sup_{x \in \mathbb{R}} |F_\alpha(x) - F_\beta(x)|.$$

Proof. By Lemma 8,

$$\mathrm{W}_\phi(\alpha, \beta) \le D\omega_\phi\left(\frac{\mathrm{W}_1(\alpha, \beta)}{D}\right).$$

On the other hand, the one-dimensional CDF representation of W_1 (Bobkov & Ledoux, 2019, Theorem 2.9) gives

$$\mathrm{W}_1(\alpha, \beta) = \int_{\mathbb{R}} |F_\alpha(x) - F_\beta(x)| \mathrm{d}x.$$

Since α and β are supported on a common interval of length at most D,

$$\mathrm{W}_1(\alpha, \beta) \le D\|F_\alpha - F_\beta\|_\infty.$$

Since ω_ϕ is increasing by Lemma 2, we conclude that

$$\mathrm{W}_\phi(\alpha, \beta) \le D\omega_\phi\left(\|F_\alpha - F_\beta\|_\infty\right).$$

□

We next derive a moment bound for the fundamental function ω_ϕ under a sub-Gaussian tail condition.

Lemma 10. *Let $n \in \mathbb{N}$ with $n \ge 1$ and let $0 \le \Delta_n \le 1$ be a random variable satisfying*

$$\mathbb{P}(\Delta_n > u) \le 2\exp(-2nu^2), \qquad u \ge 0.$$

Then, for every $p \in [1, \infty)$, there exists a constant $C_p > 0$, depending only on p, such that

$$[\mathbb{E}\omega_\phi(\Delta_n)^p]^{1/p} \le C_p\, \omega_\phi(n^{-1/2}).$$

Proof. We use a dyadic decomposition. Define

$$E_* := \left\{\Delta_n \le n^{-1/2}\right\},$$

and, for $k \ge 0$,

$$E_k := \left\{2^k n^{-1/2} < \Delta_n \le \min\left\{2^{k+1} n^{-1/2}, 1\right\}\right\}.$$

On E_*, since ω_ϕ is increasing by Lemma 2,

$$\omega_\phi(\Delta_n) \le \omega_\phi(n^{-1/2}),$$

and therefore

$$\mathbb{E}\left[\omega_\phi(\Delta_n)^p \mathbb{1}_{E_*}\right] \le \omega_\phi(n^{-1/2})^p.$$

Fix $k \ge 0$. On E_k, we claim that

$$\omega_\phi(\Delta_n) \le 2^{k+1}\omega_\phi(n^{-1/2}).$$

Indeed, if $2^{k+1}n^{-1/2} \le 1$, then this follows directly from Lemma 2. We now assume that $2^{k+1}n^{-1/2} > 1$. Applying Lemma 2 (with $a = n^{1/2}$ and $u = n^{-1/2}$), we obtain

$$\omega_\phi(\Delta_n) \le \omega_\phi(1) = \omega_\phi\left(n^{1/2}n^{-1/2}\right) \le n^{1/2}\omega_\phi(n^{-1/2}) < 2^{k+1}\omega_\phi(n^{-1/2}).$$

Hence,

$$\begin{aligned}\mathbb{E}\left[\omega_\phi(\Delta_n)^p \mathbb{1}_{E_k}\right] &\le 2^{p(k+1)}\omega_\phi(n^{-1/2})^p \mathbb{P}\left(\Delta_n > 2^k n^{-1/2}\right) \\ &\le 2^{p(k+1)+1}\omega_\phi(n^{-1/2})^p \exp(-2\cdot 4^k).\end{aligned}$$

Since $E_*, E_0, E_1, \ldots$ form a partition of the sample space,

$$\mathbb{E}\omega_\phi(\Delta_n)^p \le \omega_\phi(n^{-1/2})^p \left[1 + 2\sum_{k\ge 0} 2^{p(k+1)} \exp(-2\cdot 4^k)\right].$$

The series is finite and depends only on p. Taking the p-th root completes the proof. □

We are now ready to prove the one-sample empirical convergence bound.

Proof of Theorem 2. Fix $\theta \in \mathbb{S}^{d-1}$. Since μ is supported on a set of diameter at most D, the projected measure $\Pi^\theta_\# \mu$ is supported on an interval of length at most D.

Let F_θ denote the CDF of $\Pi^\theta_\# \mu$, and let $\widehat{F}_{n,\theta}$ denote the empirical CDF associated with $\Pi^\theta_\# \widehat{\mu}_n$. By the Dvoretzky–Kiefer–Wolfowitz inequality (Massart, 1990),

$$\mathbb{P}\left(\|\widehat{F}_{n,\theta} - F_\theta\|_\infty > u\right) \le 2\exp(-2nu^2), \quad u \ge 0.$$

Since $\|\widehat{F}_{n,\theta} - F_\theta\|_\infty \in [0,1]$, Lemma 10 yields

$$\left[\mathbb{E}\omega_\phi\left(\|\widehat{F}_{n,\theta} - F_\theta\|_\infty\right)^p\right]^{1/p} \le C_p \omega_\phi(n^{-1/2}).$$

Applying Lemma 9, we obtain

$$\mathbb{E}\,\mathrm{W}^p_\phi\left(\Pi^\theta_\# \widehat{\mu}_n, \Pi^\theta_\# \mu\right) \le C_p^p D^p \omega_\phi(n^{-1/2})^p.$$

Integrating over θ and using Tonelli's theorem,

$$\begin{aligned}\mathbb{E}\,\mathrm{SOW}^p_{\phi,p}\left(\widehat{\mu}_n, \mu\right) &= \int_{\mathbb{S}^{d-1}} \mathbb{E}\,\mathrm{W}^p_\phi\left(\Pi^\theta_\# \widehat{\mu}_n, \Pi^\theta_\# \mu\right) \mathrm{d}\sigma(\theta) \\ &\le C_p^p D^p \omega_\phi(n^{-1/2})^p.\end{aligned}$$

Taking the p-th root gives

$$\left[\mathbb{E}\,\mathrm{SOW}^p_{\phi,p}\left(\widehat{\mu}_n, \mu\right)\right]^{1/p} \le C_p D\, \omega_\phi(n^{-1/2}),$$

which proves the theorem. □

We finally prove the two-sample bound for the SOW distance.

Proof of Corollary 2. Once we have established Theorem 2, this corollary can be deduced directly by applying Nadjahi et al. (2020, Theorem 5). However, we provide the proof for completeness.

By Jensen's inequality and Theorem 2,

$$\mathbb{E}\,\mathrm{SOW}_{\phi,p}\left(\widehat{\mu}_n,\mu\right) \le \left[\mathbb{E}\,\mathrm{SOW}^p_{\phi,p}\left(\widehat{\mu}_n,\mu\right)\right]^{1/p} \le C_p D \omega_\phi(n^{-1/2}).$$

Similarly,

$$\mathbb{E}\,\mathrm{SOW}_{\phi,p}\left(\widehat{\nu}_m,\nu\right) \le C_p D \omega_\phi(m^{-1/2}).$$

Since $\mathrm{SOW}_{\phi,p}$ is a metric,

$$\left|\mathrm{SOW}_{\phi,p}\left(\widehat{\mu}_n,\widehat{\nu}_m\right) - \mathrm{SOW}_{\phi,p}(\mu,\nu)\right| \le \mathrm{SOW}_{\phi,p}\left(\widehat{\mu}_n,\mu\right) + \mathrm{SOW}_{\phi,p}\left(\widehat{\nu}_m,\nu\right).$$

Taking expectations and combining the preceding bounds gives

$$\mathbb{E}\left|\mathrm{SOW}_{\phi,p}\left(\widehat{\mu}_n,\widehat{\nu}_m\right) - \mathrm{SOW}_{\phi,p}(\mu,\nu)\right| \le C_p D\left[\omega_\phi(n^{-1/2}) + \omega_\phi(m^{-1/2})\right].$$

This completes the proof. □

E Proof of Theorem 3

We first establish several auxiliary results for the powered Luxemburg functional. These results will be used to control its variation under L^1 perturbations.

We begin by defining

$$\psi^\sharp_{\phi,p}(u) \coloneqq u \sup_{u\le v\le 1} \frac{\psi_{\phi,p}(v)}{v}, \qquad u \in (0,1], \tag{16}$$

with $\psi^\sharp_{\phi,p}(0) \coloneqq 0$, where $\psi_{\phi,p}$ is defined in equation 9.

We record some elementary properties of $\psi^\sharp_{\phi,p}$.

Lemma 11. *The function $\psi^\sharp_{\phi,p}$ is increasing on $[0,1]$, the map*

$$u \longmapsto \frac{\psi^\sharp_{\phi,p}(u)}{u}$$

is decreasing on $(0,1]$, and

$$\psi^\sharp_{\phi,p}(u) \ge \psi_{\phi,p}(u), \quad u \in [0,1].$$

Proof. For $u = 0$, both sides are equal to zero. For $u \in (0,1]$,

$$\frac{\psi^\sharp_{\phi,p}(u)}{u} = \sup_{u\le v\le 1} \frac{\psi_{\phi,p}(v)}{v},$$

which is decreasing on $(0,1]$. Moreover, choosing $v = u$ in the supremum gives

$$\psi^\sharp_{\phi,p}(u) \ge \psi_{\phi,p}(u).$$

It remains to prove that $\psi^\sharp_{\phi,p}$ is increasing. Let $0 < u_1 < u_2 \le 1$. For $v \in [u_1,u_2]$, since $\psi_{\phi,p}$ is increasing by Lemma 2,

$$\frac{u_1}{v}\psi_{\phi,p}(v) \le \psi_{\phi,p}(v) \le \psi_{\phi,p}(u_2) \le \psi^\sharp_{\phi,p}(u_2).$$

For $v \in [u_2,1]$,

$$\frac{u_1}{v}\psi_{\phi,p}(v) \le \frac{u_2}{v}\psi_{\phi,p}(v) \le \psi^\sharp_{\phi,p}(u_2).$$

Taking the supremum over $v \in [u_1,1]$ yields

$$\psi^\sharp_{\phi,p}(u_1) \le \psi^\sharp_{\phi,p}(u_2).$$

The case $u_1 = 0$ follows from $\psi^\sharp_{\phi,p}(0) = 0$. □

To connect $\psi^\sharp_{\phi,p}$ with the modulus appearing in Theorem 3, we first show that $\psi_{\phi,p}$ and $\psi^\sharp_{\phi,p}$ have the same least concave majorant on $[0,1]$.

Lemma 12. *The functions $\psi_{\phi,p}$ defined in equation 9 and $\psi^\sharp_{\phi,p}$ defined in equation 16 have the same least concave majorant.*

Proof. By Lemma 11, $\psi_{\phi,p} \le \psi^\sharp_{\phi,p}$ on $[0,1]$. Hence, every concave function dominating $\psi^\sharp_{\phi,p}$ also dominates $\psi_{\phi,p}$.

Let g be a concave function on $[0,1]$ such that

$$g(u) \ge \psi_{\phi,p}(u), \quad u \in [0,1].$$

Since $\psi_{\phi,p}(0) = 0$, we have $g(0) \ge 0$. For any $0 < u \le v \le 1$, we write

$$u = \frac{u}{v} v + \left(1 - \frac{u}{v}\right) 0.$$

By the concavity of g,

$$g(u) \ge \frac{u}{v} g(v) + \left(1 - \frac{u}{v}\right) g(0) \ge \frac{u}{v} \psi_{\phi,p}(v).$$

Taking the supremum over $v \in [u,1]$ gives

$$g(u) \ge u \sup_{u \le v \le 1} \frac{\psi_{\phi,p}(v)}{v} = \psi^\sharp_{\phi,p}(u).$$

At $u = 0$, we also have

$$g(0) \ge 0 = \psi^\sharp_{\phi,p}(0).$$

Consequently, $g(u) \ge \psi^\sharp_{\phi,p}(u) \, \forall u \in [0,1]$. Therefore, every concave function dominating $\psi_{\phi,p}$ also dominates $\psi^\sharp_{\phi,p}$.

As a result, $\psi_{\phi,p}$ and $\psi^\sharp_{\phi,p}$ are dominated by exactly the same concave functions, and consequently have the same least concave majorant. □

We now relate $\psi^\sharp_{\phi,p}$ to $\psi^{\mathrm{cav}}_{\phi,p}$ used in Theorem 3. By Lemma 12, $\psi^{\mathrm{cav}}_{\phi,p}$ is also the least concave majorant of $\psi^\sharp_{\phi,p}$. Moreover, by Lemma 11, $\psi^\sharp_{\phi,p}$ is quasiconcave ; see (Bennett et al., 1988, Definition 5.6). Therefore, Bennett et al. (1988, Proposition 5.10) yields

$$\psi^\sharp_{\phi,p}(u) \le \psi^{\mathrm{cav}}_{\phi,p}(u) \le 2\psi^\sharp_{\phi,p}(u), \qquad u \in [0,1]. \tag{17}$$

We next record a useful property of the Luxemburg norm; see, e.g., (Rao & Ren, 1991, Proposition 6, page 77). We state the version needed in our setting and provide its proof. To be more precise, we note that, for the measurable profiles in this appendix, "nonzero" means nonzero on a set of positive Lebesgue measure. The notation $f \equiv 0$ means that $f = 0$ Lebesgue-almost everywhere.

Lemma 13. *Let $0 \le h \le 1$ be a nonzero measurable function and let $b = \|h\|_\phi$. Then*

$$\int_0^1 \phi\left(\frac{h(s)}{b}\right) \mathrm{d}s = 1.$$

Proof. Define

$$G(\lambda) := \int_0^1 \phi\left(\frac{h(s)}{\lambda}\right) \mathrm{d}s, \qquad \lambda > 0.$$

Since $0 \le h \le 1$ and ϕ is continuous on $[0,\infty)$, the map G is continuous on $(0,\infty)$. Besides, since ϕ is increasing, G is decreasing. Moreover, by the definition of the Luxemburg norm,

$$b = \inf\{\lambda > 0 : G(\lambda) \le 1\}.$$

Hence, $G(\lambda) > 1$ for every $\lambda < b$. If $\lambda > b$, by the definition of the infimum, there exists $\lambda' \in [b, \lambda)$ such that $G(\lambda') \le 1$. Since G is decreasing, $G(\lambda) \le G(\lambda') \le 1$. Moreover, since G is continuous on $(0,\infty)$, we deduce $G(b) = 1$. □

We now construct a linear functional associated with the Luxemburg norm that will be useful for deriving the upper bound below.

Lemma 14. *Let $0 \le h \le 1$ be a nonzero measurable function. For $s \in [0,1]$, we define*

$$z(s) \coloneqq \frac{h(s)}{\|h\|_\phi}, \qquad \xi(s) \coloneqq \phi'_+\big(z(s)\big),$$

and

$$D_0 \coloneqq \int_0^1 z(s)\xi(s)\mathrm{d}s, \qquad y(s) \coloneqq \frac{\xi(s)}{D_0},$$

where ϕ'_+ denotes the right derivative of ϕ. Then $1 \le D_0 < +\infty$, and for every $0 \le r \le 1$,

$$\int_0^1 r(s)y(s)\mathrm{d}s \le \|r\|_\phi.$$

Moreover,

$$\int_0^1 h(s)y(s)\mathrm{d}s = \|h\|_\phi.$$

Proof. Fix $s \in [0,1]$. Since ϕ is convex, $\phi'_+(z(s))$ is a subgradient of ϕ at $z(s)$. Hence, for every $t \ge 0$,

$$\phi(t) \ge \phi(z(s)) + \phi'_+(z(s))\,(t - z(s))\,. \tag{18}$$

Taking $t = 0$ and using $\phi(0) = 0$, we obtain

$$z(s)\phi'_+(z(s)) \ge \phi(z(s)).$$

Since $\xi(s) = \phi'_+(z(s))$, it follows that

$$z(s)\xi(s) \ge \phi(z(s)).$$

Integrating over $s \in [0,1]$ and using Lemma 13, we have

$$D_0 = \int_0^1 z(s)\xi(s)\mathrm{d}s \ge \int_0^1 \phi(z(s))\mathrm{d}s = 1.$$

Since $0 \le z(s) \le \|h\|_\phi^{-1}$ for every $s \in [0,1]$ and since ϕ'_+ is finite-valued and increasing on $[0,+\infty)$, we have

$$0 \le D_0 \le \|h\|_\phi^{-1}\phi'_+\left(\|h\|_\phi^{-1}\right) < +\infty.$$

Let $0 \le r \le 1$. If $r \equiv 0$, the inequality is immediate. Hence, assume that r is nonzero and define

$$w(s) \coloneqq \frac{r(s)}{\|r\|_\phi}.$$

Again fix $s \in [0,1]$. Applying the subgradient inequality at $z(s)$ with $t = w(s)$ gives

$$\phi(w(s)) \ge \phi(z(s)) + \xi(s)\big(w(s) - z(s)\big).$$

Integrating over $[0,1]$ and using Lemma 13,

$$\int_0^1 \phi(w(s))\mathrm{d}s = \int_0^1 \phi(z(s))\mathrm{d}s = 1.$$

Therefore,

$$\int_0^1 \xi(s)w(s)\mathrm{d}s \le \int_0^1 \xi(s)z(s)\mathrm{d}s = D_0.$$

Since $r(s) = \|r\|_\phi w(s)$ and $y(s) = \xi(s)/D_0$,

$$\int_0^1 r(s)y(s)\mathrm{d}s = \frac{\|r\|_\phi}{D_0}\int_0^1 w(s)\xi(s)\mathrm{d}s \le \|r\|_\phi.$$

Finally, since $h(s) = \|h\|_\phi z(s)$, we have

$$\int_0^1 h(s)y(s)\mathrm{d}s = \frac{\|h\|_\phi}{D_0}\int_0^1 z(s)\xi(s)\mathrm{d}s = \|h\|_\phi.$$

□

The following identity expresses the regularization $\psi^{\sharp}_{\phi,p}$ directly in terms of the growth of ϕ and will be useful for the upper bound below.

Lemma 15. *For every $u \in (0,1]$,*

$$\frac{\psi^{\sharp}_{\phi,p}(u)}{u} = \sup_{\phi^{-1}(1)\le t\le \phi^{-1}(1/u)} \frac{\phi(t)}{t^p}.$$

Proof. By definition,

$$\frac{\psi^{\sharp}_{\phi,p}(u)}{u} = \sup_{u\le v\le 1} \frac{\psi_{\phi,p}(v)}{v}.$$

For $v \in [u,1]$, set

$$t := \phi^{-1}(1/v).$$

Since ϕ is strictly increasing,

$$v = \frac{1}{\phi(t)},$$

and, as v varies over $[u,1]$, t varies over $\left[\phi^{-1}(1),\ \phi^{-1}(1/u)\right]$. Moreover,

$$\frac{\psi_{\phi,p}(v)}{v} = \frac{1}{\phi^{-1}(1/v)^p}\frac{1}{v} = \frac{\phi(t)}{t^p}.$$

Taking the supremum over $v \in [u,1]$ gives the result. □

We now introduce a quantity that measures the largest possible variation between $\|f\|^p_\phi$ and $\|g\|^p_\phi$ among functions f and g that are close in L^1.

$$\Omega_{\phi,p}(u) := \sup_{\substack{0\le f,g\le 1\\ \|f-g\|_{L^1(0,1)}\le u}} \left|\|f\|^p_\phi - \|g\|^p_\phi\right|, \quad u\in[0,1]. \tag{19}$$

For the upper-bound argument in Theorem 3, it will be useful to work with pointwise ordered pairs f and g. By symmetry, one may interchange f and g so that $\|f\|_\phi \ge \|g\|_\phi$, but this does not imply $f \ge g$ pointwise. However, replacing f and g by their pointwise maximum and minimum preserves their L^1-distance and can only increase the difference of their powered Luxemburg norms. We formalize this observation in the following lemma.

Lemma 16. *For every $u \in [0,1]$,*

$$\Omega_{\phi,p}(u) = \sup_{\substack{0\le g\le f\le 1\\ \|f-g\|_{L^1(0,1)}\le u}} \left(\|f\|^p_\phi - \|g\|^p_\phi\right).$$

Proof. Let (f,g) be any pair satisfying

$$0\le f,g\le 1, \quad \|f-g\|_{L^1(0,1)}\le u.$$

Define $F := \max\{f,g\}$, $\quad G := \min\{f,g\}$. Then, $0\le G\le F\le 1$ and

$$F-G = |f-g|, \quad \|F-G\|_{L^1(0,1)} = \|f-g\|_{L^1(0,1)} \le u.$$

Moreover, by the monotonicity of the Luxemburg norm,

$$\|F\|_\phi \ge \max\{\|f\|_\phi, \|g\|_\phi\}, \quad \|G\|_\phi \le \min\{\|f\|_\phi, \|g\|_\phi\}.$$

Hence,

$$\|F\|^p_\phi - \|G\|^p_\phi \ge \max\{\|f\|^p_\phi, \|g\|^p_\phi\} - \min\{\|f\|^p_\phi, \|g\|^p_\phi\} = \left|\|f\|^p_\phi - \|g\|^p_\phi\right|.$$

It follows that

$$\Omega_{\phi,p}(u) \le \sup_{\substack{0\le G\le F\le 1\\ \|F-G\|_{L^1(0,1)}\le u}} \left(\|F\|^p_\phi - \|G\|^p_\phi\right).$$

Conversely, every ordered pair $0 \le G \le F \le 1$ satisfying $\|F - G\|_{L^1(0,1)} \le u$ is admissible in the definition of $\Omega_{\phi,p}(u)$. Moreover, since $G \le F$, the monotonicity of the Luxemburg norm gives $\|G\|_\phi \le \|F\|_\phi$. Hence,

$$\|F\|_\phi^p - \|G\|_\phi^p = \left| \|F\|_\phi^p - \|G\|_\phi^p \right| \le \Omega_{\phi,p}(u).$$

Taking the supremum over such (F, G) gives the reverse inequality. Relabeling F and G as f, g completes the proof. □

With this reduction in hand, we are now ready to establish the main technical estimate for $\Omega_{\phi,p}$.

Proposition 6. *For every $u \in [0, 1]$,*

$$\psi_{\phi,p}^\sharp(u) \le \Omega_{\phi,p}(u) \le p2^p \psi_{\phi,p}^\sharp(u).$$

Consequently,

$$\frac{1}{2}\psi_{\phi,p}^{\mathrm{cav}}(u) \le \Omega_{\phi,p}(u) \le p2^p \psi_{\phi,p}^{\mathrm{cav}}(u).$$

Proof. The case $u = 0$ is immediate. Hence, let $u \in (0, 1]$.

Lower bound. Fix $v \in [u, 1]$ and choose a measurable set $A \subset (0, 1)$ with $|A| = v$. Define

$$f := \mathbb{1}_A, \quad g := \left(1 - \frac{u}{v}\right) \mathbb{1}_A.$$

Then $0 \le g \le f \le 1$ and

$$\|f - g\|_{L^1(0,1)} = \frac{u}{v}|A| = u.$$

Since $\|\mathbb{1}_A\|_\phi = \omega_\phi(v)$, the homogeneity of the Luxemburg norm gives

$$\|f\|_\phi = \omega_\phi(v), \quad \|g\|_\phi = \left(1 - \frac{u}{v}\right) \omega_\phi(v).$$

Therefore,

$$\|f\|_\phi^p - \|g\|_\phi^p = \omega_\phi(v)^p \left[1 - \left(1 - \frac{u}{v}\right)^p\right] = \psi_{\phi,p}(v) \left[1 - \left(1 - \frac{u}{v}\right)^p\right].$$

Since $0 \le 1 - u/v \le 1$ and $p \ge 1$,

$$1 - \left(1 - \frac{u}{v}\right)^p \ge \frac{u}{v}.$$

Hence

$$\Omega_{\phi,p}(u) \ge \frac{u}{v} \psi_{\phi,p}(v).$$

Taking the supremum over $v \in [u, 1]$ yields

$$\Omega_{\phi,p}(u) \ge u \sup_{u \le v \le 1} \frac{\psi_{\phi,p}(v)}{v} = \psi_{\phi,p}^\sharp(u).$$

Upper bound. By Lemma 16, it suffices to consider ordered pairs. Fix $0 \le g \le f \le 1$ satisfying

$$\|f - g\|_{L^1(0,1)} \le u.$$

If $f \equiv 0$, then $g \equiv 0$ and the claim is trivial. Hence, assume $f \not\equiv 0$.

We distinguish two cases.

Case 1. We assume

$$\|f\|_\phi \le \frac{2}{\phi^{-1}(1/u)}.$$

Then, for $p \in [1, +\infty)$, using Lemma 11, we have

$$\begin{aligned}\|f\|_\phi^p - \|g\|_\phi^p &\leq \|f\|_\phi^p \\ &\leq \frac{2^p}{\left[\phi^{-1}\left(1/u\right)\right]^p} = 2^p \psi_{\phi,p}\left(u\right) \\ &\leq 2^p \psi_{\phi,p}^\sharp(u) \leq p 2^p \psi_{\phi,p}^\sharp(u).\end{aligned}$$

Case 2. We now assume

$$\|f\|_\phi > \frac{2}{\phi^{-1}(1/u)}.$$

Applying Lemma 14 to f, there exists a function y such that

$$\int_0^1 f(s)y(s)\mathrm{d}s = \|f\|_\phi$$

and

$$\int_0^1 g(s)y(s)\mathrm{d}s \leq \|g\|_\phi.$$

Therefore,

$$\begin{aligned}\|f\|_\phi - \|g\|_\phi &\leq \int_0^1 f(s)y(s)\mathrm{d}s - \int_0^1 g(s)y(s)\mathrm{d}s \\ &= \int_0^1 \left(f(s) - g(s)\right) y(s)\mathrm{d}s.\end{aligned}$$

Since $0 \leq \|g\|_\phi \leq \|f\|_\phi$, the inequality $a^p - b^p \leq p a^{p-1}(a-b)$, for all $0 \leq b \leq a$ gives

$$\|f\|_\phi^p - \|g\|_\phi^p \leq p\|f\|_\phi^{p-1} \int_0^1 \left(f(s) - g(s)\right) y(s)\mathrm{d}s. \tag{20}$$

Pointwise estimate. We now claim that

$$\|f\|_\phi^{p-1} y(s) \leq 2^p \frac{\psi_{\phi,p}^\sharp(u)}{u}, \quad \forall s \in [0,1] \text{ such that } f(s) > 0.$$

By Lemma 14,

$$y(s) = \frac{\phi'_+\left(f(s)/\|f\|_\phi\right)}{D_0}, \quad D_0 \geq 1.$$

Since ϕ is increasing, $\phi'_+ \geq 0$. Hence, $D_0 \geq 1$ implies

$$y(s) \leq \phi'_+\left(\frac{f(s)}{\|f\|_\phi}\right).$$

Therefore, it is enough to show that

$$\|f\|_\phi^{p-1} \phi'_+\left(\frac{f(s)}{\|f\|_\phi}\right) \leq 2^p \frac{\psi_{\phi,p}^\sharp(u)}{u}.$$

Fix $s \in [0,1]$ such that $f(s) > 0$. From equation 18, we deduce

$$\phi'_+\left(\frac{f(s)}{\|f\|_\phi}\right) \leq \frac{\phi\left(2f(s)/\|f\|_\phi\right) - \phi\left(f(s)/\|f\|_\phi\right)}{f(s)/\|f\|_\phi} \leq \frac{\phi\left(2f(s)/\|f\|_\phi\right)}{f(s)/\|f\|_\phi}. \tag{21}$$

First, suppose that

$$0 < \frac{f(s)}{\|f\|_\phi} \leq \frac{\phi^{-1}(1)}{2}.$$

Since

$$0 \leq \frac{2f(s)}{\|f\|_\phi \phi^{-1}(1)} \leq 1,$$

the convexity of ϕ, together with $\phi(0) = 0$ and $\phi\left(\phi^{-1}(1)\right) = 1$ gives

$$\phi\left(\frac{2f(s)}{\|f\|_\phi}\right) \leq \frac{2f(s)}{\|f\|_\phi \phi^{-1}(1)}.$$

Therefore,

$$\phi'_+\left(\frac{f(s)}{\|f\|_\phi}\right) \leq \frac{2}{\phi^{-1}(1)}.$$

Moreover, since $0 \leq f \leq 1$,

$$\|f\|_\phi \leq \|1\|_\phi = 1/\phi^{-1}(1).$$

Hence

$$\|f\|_\phi^{p-1} \phi'_+\left(\frac{f(s)}{\|f\|_\phi}\right) \leq \frac{2}{[\phi^{-1}(1)]^p}.$$

By Lemma 15, we have

$$\frac{\psi^\sharp_{\phi,p}(u)}{u} = \sup_{\phi^{-1}(1) \leq t \leq \phi^{-1}(1/u)} \frac{\phi(t)}{t^p} \geq \frac{1}{[\phi^{-1}(1)]^p}.$$

Thus, in this case

$$\|f\|_\phi^{p-1} \phi'_+\left(\frac{f(s)}{\|f\|_\phi}\right) \leq 2\frac{\psi^\sharp_{\phi,p}(u)}{u} \leq 2^p \frac{\psi^\sharp_{\phi,p}(u)}{u}.$$

It remains to consider the case where

$$\frac{f(s)}{\|f\|_\phi} > \frac{\phi^{-1}(1)}{2}.$$

Since we are in the case where

$$\|f\|_\phi > \frac{2}{\phi^{-1}(1/u)},$$

and $f(s) \leq 1$, we have

$$\phi^{-1}(1) < \frac{2f(s)}{\|f\|_\phi} < \phi^{-1}(1/u).$$

Hence, using Lemma 15,

$$\phi\left(\frac{2f(s)}{\|f\|_\phi}\right) \leq \frac{\psi^\sharp_{\phi,p}(u)}{u}\left(\frac{2f(s)}{\|f\|_\phi}\right)^p.$$

Therefore, from equation 21, we have

$$\phi'_+\left(\frac{f(s)}{\|f\|_\phi}\right) \leq \frac{\phi\left(2f(s)/\|f\|_\phi\right)}{f(s)/\|f\|_\phi} \leq 2^p \frac{\psi^\sharp_{\phi,p}(u)}{u}\left(\frac{f(s)}{\|f\|_\phi}\right)^{p-1}.$$

Consequently,

$$\|f\|_\phi^{p-1} \phi'_+\left(\frac{f(s)}{\|f\|_\phi}\right) \leq 2^p \frac{\psi^\sharp_{\phi,p}(u)}{u} f(s)^{p-1} \leq 2^p \frac{\psi^\sharp_{\phi,p}(u)}{u},$$

where the last inequality comes from $0 \leq f(s) \leq 1$ and $p \geq 1$.

Thus, for every $s \in [0,1]$ such that $f(s) > 0$, we have

$$\|f\|_\phi^{p-1} y(s) \leq 2^p \frac{\psi^\sharp_{\phi,p}(u)}{u}.$$

Since $0 \le g \le f$, we have $f - g = 0$ on the set $\{f = 0\}$. Substituting the pointwise bound into equation 20 yields

$$\begin{aligned}
\|f\|_\phi^p - \|g\|_\phi^p &\le p\|f\|_\phi^{p-1} \int_0^1 (f(s) - g(s))\, y(s) \mathrm{d}s \\
&= p\|f\|_\phi^{p-1} \int_{\{s\in[0,1]:\, f(s)>0\}} (f(s) - g(s))\, y(s) \mathrm{d}s \\
&\le p2^p \frac{\psi_{\phi,p}^\sharp(u)}{u} \int_{\{s\in[0,1]:\, f(s)>0\}} (f(s) - g(s))\, \mathrm{d}s \\
&= p2^p \frac{\psi_{\phi,p}^\sharp(u)}{u} \int_0^1 (f(s) - g(s))\, \mathrm{d}s \\
&= p2^p \frac{\psi_{\phi,p}^\sharp(u)}{u} \|f - g\|_{L^1(0,1)} \\
&\le p2^p \psi_{\phi,p}^\sharp(u).
\end{aligned}$$

Taking the supremum over all admissible pairs gives

$$\Omega_{\phi,p}(u) \le p2^p \psi_{\phi,p}^\sharp(u).$$

Together with the lower bound proved above, this establishes the first assertion. The second assertion follows by combining this result with equation 17. □

With the estimate for $\Omega_{\phi,p}$ established, we are now ready to prove Theorem 3.

Proof of Theorem 3. We have

$$\begin{aligned}
\mathbb{E}\left|\mathrm{SOW}_{\phi,p}^p(\widehat{\mu}_n, \widehat{\nu}_m) - \mathrm{SOW}_{\phi,p}^p(\mu,\nu)\right| &= \mathbb{E}\left|\int_{\mathbb{S}^{d-1}} \mathrm{W}_\phi^p\left(\Pi_\#^\theta \widehat{\mu}_n, \Pi_\#^\theta \widehat{\nu}_m\right) - \mathrm{W}_\phi^p\left(\Pi_\#^\theta \mu, \Pi_\#^\theta \nu\right) \mathrm{d}\sigma(\theta)\right| \\
&\le \int_{\mathbb{S}^{d-1}} \mathbb{E}\left|\mathrm{W}_\phi^p\left(\Pi_\#^\theta \widehat{\mu}_n, \Pi_\#^\theta \widehat{\nu}_m\right) - \mathrm{W}_\phi^p\left(\Pi_\#^\theta \mu, \Pi_\#^\theta \nu\right)\right| \mathrm{d}\sigma(\theta),
\end{aligned}$$

where the inequality follows from the triangle inequality and Tonelli's theorem.

Fix $\theta \in \mathbb{S}^{d-1}$. We now bound

$$\left|\mathrm{W}_\phi^p\left(\Pi_\#^\theta \widehat{\mu}_n, \Pi_\#^\theta \widehat{\nu}_m\right) - \mathrm{W}_\phi^p\left(\Pi_\#^\theta \mu, \Pi_\#^\theta \nu\right)\right|.$$

For a probability measure α on $\mathbb{R}$, let F_α denote its cumulative distribution function and let F_α^{-1} denote its quantile function. With this convention, define

$$\widehat{h}_\theta(s) := \frac{\left|F^{-1}_{\Pi_\#^\theta \widehat{\mu}_n}(s) - F^{-1}_{\Pi_\#^\theta \widehat{\nu}_m}(s)\right|}{D}, \quad s \in (0,1),$$

and

$$h_\theta(s) := \frac{\left|F^{-1}_{\Pi_\#^\theta \mu}(s) - F^{-1}_{\Pi_\#^\theta \nu}(s)\right|}{D}, \quad s \in (0,1).$$

By the one-dimensional quantile representation and the homogeneity of the Luxemburg norm, we have

$$\left|\mathrm{W}_\phi^p\left(\Pi_\#^\theta \widehat{\mu}_n, \Pi_\#^\theta \widehat{\nu}_m\right) - \mathrm{W}_\phi^p\left(\Pi_\#^\theta \mu, \Pi_\#^\theta \nu\right)\right| = D^p \left|\|\widehat{h}_\theta\|_\phi^p - \|h_\theta\|_\phi^p\right|.$$

Since μ and ν are supported on K, the empirical measures $\widehat{\mu}_n$ and $\widehat{\nu}_m$ are also supported on K. Since $\theta \in \mathbb{S}^{d-1}$, the projected supports have diameter at most D. Hence,

$$0 \le \widehat{h}_\theta(s), h_\theta(s) \le 1, \quad s \in (0,1).$$

As a result,

$$0 \le \|\widehat{h}_\theta - h_\theta\|_{L^1(0,1)} \le 1.$$

By the definition of $\Omega_{\phi,p}$ in equation 19, we have

$$\left| \|\widehat{h}_\theta\|_\phi^p - \|h_\theta\|_\phi^p \right| \le \Omega_{\phi,p}\left(\|\widehat{h}_\theta - h_\theta\|_{L^1(0,1)} \right).$$

Applying Proposition 6, we have

$$\Omega_{\phi,p}\left(\|\widehat{h}_\theta - h_\theta\|_{L^1(0,1)} \right) \le p2^p \psi_{\phi,p}^{\mathrm{cav}}\left(\|\widehat{h}_\theta - h_\theta\|_{L^1(0,1)} \right).$$

As a consequence,

$$\mathbb{E}\left| \mathrm{W}_\phi^p\left(\Pi_\#^\theta \widehat{\mu}_n, \Pi_\#^\theta \widehat{\nu}_m\right) - \mathrm{W}_\phi^p\left(\Pi_\#^\theta \mu, \Pi_\#^\theta \nu\right) \right| \le p2^p D^p \mathbb{E}\left[\psi_{\phi,p}^{\mathrm{cav}}\left(\|\widehat{h}_\theta - h_\theta\|_{L^1(0,1)} \right) \right].$$

Since $\psi_{\phi,p}^{\mathrm{cav}}$ is concave, Jensen's inequality yields

$$\mathbb{E}\left[\psi_{\phi,p}^{\mathrm{cav}}\left(\|\widehat{h}_\theta - h_\theta\|_{L^1(0,1)} \right) \right] \le \psi_{\phi,p}^{\mathrm{cav}}\left(\mathbb{E}\|\widehat{h}_\theta - h_\theta\|_{L^1(0,1)} \right).$$

We now bound $\mathbb{E}\|\widehat{h}_\theta - h_\theta\|_{L^1(0,1)}$. By the triangle inequality,

$$\begin{aligned}
\|\widehat{h}_\theta - h_\theta\|_{L^1(0,1)} &= \frac{1}{D}\int_0^1 \left| \left| F^{-1}_{\Pi_\#^\theta \widehat{\mu}_n}(s) - F^{-1}_{\Pi_\#^\theta \widehat{\nu}_m}(s) \right| - \left| F^{-1}_{\Pi_\#^\theta \mu}(s) - F^{-1}_{\Pi_\#^\theta \nu}(s) \right| \right| \mathrm{d}s \\
&\le \frac{1}{D}\int_0^1 \left| F^{-1}_{\Pi_\#^\theta \widehat{\mu}_n}(s) - F^{-1}_{\Pi_\#^\theta \mu}(s) \right| \mathrm{d}s + \frac{1}{D}\int_0^1 \left| F^{-1}_{\Pi_\#^\theta \widehat{\nu}_m}(s) - F^{-1}_{\Pi_\#^\theta \nu}(s) \right| \mathrm{d}s \\
&= \frac{1}{D}\left[\mathrm{W}_1\left(\Pi_\#^\theta \widehat{\mu}_n, \Pi_\#^\theta \mu\right) + \mathrm{W}_1\left(\Pi_\#^\theta \widehat{\nu}_m, \Pi_\#^\theta \nu\right) \right],
\end{aligned}$$

where the last equality follows from the one-dimensional quantile representation of the Wasserstein distance (see, e.g., (Bobkov & Ledoux, 2019, Theorem 2.10)).

Taking expectations gives

$$\mathbb{E}\left[\|\widehat{h}_\theta - h_\theta\|_{L^1(0,1)} \right] \le \frac{1}{D}\left[\mathbb{E}\,\mathrm{W}_1\left(\Pi_\#^\theta \widehat{\mu}_n, \Pi_\#^\theta \mu\right) + \mathbb{E}\,\mathrm{W}_1\left(\Pi_\#^\theta \widehat{\nu}_m, \Pi_\#^\theta \nu\right) \right].$$

Applying Bobkov & Ledoux (2019, Theorem 3.2), we have

$$\mathbb{E}\,\mathrm{W}_1\left(\Pi_\#^\theta \widehat{\mu}_n, \Pi_\#^\theta \mu\right) \le \frac{1}{\sqrt{n}}\int_{\mathbb{R}} \sqrt{F_{\Pi_\#^\theta \mu}(x)\left(1 - F_{\Pi_\#^\theta \mu}(x)\right)}\mathrm{d}x \le \frac{D}{2\sqrt{n}},$$

and similarly,

$$\mathbb{E}\,\mathrm{W}_1\left(\Pi_\#^\theta \widehat{\nu}_m, \Pi_\#^\theta \nu\right) \le \frac{D}{2\sqrt{m}}.$$

Consequently,

$$\mathbb{E}\left[\|\widehat{h}_\theta - h_\theta\|_{L^1(0,1)} \right] \le \frac{1}{2}\left(n^{-1/2} + m^{-1/2} \right).$$

Since $\psi_{\phi,p}^{\mathrm{cav}}$ is increasing by Lemma 3, we obtain

$$\mathbb{E}\left| \mathrm{W}_\phi^p\left(\Pi_\#^\theta \widehat{\mu}_n, \Pi_\#^\theta \widehat{\nu}_m\right) - \mathrm{W}_\phi^p\left(\Pi_\#^\theta \mu, \Pi_\#^\theta \nu\right) \right| \le p2^p D^p \psi_{\phi,p}^{\mathrm{cav}}\left(\frac{1}{2}n^{-1/2} + \frac{1}{2}m^{-1/2} \right).$$

Integrating the above bound over $\theta \in \mathbb{S}^{d-1}$ and using that σ is a probability measure yields the desired result. □

F Proofs for Minimax Optimality

F.1 Le Cam Two-Point Bounds

The minimax lower bounds are based on Le Cam's two-point method (LeCam, 1973; Le Cam, 2012). We first state the one-sample version, followed by a two-sample version for real-valued functionals.

Throughout this appendix, for $\mu, \nu \in \mathcal{P}(K)$, we write $\mathbb{P}_\mu$ and $\mathbb{E}_\mu$ for probability and expectation with respect to $\mu^{\otimes n}$. Likewise, $\mathbb{P}_{\mu,\nu}$ and $\mathbb{E}_{\mu,\nu}$ denote probability and expectation with respect to $\mu^{\otimes n} \otimes \nu^{\otimes m}$.

Lemma 17 (One-sample two-point bound)**.** *Let $\mu_0, \mu_1 \in \mathcal{P}(K)$. Then*

$$\inf_{\widetilde{\mu}_n} \sup_{\mu \in \mathcal{P}(K)} \left[\mathbb{E}_\mu \operatorname{SOW}_{\phi,p}^p (\widetilde{\mu}_n, \mu)\right]^{1/p} \geq \frac{\operatorname{SOW}_{\phi,p}(\mu_0, \mu_1)}{2^{1+1/p}} \left[1 - \sqrt{\frac{n \operatorname{KL}(\mu_0 \| \mu_1)}{2}}\right].$$

Proof. Let $\widetilde{\mu}_n$ be any estimator based on $X_1, \ldots, X_n$. Since $\mu_0, \mu_1 \in \mathcal{P}(K)$, we have

$$\sup_{\mu \in \mathcal{P}(K)} \left[\mathbb{E}_\mu \operatorname{SOW}_{\phi,p}^p (\widetilde{\mu}_n, \mu)\right]^{1/p} \geq \max_{i \in \{0,1\}} \left[\mathbb{E}_{\mu_i} \operatorname{SOW}_{\phi,p}^p (\widetilde{\mu}_n, \mu_i)\right]^{1/p}.$$

Define the test

$$\Psi(\widetilde{\mu}_n) := \begin{cases} 0, & \text{if } \operatorname{SOW}_{\phi,p}(\widetilde{\mu}_n, \mu_0) \leq \operatorname{SOW}_{\phi,p}(\widetilde{\mu}_n, \mu_1), \\ 1, & \text{if } \operatorname{SOW}_{\phi,p}(\widetilde{\mu}_n, \mu_0) > \operatorname{SOW}_{\phi,p}(\widetilde{\mu}_n, \mu_1). \end{cases}$$

Consider first the event $\{\Psi(\widetilde{\mu}_n) = 1\}$. By the definition of $\Psi(\widetilde{\mu}_n)$,

$$\operatorname{SOW}_{\phi,p}(\widetilde{\mu}_n, \mu_0) > \operatorname{SOW}_{\phi,p}(\widetilde{\mu}_n, \mu_1).$$

Hence, by the triangle inequality for $\operatorname{SOW}_{\phi,p}$,

$$\begin{aligned} \operatorname{SOW}_{\phi,p}(\mu_0, \mu_1) &\leq \operatorname{SOW}_{\phi,p}(\mu_0, \widetilde{\mu}_n) + \operatorname{SOW}_{\phi,p}(\widetilde{\mu}_n, \mu_1) \\ &< 2 \operatorname{SOW}_{\phi,p}(\widetilde{\mu}_n, \mu_0). \end{aligned}$$

Therefore, on the event $\{\Psi(\widetilde{\mu}_n) = 1\}$,

$$\operatorname{SOW}_{\phi,p}(\widetilde{\mu}_n, \mu_0) \geq \frac{\operatorname{SOW}_{\phi,p}(\mu_0, \mu_1)}{2}.$$

Raising both sides to the power p gives

$$\operatorname{SOW}_{\phi,p}^p (\widetilde{\mu}_n, \mu_0) \geq \left(\frac{\operatorname{SOW}_{\phi,p}(\mu_0, \mu_1)}{2}\right)^p \qquad \text{on } \{\Psi(\widetilde{\mu}_n) = 1\}.$$

It follows that

$$\operatorname{SOW}_{\phi,p}^p (\widetilde{\mu}_n, \mu_0) \geq \left(\frac{\operatorname{SOW}_{\phi,p}(\mu_0, \mu_1)}{2}\right)^p \mathbb{1}_{\{\Psi(\widetilde{\mu}_n) = 1\}}.$$

Taking expectation under $\mu_0^{\otimes n}$ yields

$$\mathbb{E}_{\mu_0} \operatorname{SOW}_{\phi,p}^p (\widetilde{\mu}_n, \mu_0) \geq \left(\frac{\operatorname{SOW}_{\phi,p}(\mu_0, \mu_1)}{2}\right)^p \mathbb{P}_{\mu_0}(\Psi(\widetilde{\mu}_n) = 1).$$

Taking the p-th root, we obtain

$$\left[\mathbb{E}_{\mu_0} \operatorname{SOW}_{\phi,p}^p (\widetilde{\mu}_n, \mu_0)\right]^{1/p} \geq \frac{\operatorname{SOW}_{\phi,p}(\mu_0, \mu_1)}{2} \mathbb{P}_{\mu_0}(\Psi(\widetilde{\mu}_n) = 1)^{1/p}.$$

Similarly, we can show that

$$\left[\mathbb{E}_{\mu_1} \operatorname{SOW}_{\phi,p}^p (\widetilde{\mu}_n, \mu_1)\right]^{1/p} \geq \frac{\operatorname{SOW}_{\phi,p}(\mu_0, \mu_1)}{2} \mathbb{P}_{\mu_1}(\Psi(\widetilde{\mu}_n) = 0)^{1/p}.$$

Combining these estimates with the initial bound and using $\max\{a^{1/p}, b^{1/p}\} \geq \left(\frac{a+b}{2}\right)^{1/p}$ for $a, b \geq 0$, we obtain

$$\sup_{\mu \in \mathcal{P}(K)} \left[\mathbb{E}_\mu \operatorname{SOW}_{\phi,p}^p (\widetilde{\mu}_n, \mu)\right]^{1/p} \geq \frac{\operatorname{SOW}_{\phi,p}(\mu_0, \mu_1)}{2^{1+1/p}} \left[\mathbb{P}_{\mu_0}(\Psi(\widetilde{\mu}_n) = 1) + \mathbb{P}_{\mu_1}(\Psi(\widetilde{\mu}_n) = 0)\right]^{1/p}.$$

By the definition of total variation (Tsybakov, 2009, Definition 2.4), we have

$$\sup_{\mu \in \mathcal{P}(K)} \left[\mathbb{E}_\mu \operatorname{SOW}_{\phi,p}^p (\widetilde{\mu}_n, \mu)\right]^{1/p} \geq \frac{\operatorname{SOW}_{\phi,p}(\mu_0, \mu_1)}{2^{1+1/p}} \left[1 - \operatorname{TV}\left(\mu_0^{\otimes n}, \mu_1^{\otimes n}\right)\right]^{1/p}.$$

Since $0 \le 1 - \mathrm{TV}\left(\mu_0^{\otimes n}, \mu_1^{\otimes n}\right) \le 1$ and $p \ge 1$, we have

$$\sup_{\mu \in \mathcal{P}(K)} \left[\mathbb{E}_\mu \, \mathrm{SOW}_{\phi,p}^p \left(\widetilde{\mu}_n, \mu\right)\right]^{1/p} \ge \frac{\mathrm{SOW}_{\phi,p}(\mu_0, \mu_1)}{2^{1+1/p}} \left[1 - \mathrm{TV}\left(\mu_0^{\otimes n}, \mu_1^{\otimes n}\right)\right].$$

By Pinsker's inequality (Tsybakov, 2009, Lemma 2.5),

$$\mathrm{TV}\left(\mu_0^{\otimes n}, \mu_1^{\otimes n}\right) \le \sqrt{\frac{1}{2} \mathrm{KL}\left(\mu_0^{\otimes n} \| \mu_1^{\otimes n}\right)}.$$

Using the additivity of the Kullback–Leibler divergence over product measures, we have

$$\sup_{\mu \in \mathcal{P}(K)} \left[\mathbb{E}_\mu \, \mathrm{SOW}_{\phi,p}^p \left(\widetilde{\mu}_n, \mu\right)\right]^{1/p} \ge \frac{\mathrm{SOW}_{\phi,p}(\mu_0, \mu_1)}{2^{1+1/p}} \left[1 - \sqrt{\frac{n \, \mathrm{KL}(\mu_0 \| \mu_1)}{2}}\right].$$

Taking the infimum over all estimators $\widetilde{\mu}_n$ completes the proof. □

We next state the corresponding two-sample bound for real-valued functionals.

Lemma 18 (Two-sample two-point bound)**.** *Let $(\mu_0, \nu_0), (\mu_1, \nu_1) \in \mathcal{P}(K) \times \mathcal{P}(K)$, and let*

$$T : \mathcal{P}(K) \times \mathcal{P}(K) \longrightarrow \mathbb{R}$$

be a real-valued functional. Define

$$T_0 \coloneqq T(\mu_0, \nu_0), \quad T_1 \coloneqq T(\mu_1, \nu_1).$$

Then

$$\inf_{\widehat{T}} \sup_{\mu, \nu \in \mathcal{P}(K)} \mathbb{E}_{\mu,\nu} \left|\widehat{T} - T(\mu, \nu)\right| \ge \frac{|T_1 - T_0|}{4} \left[1 - \sqrt{\frac{n \, \mathrm{KL}(\mu_0 \| \mu_1) + m \, \mathrm{KL}(\nu_0 \| \nu_1)}{2}}\right].$$

Proof. Let $\widehat{T} = \widehat{T}(X_1, \ldots, X_n, Y_1, \ldots, Y_m)$ be any estimator based on $X_1, \ldots, X_n, Y_1, \ldots, Y_m$. We have

$$\begin{aligned}\sup_{\mu, \nu \in \mathcal{P}(K)} \mathbb{E}_{\mu,\nu} \left|\widehat{T} - T(\mu, \nu)\right| &\ge \max\left\{\mathbb{E}_{\mu_0,\nu_0} |\widehat{T} - T_0|, \mathbb{E}_{\mu_1,\nu_1} |\widehat{T} - T_1|\right\} \\ &\ge \frac{1}{2} \mathbb{E}_{\mu_0,\nu_0} |\widehat{T} - T_0| + \frac{1}{2} \mathbb{E}_{\mu_1,\nu_1} |\widehat{T} - T_1|.\end{aligned}$$

Define

$$\Psi(\widehat{T}) \coloneqq \begin{cases} 0, & \text{if } |\widehat{T} - T_0| \le |\widehat{T} - T_1|, \\ 1, & \text{if } |\widehat{T} - T_0| > |\widehat{T} - T_1|. \end{cases}$$

On the event $\{\Psi(\widehat{T}) = 1\}$, we have

$$|\widehat{T} - T_0| > |\widehat{T} - T_1|.$$

Hence, by the triangle inequality,

$$|T_1 - T_0| \le |T_1 - \widehat{T}| + |\widehat{T} - T_0| \le 2|\widehat{T} - T_0|.$$

Therefore,

$$\mathbb{E}_{\mu_0,\nu_0} |\widehat{T} - T_0| \ge \frac{|T_1 - T_0|}{2} \mathbb{P}_{\mu_0,\nu_0} \left(\Psi(\widehat{T}) = 1\right).$$

Similarly,

$$\mathbb{E}_{\mu_1,\nu_1} |\widehat{T} - T_1| \ge \frac{|T_1 - T_0|}{2} \mathbb{P}_{\mu_1,\nu_1} \left(\Psi(\widehat{T}) = 0\right).$$

Consequently,

$$\begin{aligned}\sup_{\mu,\nu\in\mathcal{P}(K)} \mathbb{E}_{\mu,\nu}\left|\widehat{T}-T(\mu,\nu)\right| &\geq \frac{|T_1-T_0|}{4}\left[\mathbb{P}_{\mu_0,\nu_0}\left(\Psi(\widehat{T})=1\right)+\mathbb{P}_{\mu_1,\nu_1}\left(\Psi(\widehat{T})=0\right)\right]\\ &= \frac{|T_1-T_0|}{4}\left[1-\mathbb{P}_{\mu_0,\nu_0}\left(\Psi(\widehat{T})=0\right)+\mathbb{P}_{\mu_1,\nu_1}\left(\Psi(\widehat{T})=0\right)\right]\\ &\geq \frac{|T_1-T_0|}{4}\left[1-\mathrm{TV}\left(\mu_0^{\otimes n}\otimes\nu_0^{\otimes m},\mu_1^{\otimes n}\otimes\nu_1^{\otimes m}\right)\right]\\ &\geq \frac{|T_1-T_0|}{4}\left[1-\sqrt{\frac{1}{2}\mathrm{KL}\left(\mu_0^{\otimes n}\otimes\nu_0^{\otimes m}\|\mu_1^{\otimes n}\otimes\nu_1^{\otimes m}\right)}\right]\\ &= \frac{|T_1-T_0|}{4}\left[1-\sqrt{\frac{n\,\mathrm{KL}(\mu_0\|\mu_1)+m\,\mathrm{KL}(\nu_0\|\nu_1)}{2}}\right].\end{aligned}$$

The second inequality follows from the definition of total variation (Tsybakov, 2009, Definition 2.4), and the third inequality follows from Pinsker's inequality (Tsybakov, 2009, Lemma 2.5). The last equality follows from the additivity of the Kullback–Leibler divergence over product measures (Tsybakov, 2009, Page 85). Taking the infimum over all estimators $\widehat{T}$ completes the proof. □

With this general framework in place, we now construct suitable pairs of distributions and apply the preceding two-point bounds to the one-sample and two-sample SOW estimation problems. Before doing so, we recall the useful spherical identity

$$\int_{\mathbb{S}^{d-1}} |\langle\theta,z\rangle|^p \,\mathrm{d}\sigma(\theta) = \kappa_{p,d}\|z\|^p, \tag{22}$$

where σ denotes the uniform probability measure on $\mathbb{S}^{d-1}$ and $\kappa_{p,d}>0$ is a constant depending only on p and d (see, e.g., (Peyré, 2025, Section 11.2)).

Besides, since K is compact, the continuous map $(x,y)\mapsto\|x-y\|$ attains its maximum on $K\times K$. Hence, there exist $x_0,x_1\in K$ such that

$$\|x_0-x_1\| = \mathrm{diam}(K) = D.$$

Throughout the remainder of this section, we fix such a pair (x_0,x_1) and use it in all subsequent lower-bound constructions.

F.2 One-sample estimation

We first prove the one-sample lower bound. We choose

$$\mu_0 = \frac{1}{2}\delta_{x_0}+\frac{1}{2}\delta_{x_1},\quad \mu_1 = \left(\frac{1}{2}-\frac{1}{4\sqrt{n}}\right)\delta_{x_0}+\left(\frac{1}{2}+\frac{1}{4\sqrt{n}}\right)\delta_{x_1}.$$

For this choice, we have

$$\begin{aligned}\mathrm{SOW}_{\phi,p}^p(\mu_0,\mu_1) &= \int_{\mathbb{S}^{d-1}} \mathrm{W}_\phi^p\left(\Pi_\#^\theta\mu_0,\Pi_\#^\theta\mu_1\right)\mathrm{d}\sigma(\theta)\\ &= \int_{\mathbb{S}^{d-1}}\left[D\left|\left\langle\theta,\frac{x_0-x_1}{\|x_0-x_1\|}\right\rangle\right|\omega_\phi\left(\frac{1}{4\sqrt{n}}\right)\right]^p\mathrm{d}\sigma(\theta)\\ &= D^p\omega_\phi\left(\frac{1}{4\sqrt{n}}\right)^p\kappa_{p,d}.\end{aligned}$$

Therefore,

$$\mathrm{SOW}_{\phi,p}\left(\mu_0,\mu_1\right) = D\omega_\phi\left(\frac{1}{4\sqrt{n}}\right)\kappa_{p,d}^{1/p}.$$

Moreover, applying Lemma 2 with $a=4$ and $u=1/(4\sqrt{n})$ gives

$$\omega_\phi\left(n^{-1/2}\right)\leq 4\omega_\phi\left(\frac{1}{4\sqrt{n}}\right).$$

Hence,

$$\mathrm{SOW}_{\phi,p}\left(\mu_0,\mu_1\right) \geq \frac{\kappa_{p,d}^{1/p} D}{4}\omega_\phi\left(n^{-1/2}\right).$$

On the other hand, we also have

$$\begin{aligned}\mathrm{KL}(\mu_0\|\mu_1) &= \frac{1}{2}\log\frac{1/2}{1/2+1/(4\sqrt{n})} + \frac{1}{2}\log\frac{1/2}{1/2-1/(4\sqrt{n})} \\ &\leq \frac{1}{2}\left(\frac{1/2}{1/2+1/(4\sqrt{n})}-1\right) + \frac{1}{2}\left(\frac{1/2}{1/2-1/(4\sqrt{n})}-1\right) \\ &= \frac{1/(16n)}{1/4-1/(16n)},\end{aligned}$$

where the inequality comes from the fact that $\log t \leq t-1\,\forall t>0$. Since $n\geq 1$,

$$\frac{1}{4}-\frac{1}{16n} \geq \frac{3}{16}.$$

As a result,

$$\mathrm{KL}\left(\mu_0\|\mu_1\right) \leq \frac{1}{3n}.$$

Combining the preceding estimates with Lemma 17, we obtain

$$\inf_{\widetilde{\mu}_n}\sup_{\mu\in\mathcal{P}(K)}\left[\mathbb{E}_\mu\,\mathrm{SOW}_{\phi,p}^p\left(\widetilde{\mu}_n,\mu\right)\right]^{1/p} \geq \frac{\kappa_{p,d}^{1/p}}{2^{3+1/p}}\left(1-\frac{1}{\sqrt{6}}\right)D\omega_\phi\left(n^{-1/2}\right).$$

The matching upper bound follows directly from Theorem 2 by taking the empirical measure $\widehat{\mu}_n$ as the estimator. Combining the upper and lower bounds yields

$$\inf_{\widetilde{\mu}_n}\sup_{\mu\in\mathcal{P}(K)}\left[\mathbb{E}_\mu\,\mathrm{SOW}_{\phi,p}^p\left(\widetilde{\mu}_n,\mu\right)\right]^{1/p} \asymp_{d,p} D\omega_\phi\left(n^{-1/2}\right).$$

F.3 Two-sample estimation

We next prove the two-sample statement in Theorem 4. Without loss of generality, assume that $n\leq m$.

F.3.1 Two-sample estimation of the SOW distance

We choose

$$\mu_0 = \frac{1}{2}\delta_{x_0}+\frac{1}{2}\delta_{x_1},\quad \nu_0 = \frac{1}{2}\delta_{x_0}+\frac{1}{2}\delta_{x_1},$$

and

$$\mu_1 = \left(\frac{1}{2}-\frac{1}{4\sqrt{n}}\right)\delta_{x_0}+\left(\frac{1}{2}+\frac{1}{4\sqrt{n}}\right)\delta_{x_1},\quad \nu_1=\nu_0.$$

As in the one-sample case, our construction gives

$$\mathrm{SOW}_{\phi,p}\left(\mu_0,\nu_0\right)=0,\quad \mathrm{SOW}_{\phi,p}(\mu_1,\nu_1) = D\kappa_{p,d}^{1/p}\omega_\phi\left(\frac{1}{4\sqrt{n}}\right) \geq \frac{D\kappa_{p,d}^{1/p}}{4}\omega_\phi\left(n^{-1/2}\right).$$

For the KL term, since $\nu_0=\nu_1$,

$$\mathrm{KL}\left(\nu_0\|\nu_1\right)=0.$$

The one-sample calculation also gives

$$\mathrm{KL}\left(\mu_0\|\mu_1\right) \leq \frac{1}{3n}.$$

Therefore,

$$n\,\mathrm{KL}\left(\mu_0\|\mu_1\right)+m\,\mathrm{KL}\left(\nu_0\|\nu_1\right) \leq n\frac{1}{3n}+0=\frac{1}{3}.$$

Applying Lemma 18 to the functional $T(\mu,\nu) = \mathrm{SOW}_{\phi,p}(\mu,\nu)$, we obtain

$$\inf_{\widehat{T}} \sup_{\mu,\nu\in\mathcal{P}(K)} \mathbb{E}_{\mu,\nu}\left|\widehat{T} - \mathrm{SOW}_{\phi,p}(\mu,\nu)\right| \geq \frac{\kappa_{p,d}^{1/p}}{16}\left(1-\frac{1}{\sqrt{6}}\right) D\omega_\phi\left(n^{-1/2}\right).$$

Since $n \leq m$ and ω_ϕ is increasing by Lemma 2,

$$\omega_\phi\left(n^{-1/2}\right) \geq \omega_\phi\left(m^{-1/2}\right).$$

Therefore,

$$\omega_\phi\left(n^{-1/2}\right) \geq \frac{1}{2}\left[\omega_\phi\left(n^{-1/2}\right) + \omega_\phi\left(m^{-1/2}\right)\right].$$

Consequently,

$$\inf_{\widehat{T}} \sup_{\mu,\nu\in\mathcal{P}(K)} \mathbb{E}_{\mu,\nu}\left|\widehat{T} - \mathrm{SOW}_{\phi,p}(\mu,\nu)\right| \geq \frac{\kappa_{p,d}^{1/p}}{32}\left(1-\frac{1}{\sqrt{6}}\right) D\left[\omega_\phi\left(n^{-1/2}\right) + \omega_\phi\left(m^{-1/2}\right)\right].$$

The matching upper bound follows from Corollary 2. Combining the upper and lower bounds yields

$$\inf_{\widehat{T}} \sup_{\mu,\nu\in\mathcal{P}(K)} \mathbb{E}_{\mu,\nu}\left|\widehat{T} - \mathrm{SOW}_{\phi,p}(\mu,\nu)\right| \asymp_{d,p} D\left[\omega_\phi\left(n^{-1/2}\right) + \omega_\phi\left(m^{-1/2}\right)\right].$$

F.3.2 Two-sample estimation of the powered SOW functional

The construction is more delicate for the powered functional. In contrast to the SOW distance considered above, a direct perturbation at the statistical scale $n^{-1/2}$ only yields $\psi_{\phi,p}\left(n^{-1/2}\right)$, whereas the sharp rate is controlled by $\psi_{\phi,p}^{\mathrm{cav}}\left(n^{-1/2}\right)$. Bridging this gap is the main technical difficulty in the lower-bound argument. We resolve it through the following useful localization result.

Lemma 19. *For every $h \in (0, 1/4]$, there exist $0 \leq q_0 < q_1 \leq \frac{1}{2}$ such that $q_1 - q_0 \leq h$ and*

$$\psi_{\phi,p}(q_1) - \psi_{\phi,p}(q_0) \geq 2^{-(p+2)}\psi_{\phi,p}^{\mathrm{cav}}(h).$$

Proof. The main idea is to find $v^* \in [h, 1/2]$ and a constant $C_p > 0$ depending only on p such that

$$h\frac{\psi_{\phi,p}(v^*)}{v^*} \geq C_p\psi_{\phi,p}^{\mathrm{cav}}(h).$$

Since $\psi_{\phi,p}(0) = 0$, the ratio $\psi_{\phi,p}(v^*)/v^*$ represents the average increase of $\psi_{\phi,p}$ over the interval $[0, v^*]$. We then partition $[0, v^*]$ into subintervals of length at most h. Since the sum of the increments of $\psi_{\phi,p}$ over these subintervals is exactly $\psi_{\phi,p}(v^*)$, at least one of them must have an increment comparable to this average. This yields the desired pair q_0, q_1.

To carry out the first step, we fix $h \in (0, 1/4]$. We then define

$$S_h \coloneqq \sup_{h\leq v\leq 1} \frac{\psi_{\phi,p}(v)}{v}, \quad S_{h,1/2} \coloneqq \sup_{h\leq v\leq 1/2} \frac{\psi_{\phi,p}(v)}{v}.$$

We first compare these two suprema. Since $h \in (0, 1/4]$,

$$S_h = \max\left\{S_{h,1/2}, \sup_{1/2\leq v\leq 1} \frac{\psi_{\phi,p}(v)}{v}\right\}.$$

It remains to control the contribution from $v \in [1/2, 1]$. By the definition of $S_{h,1/2}$, we have

$$S_{h,1/2} \geq 2\psi_{\phi,p}(1/2) = \frac{2}{[\phi^{-1}(2)]^p}.$$

Applying Lemma 2 (with $a = 2$ and $u = 1/2$) gives

$$\phi^{-1}(2) \leq 2\phi^{-1}(1).$$

Consequently,

$$S_{h,1/2} \geq \frac{2^{1-p}}{[\phi^{-1}(1)]^p}.$$

On the other hand, for $v \in [1/2, 1]$, since $\psi_{\phi,p}$ is increasing by Lemma 2,

$$\frac{\psi_{\phi,p}(v)}{v} \leq 2\psi_{\phi,p}(1) = \frac{2}{[\phi^{-1}(1)]^p}.$$

Hence,

$$\sup_{v\in[1/2,1]} \frac{\psi_{\phi,p}(v)}{v} \leq \frac{2}{[\phi^{-1}(1)]^p} \leq 2^p S_{h,1/2}.$$

As a result, $S_h \leq 2^p S_{h,1/2}$.

By Lemma 2, $\psi_{\phi,p}$ is continuous on $[h, 1/2]$. Hence, the map $v \mapsto \psi_{\phi,p}(v)/v$ is continuous on $[h, 1/2]$. Therefore, there exists $v^* \in [h, 1/2]$ such that

$$\frac{\psi_{\phi,p}(v^*)}{v^*} = S_{h,1/2}.$$

Since $hS_h = \psi^\sharp_{\phi,p}(h)$, combining $S_h \leq 2^p S_{h,1/2}$ with equation 17 gives

$$h\frac{\psi_{\phi,p}(v^*)}{v^*} = hS_{h,1/2} \geq 2^{-p} hS_h \geq 2^{-(p+1)}\psi^{\mathrm{cav}}_{\phi,p}(h).$$

We now partition $[0, v^*]$ into

$$k \coloneqq \left\lceil \frac{v^*}{h} \right\rceil$$

equal subintervals with endpoints

$$t_i \coloneqq \frac{iv^*}{k}, \qquad i = 0, \ldots, k.$$

Then $t_i - t_{i-1} = v^*/k \leq h$. Moreover,

$$\sum_{i=1}^{k} [\psi_{\phi,p}(t_i) - \psi_{\phi,p}(t_{i-1})] = \psi_{\phi,p}(v^*),$$

since $\psi_{\phi,p}(0) = 0$. Hence, for some $i \in \{1, \ldots, k\}$,

$$\psi_{\phi,p}(t_i) - \psi_{\phi,p}(t_{i-1}) \geq \frac{\psi_{\phi,p}(v^*)}{k}.$$

Since $v^* \geq h$ and $k \leq v^*/h + 1$, we deduce

$$\psi_{\phi,p}(t_i) - \psi_{\phi,p}(t_{i-1}) \geq \frac{h}{2v^*}\psi_{\phi,p}(v^*) \geq 2^{-(p+2)}\psi^{\mathrm{cav}}_{\phi,p}(h).$$

Setting $q_0 = t_{i-1}$ and $q_1 = t_i$ gives

$$0 \leq q_0 < q_1 \leq \frac{1}{2}, \quad q_1 - q_0 \leq h.$$

□

With this lemma in hand, we are now ready to prove the lower bound for the powered SOW functional in Theorem 4.

Proof. Assume without loss of generality that $n \leq m$. Applying Lemma 19 with $h = \frac{1}{4\sqrt{n}}$, there exist $0 \leq q_0 < q_1 \leq 1/2$ such that

$$q_1 - q_0 \leq \frac{1}{4\sqrt{n}}$$

and

$$\psi_{\phi,p}(q_1) - \psi_{\phi,p}(q_0) \geq 2^{-(p+2)} \psi_{\phi,p}^{\mathrm{cav}}\left(\frac{1}{4\sqrt{n}}\right).$$

We then choose

$$a_i = \frac{1}{4} + q_i, \qquad i = 0, 1,$$

and define

$$\mu_i = (1 - a_i)\delta_{x_0} + a_i \delta_{x_1}, \quad \nu_0 = \nu_1 = \frac{3}{4}\delta_{x_0} + \frac{1}{4}\delta_{x_1}.$$

This choice gives $\mathrm{KL}(\nu_0 \| \nu_1) = 0$. Besides, since $a_0, a_1 \in [1/4, 3/4]$, using $\log t \leq t - 1$ for all $t > 0$ gives

$$\begin{aligned}
\mathrm{KL}(\mu_0 \| \mu_1) &= a_0 \log \frac{a_0}{a_1} + (1 - a_0) \log \frac{1 - a_0}{1 - a_1} \\
&\leq \frac{(a_1 - a_0)^2}{a_1(1 - a_1)} \\
&\leq \frac{16}{3}(a_1 - a_0)^2 \\
&\leq \frac{1}{3n}.
\end{aligned}$$

Hence,

$$n\,\mathrm{KL}(\mu_0 \| \mu_1) + m\,\mathrm{KL}(\nu_0 \| \nu_1) \leq \frac{1}{3}.$$

Similarly, we also have

$$\mathrm{SOW}_{\phi,p}^p(\mu_i, \nu_i) = \kappa_{p,d} D^p \psi_{\phi,p}(q_i), \quad i = 0, 1.$$

Combining with Lemma 19, we obtain

$$\begin{aligned}
\left|\mathrm{SOW}_{\phi,p}^p(\mu_1, \nu_1) - \mathrm{SOW}_{\phi,p}^p(\mu_0, \nu_0)\right| &= \kappa_{p,d} D^p \left[\psi_{\phi,p}(q_1) - \psi_{\phi,p}(q_0)\right] \\
&\geq 2^{-(p+2)} \kappa_{p,d} D^p \psi_{\phi,p}^{\mathrm{cav}}\left(\frac{1}{4\sqrt{n}}\right).
\end{aligned}$$

On the other hand, since $n \leq m$

$$0 < \frac{1}{4}\left(\frac{1}{2}n^{-1/2} + \frac{1}{2}m^{-1/2}\right) \leq \frac{1}{4\sqrt{n}} \leq \frac{1}{4}.$$

By Lemma 3, $\psi_{\phi,p}^{\mathrm{cav}}$ is increasing on $[0, 1]$ and satisfies $\psi_{\phi,p}^{\mathrm{cav}}(0) = 0$. Therefore, by its concavity,

$$\psi_{\phi,p}^{\mathrm{cav}}\left(\frac{1}{4\sqrt{n}}\right) \geq \frac{1}{4}\psi_{\phi,p}^{\mathrm{cav}}\left(\frac{1}{2}n^{-1/2} + \frac{1}{2}m^{-1/2}\right).$$

Consequently,

$$\left|\mathrm{SOW}_{\phi,p}^p(\mu_1, \nu_1) - \mathrm{SOW}_{\phi,p}^p(\mu_0, \nu_0)\right| \geq 2^{-(p+4)} \kappa_{p,d} D^p \psi_{\phi,p}^{\mathrm{cav}}\left(\frac{1}{2}n^{-1/2} + \frac{1}{2}m^{-1/2}\right).$$

Plugging this estimate, together with the bound on the KL term, into Lemma 18 for the functional $T(\mu, \nu) = \mathrm{SOW}_{\phi,p}^p(\mu, \nu)$ yields

$$\inf_{\widehat{T}} \sup_{\mu,\nu \in \mathcal{P}(K)} \mathbb{E}_{\mu,\nu} \left|\widehat{T} - \mathrm{SOW}_{\phi,p}^p(\mu, \nu)\right| \geq 2^{-(p+6)} \left(1 - \frac{1}{\sqrt{6}}\right) \kappa_{p,d} D^p \psi_{\phi,p}^{\mathrm{cav}}\left(\frac{1}{2}n^{-1/2} + \frac{1}{2}m^{-1/2}\right).$$

This lower bound matches the upper bound in Theorem 3 up to a multiplicative constant depending only on d and p, thereby completing the proof. □

G COMPUTATIONAL DETAILS

G.1 CLOSED-FORM COMPUTATION FOR ϕ_1

Algorithm 1 approximates $\mathrm{SOW}^p_{\phi,p}(\widehat{\mu}_n, \widehat{\nu}_m)$ by combining Monte Carlo integration with bisection for the projected OW distances. When these distances admit closed-form expressions, the bisection steps can be omitted, leaving only the Monte Carlo approximation.

Recall that, for $\theta_1, \dots, \theta_L \overset{\text{i.i.d.}}{\sim} \sigma$, the Monte Carlo approximation is

$$\widehat{\mathrm{SOW}}^p_{\phi,p}(\widehat{\mu}_n, \widehat{\nu}_m, L) = \frac{1}{L}\sum_{l=1}^{L}\left[\inf\left\{\lambda > 0 : \int_0^1 \phi\left(\frac{|\Delta_{\theta_l}(u)|}{\lambda}\right)\mathrm{d}u \le 1\right\}\right]^p,$$

where Δ_θ is the projected quantile gap defined in equation 11. Thus, for each direction, it remains to evaluate the scalar infimum inside the sum. Bisection provides a numerical approximation when no closed-form expression is available.

We consider the Orlicz function $\phi_1(t) = \frac{t^2+t^4}{2}$ with $t \ge 0$, used in our experiments. Fix a direction θ and define

$$A_\theta \coloneqq \int_0^1 |\Delta_\theta(u)|^2 \,\mathrm{d}u, \quad B_\theta \coloneqq \int_0^1 |\Delta_\theta(u)|^4 \,\mathrm{d}u.$$

For every $\lambda > 0$, the constraint in the infimum becomes

$$\int_0^1 \phi_1\left(\frac{|\Delta_\theta(u)|}{\lambda}\right)\mathrm{d}u \le 1 \Leftrightarrow \frac{A_\theta}{2\lambda^2} + \frac{B_\theta}{2\lambda^4} \le 1$$
$$\Leftrightarrow \lambda^2 \ge \frac{A_\theta + \sqrt{A_\theta^2 + 8B_\theta}}{4}.$$

Taking the infimum over $\lambda > 0$ yields

$$\mathrm{W}_{\phi_1}\left(\Pi^\theta_\# \widehat{\mu}_n, \Pi^\theta_\# \widehat{\nu}_m\right) = \sqrt{\frac{A_\theta + \sqrt{A_\theta^2 + 8B_\theta}}{4}}.$$

Consequently, each projected distance can be computed directly from A_θ and B_θ. Then, the Monte Carlo approximation can be evaluated without bisection.

G.2 PROOF OF PROPOSITION 5

Throughout this proof, the observed samples are held fixed, and all expectations are taken over the projection directions.

Using the triangle inequality, we obtain

$$\mathbb{E}\left|\widehat{\mathrm{SOW}}^p_{\phi,p}(\widehat{\mu}_n, \widehat{\nu}_m, L, T) - \mathrm{SOW}^p_{\phi,p}(\widehat{\mu}_n, \widehat{\nu}_m)\right|$$
$$\le \mathbb{E}\left|\widehat{\mathrm{SOW}}^p_{\phi,p}(\widehat{\mu}_n, \widehat{\nu}_m, L, T) - \widehat{\mathrm{SOW}}^p_{\phi,p}(\widehat{\mu}_n, \widehat{\nu}_m, L)\right| + \mathbb{E}\left|\widehat{\mathrm{SOW}}^p_{\phi,p}(\widehat{\mu}_n, \widehat{\nu}_m, L) - \mathrm{SOW}^p_{\phi,p}(\widehat{\mu}_n, \widehat{\nu}_m)\right|.$$

Now, we control the first term $\mathbb{E}\left|\widehat{\mathrm{SOW}}^p_{\phi,p}(\widehat{\mu}_n, \widehat{\nu}_m, L, T) - \widehat{\mathrm{SOW}}^p_{\phi,p}(\widehat{\mu}_n, \widehat{\nu}_m, L)\right|$. We have:

$$\begin{aligned}
&\mathbb{E}\left|\widehat{\mathrm{SOW}}^p_{\phi,p}(\widehat{\mu}_n, \widehat{\nu}_m, L, T) - \widehat{\mathrm{SOW}}^p_{\phi,p}(\widehat{\mu}_n, \widehat{\nu}_m, L)\right| \\
&= \mathbb{E}\left|\frac{1}{L}\sum_{l=1}^{L}\widehat{\mathrm{W}}^p_\phi(\Pi^{\theta_l}_\#\widehat{\mu}_n, \Pi^{\theta_l}_\#\widehat{\nu}_m, T) - \frac{1}{L}\sum_{l=1}^{L}\mathrm{W}^p_\phi(\Pi^{\theta_l}_\#\widehat{\mu}_n, \Pi^{\theta_l}_\#\widehat{\nu}_m)\right| \\
&= \mathbb{E}\left|\frac{1}{L}\sum_{l=1}^{L}\left(\widehat{\mathrm{W}}^p_\phi(\Pi^{\theta_l}_\#\widehat{\mu}_n, \Pi^{\theta_l}_\#\widehat{\nu}_m, T) - \mathrm{W}^p_\phi(\Pi^{\theta_l}_\#\widehat{\mu}_n, \Pi^{\theta_l}_\#\widehat{\nu}_m)\right)\right| \\
&\le \frac{1}{L}\sum_{l=1}^{L}\mathbb{E}\left|\widehat{\mathrm{W}}^p_\phi(\Pi^{\theta_l}_\#\widehat{\mu}_n, \Pi^{\theta_l}_\#\widehat{\nu}_m, T) - \mathrm{W}^p_\phi(\Pi^{\theta_l}_\#\widehat{\mu}_n, \Pi^{\theta_l}_\#\widehat{\nu}_m)\right|.
\end{aligned}$$

From equation 12, we have

$$0 \le L_\theta \le \mathrm{W}_\phi(\Pi^\theta_\# \widehat{\mu}_n, \Pi^\theta_\# \widehat{\nu}_m) \le U_\theta, \qquad \theta \in \mathbb{S}^{d-1},$$

so that $[L_\theta, U_\theta]$ is a valid initial bracket for Algorithm 1. Since ϕ is increasing, $\lambda \mapsto \phi(|\Delta_\theta(u)|/\lambda)$ is decreasing for each u. Therefore each bisection step retains $\mathrm{W}_\phi(\Pi^{\theta_l}_\# \widehat{\mu}_n, \Pi^{\theta_l}_\# \widehat{\nu}_m)$ in the current bracket while halving its width. In particular $\widehat{\mathrm{W}}_\phi(\Pi^{\theta_l}_\# \widehat{\mu}_n, \Pi^{\theta_l}_\# \widehat{\nu}_m, T) \in [L_{\theta_l}, U_{\theta_l}]$, and the error of the bisection method (Burden, 2011) is

$$\left|\widehat{\mathrm{W}}_\phi(\Pi^{\theta_l}_\# \widehat{\mu}_n, \Pi^{\theta_l}_\# \widehat{\nu}_m, T) - \mathrm{W}_\phi(\Pi^{\theta_l}_\# \widehat{\mu}_n, \Pi^{\theta_l}_\# \widehat{\nu}_m)\right| \le \frac{|U_{\theta_l} - L_{\theta_l}|}{2^T}.$$

If $L_{\theta_l} = U_{\theta_l}$, the algorithm returns their common value without bisection, so the error is zero and the same bound still holds.

Moreover, since $p \ge 1$, the map $s \mapsto s^p$ has increasing derivative ps^{p-1} on $(0, \infty)$, so the mean value theorem gives $|a^p - b^p| \le pM^{p-1}|a - b|$ for all $a, b \in [0, M]$. Applying this with $M = U_{\theta_l}$, we obtain:

$$\begin{aligned}
&\mathbb{E}\left[\left|\widehat{\mathrm{SOW}}^p_{\phi,p}(\widehat{\mu}_n, \widehat{\nu}_m, L, T) - \widehat{\mathrm{SOW}}^p_{\phi,p}(\widehat{\mu}_n, \widehat{\nu}_m, L)\right|\right] \\
&\le \frac{p}{L}\sum_{l=1}^{L} \mathbb{E}\left[\left|\widehat{\mathrm{W}}_\phi(\Pi^{\theta_l}_\# \widehat{\mu}_n, \Pi^{\theta_l}_\# \widehat{\nu}_m, T) - \mathrm{W}_\phi(\Pi^{\theta_l}_\# \widehat{\mu}_n, \Pi^{\theta_l}_\# \widehat{\nu}_m)\right| U^{p-1}_{\theta_l}\right] \\
&\le \frac{p}{L}\sum_{l=1}^{L} \mathbb{E}\left[\frac{|U_{\theta_l} - L_{\theta_l}|}{2^T} U^{p-1}_{\theta_l}\right] \\
&\le \frac{p}{L}\sum_{l=1}^{L} \mathbb{E}\left[\frac{U_{\theta_l}}{2^T} U^{p-1}_{\theta_l}\right] \\
&= \frac{p}{L2^T}\sum_{l=1}^{L} \mathbb{E}\left[U^p_{\theta_l}\right] = \frac{p}{2^T}\int_{\mathbb{S}^{d-1}} U^p_\theta \,\mathrm{d}\sigma(\theta),
\end{aligned}$$

where the third inequality uses $0 \le L_{\theta_l} \le U_{\theta_l}$, and the last equality holds since $\theta_1, \ldots, \theta_L \overset{\text{i.i.d.}}{\sim} \sigma$.

Now, we control the second term $\mathbb{E}\left|\widehat{\mathrm{SOW}}^p_{\phi,p}(\widehat{\mu}_n, \widehat{\nu}_m, L) - \mathrm{SOW}^p_{\phi,p}(\widehat{\mu}_n, \widehat{\nu}_m)\right|$. We follow the argument of Nadjahi et al. (2020, Theorem 6), with W_ϕ as the base distance.

By the Cauchy–Schwarz inequality,

$$\begin{aligned}
&\mathbb{E}\left|\widehat{\mathrm{SOW}}^p_{\phi,p}(\widehat{\mu}_n, \widehat{\nu}_m, L) - \mathrm{SOW}^p_{\phi,p}(\widehat{\mu}_n, \widehat{\nu}_m)\right| \\
&= \mathbb{E}\left|\frac{1}{L}\sum_{l=1}^{L} \mathrm{W}^p_\phi(\Pi^{\theta_l}_\# \widehat{\mu}_n, \Pi^{\theta_l}_\# \widehat{\nu}_m) - \mathbb{E}\left[\mathrm{W}^p_\phi(\Pi^\theta_\# \widehat{\mu}_n, \Pi^\theta_\# \widehat{\nu}_m)\right]\right| \\
&\le \sqrt{\mathbb{E}\left|\frac{1}{L}\sum_{l=1}^{L} \mathrm{W}^p_\phi(\Pi^{\theta_l}_\# \widehat{\mu}_n, \Pi^{\theta_l}_\# \widehat{\nu}_m) - \mathbb{E}\left[\mathrm{W}^p_\phi(\Pi^\theta_\# \widehat{\mu}_n, \Pi^\theta_\# \widehat{\nu}_m)\right]\right|^2} \\
&= \sqrt{\mathrm{Var}\left[\frac{1}{L}\sum_{l=1}^{L} \mathrm{W}^p_\phi(\Pi^{\theta_l}_\# \widehat{\mu}_n, \Pi^{\theta_l}_\# \widehat{\nu}_m)\right]} = \frac{1}{\sqrt{L}}\sqrt{\mathrm{Var}_{\theta\sim\sigma}\, \mathrm{W}^p_\phi(\Pi^\theta_\# \widehat{\mu}_n, \Pi^\theta_\# \widehat{\nu}_m)}.
\end{aligned}$$

Combining results, we have:

$$\begin{aligned}
&\mathbb{E}\left|\widehat{\mathrm{SOW}}^p_{\phi,p}(\widehat{\mu}_n, \widehat{\nu}_m, L, T) - \mathrm{SOW}^p_{\phi,p}(\widehat{\mu}_n, \widehat{\nu}_m)\right| \\
&\le \frac{p}{2^T}\int_{\mathbb{S}^{d-1}} U^p_\theta \,\mathrm{d}\sigma(\theta) + \frac{1}{\sqrt{L}}\,\mathrm{Var}_{\theta\sim\sigma}\left[\mathrm{W}^p_\phi(\Pi^\theta_\# \widehat{\mu}_n, \Pi^\theta_\# \widehat{\nu}_m)\right]^{1/2},
\end{aligned}$$

which completes the proof.

H BEYOND BOUNDED SUPPORT

The statistical results in the main text are developed under a common bounded-support assumption. A natural question is whether this assumption can be relaxed. For the Wasserstein and sliced Wasserstein distances, a standard route is to replace bounded support by finite higher-order moments (see, for example, (Fournier & Guillin, 2015; Nadjahi et al., 2020)). Although such conditions allow genuinely unbounded distributions, the same approach can fail for a general Orlicz generator ϕ. The purpose of this appendix is not to characterize the largest class of unbounded distributions for which empirical $\mathrm{SOW}_{\phi,p}$ convergence holds. Instead, in this discussion, we ask a more modest question: If $\mathbb{E}\,\mathrm{SOW}^p_{\phi,p}(\widehat{\mu}_n,\mu)\to 0$ as $n\to\infty$, where must the target law μ lie? We first give a counterexample to demonstrate that polynomial moment conditions can still be insufficient, then isolate the obstruction suggested by that example, and finally derive a necessary tail-integrability condition.

H.1 WHY POLYNOMIAL MOMENT CONDITIONS CAN STILL BE INSUFFICIENT

Example 1 (All polynomial moments may still be insufficient)**.** *Let $d=1$ and let $\phi(t)=e^t-t-1$ for $t\geq 0$. Consider the exponential distribution with rate 1, denoted by $\mu=\mathrm{Exp}(1)$, and let $X\sim\mu$. First, μ has finite moments of every polynomial order. Indeed, for every $q>0$,*

$$\mathbb{E}\left[X^q\right]=\Gamma(q+1)<+\infty,$$

where Γ denotes the gamma function.

We next show that $\mu\in\mathcal{P}_\phi(\mathbb{R})$. For every $\lambda>1$, we have

$$\mathbb{E}e^{X/\lambda}=\int_0^{+\infty}e^{x/\lambda}e^{-x}\mathrm{d}x=\frac{\lambda}{\lambda-1}.$$

Since $\mathbb{E}[X]=1$, it follows that, for every $\lambda>1$,

$$\mathbb{E}\phi(|X|/\lambda)=\mathbb{E}\phi(X/\lambda)=\mathbb{E}e^{X/\lambda}-\mathbb{E}\left(X/\lambda\right)-1=\frac{\lambda}{\lambda-1}-\frac{1}{\lambda}-1<+\infty.$$

In particular, taking $\lambda=2$, we obtain

$$\mathbb{E}\phi\left(\frac{|X|}{2}\right)=\frac{1}{2}\leq 1.$$

Therefore, $\mu\in\mathcal{P}_\phi(\mathbb{R})$.

Nevertheless, let ν be any probability measure with bounded support. More precisely, assume that, for some $M>0$,

$$\mathrm{supp}\left(\nu\right)\subset[-M,M].$$

Let (X,Y) be an arbitrary coupling of μ and ν. Since $Y\leq M$ almost surely, on the event $\{X>M\}$ we have

$$|X-Y|\geq X-M.$$

Fix $0<\lambda\leq 1$. By the monotonicity of ϕ, we obtain

$$\begin{aligned}\mathbb{E}\phi\left(\frac{|X-Y|}{\lambda}\right)&\geq\int_M^{+\infty}\phi\left(\frac{x-M}{\lambda}\right)e^{-x}\mathrm{d}x\\&=\int_M^{+\infty}\left[e^{(x-M)/\lambda}-\frac{x-M}{\lambda}-1\right]e^{-x}\mathrm{d}x.\end{aligned}$$

The contributions of the last two terms are finite, whereas the exponential term is not integrable. Indeed,

$$\int_M^{+\infty}e^{(x-M)/\lambda}e^{-x}\mathrm{d}x=e^{-M/\lambda}\int_M^{+\infty}e^{(1/\lambda-1)x}\mathrm{d}x=+\infty,$$

because $1/\lambda-1\geq 0$. Therefore, no $\lambda\leq 1$ is admissible in the definition of the Luxemburg norm, and hence

$$\|X-Y\|_\phi\geq 1.$$

Since the coupling (X, Y) was arbitrary, we conclude that

$$\mathrm{W}_\phi(\mu, \nu) \geq 1.$$

Now, for every sample realization, the empirical measure $\widehat{\mu}_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ has bounded support. Applying the preceding argument with $\nu = \widehat{\mu}_n$ gives

$$\mathrm{W}_\phi(\mu, \widehat{\mu}_n) \geq 1, \quad \textit{for every } n \geq 1.$$

In dimension one, $\mathrm{SOW}_{\phi,p} = \mathrm{W}_\phi$, and hence

$$\mathbb{E}\left[\mathrm{SOW}^p_{\phi,p}(\widehat{\mu}_n, \mu)\right] \geq 1, \quad \textit{for every } n \geq 1.$$

In this example, although μ already belongs to $\mathcal{P}_\phi(\mathbb{R})$ and it has finite moments of every polynomial order, the empirical measures do not converge to μ in $\mathrm{SOW}_{\phi,p}$ for any sample realization. Thus one-scale Orlicz integrability, even together with finite moments of every polynomial order, can still be insufficient. However, the proof also suggests why the approximation fails: empirical measures are boundedly supported, while the far tail of the target law remains expensive in the Luxemburg geometry. We now isolate this mechanism.

H.2 Insights from the Counterexample

We begin with the one-dimensional setting. Let μ be a probability measure on $\mathbb{R}$, and let $X \sim \mu$. Assume that μ can be approximated in W_ϕ by compactly supported probability measures. More precisely, suppose that there exists a sequence $(\nu_k)_{k\in\mathbb{N}}$ of compactly supported probability measures satisfying

$$\mathrm{W}_\phi(\mu, \nu_k) \to 0 \quad \text{as } k \to +\infty. \tag{23}$$

For each k, choose $0 < M_k < +\infty$ such that

$$\mathrm{supp}(\nu_k) \subset [-M_k, M_k].$$

Let (X, Y_k) be any coupling of μ and ν_k. Since $|Y_k| \leq M_k$ almost surely,

$$|X - Y_k| \geq |X| - |Y_k| \geq |X| - M_k.$$

Because the left-hand side is nonnegative,

$$|X - Y_k| \geq (|X| - M_k)_+,$$

where we denote $(a)_+ = \max\{a, 0\}$. By monotonicity of the Luxemburg norm,

$$\|X - Y_k\|_\phi \geq \left\|(|X| - M_k)_+\right\|_\phi.$$

Consequently,

$$\mathrm{W}_\phi(\mu, \nu_k) \geq \|\,(|X| - M_k)_+\,\|_\phi.$$

Combining with equation 23 yields

$$\|\,(|X| - M_k)_+\,\|_\phi \to 0 \quad \text{as } k \to +\infty. \tag{24}$$

Now define

$$\tau(R) \coloneqq \|\,(|X| - R)_+\,\|_\phi, \quad R \geq 0.$$

It is easy to check that the function $R \mapsto \tau(R)$ is decreasing. For every $\varepsilon > 0$, equation 24 provides an index k such that $\tau(M_k) < \varepsilon$. Hence, for every $R \geq M_k$, we have $\tau(R) \leq \tau(M_k) < \varepsilon$. As a consequence,

$$\|\,(|X| - R)_+\,\|_\phi \to 0 \quad \text{as } R \to +\infty.$$

This is the first structural conclusion suggested by Example 1: if boundedly supported measures can approach the target law in Orlicz–Wasserstein distance, then the far tail of the target must disappear in Luxemburg norm.

Informal interpretation. Recall that the Luxemburg norm is

$$\|X\|_\phi = \inf\left\{\lambda > 0 : \mathbb{E}\phi\left(\frac{|X|}{\lambda}\right) \le 1\right\}.$$

If the tail $(|X| - R)_+$ is to become arbitrarily small in this norm, then one must eventually control it at arbitrarily small Luxemburg scales. A small scale ϵ places a large factor $1/\epsilon$ inside ϕ. This suggests that one-scale integrability may not be enough. Therefore, a natural candidate is integrability at every positive multiplicative scale, namely

$$\mathbb{E}\phi\left(a\,|X|\right) < +\infty \quad \text{for every } a > 0.$$

The next lemma verifies that intuition.

Lemma 20. *Let Z be a real-valued random variable. If $\lim_{R\to+\infty} \|\,(|Z| - R)_+\,\|_\phi = 0$, then $\mathbb{E}\phi\left(a\,|Z|\right) < +\infty$ for every $a > 0$.*

Proof. Fix $a > 0$. Since $\lim_{R\to+\infty} \|\,(|Z| - R)_+\,\|_\phi = 0$, we can choose R large enough such that

$$\|\,(|Z| - R)_+\,\|_\phi < 1/(2a).$$

Since $\mathbb{R}$ is densely ordered, we may choose $t \in \mathbb{R}$ such that

$$\|\,(|Z| - R)_+\,\|_\phi < t < 1/(2a).$$

We first justify that t is an admissible Luxemburg scale. Since $t > \|\,(|Z| - R)_+\,\|_\phi$, there exists $0 < \lambda_0 < t$ such that

$$\mathbb{E}\phi\left(\frac{(|Z| - R)_+}{\lambda_0}\right) \le 1.$$

Since $t > \lambda_0$, the monotonicity of ϕ yields

$$\mathbb{E}\phi\left(\frac{(|Z| - R)_+}{t}\right) \le 1.$$

Besides, since $2a < 1/t$, again, the monotonicity of ϕ gives

$$\mathbb{E}\phi\left(2a\,(|Z| - R)_+\right) \le 1. \tag{25}$$

Finally, since $|Z| \le R + (|Z| - R)_+$, using monotonicity and convexity of ϕ, we have

$$\phi\left(a\,|Z|\right) \le \frac{1}{2}\phi\left(2aR\right) + \frac{1}{2}\phi\left[2a\,(|Z| - R)_+\right].$$

Taking expectation and using equation 25,

$$\mathbb{E}\phi\left(a\,|Z|\right) \le \frac{1}{2}\phi\left(2aR\right) + \frac{1}{2} < +\infty.$$

Since $a > 0$ is arbitrary, we complete the proof. □

With this ingredient in hand, we now state a necessary condition for vanishing expected empirical $\mathrm{SOW}^p_{\phi,p}$ loss.

Proposition 7. *Let $p \in [1, +\infty)$ and let $\mu \in \mathcal{P}_\phi(\mathbb{R}^d)$. Let $\widehat{\mu}_n$ denote the empirical measure of n i.i.d. samples drawn from μ. If $\lim_{n\to+\infty} \mathbb{E}\,\mathrm{SOW}^p_{\phi,p}\left(\widehat{\mu}_n, \mu\right) = 0$, then $\mu \in \mathcal{P}^{\mathrm{all}}_\phi\left(\mathbb{R}^d\right)$, where*

$$\mathcal{P}^{\mathrm{all}}_\phi\left(\mathbb{R}^d\right) \coloneqq \left\{P \in \mathcal{P}\left(\mathbb{R}^d\right) : \int_{\mathbb{R}^d} \phi\left(\frac{\|x\|}{\lambda}\right) \mathrm{d}P(x) < +\infty \quad \forall \lambda > 0\right\}.$$

Proof. Since $\lim_{n\to+\infty} \mathbb{E}\,\mathrm{SOW}^p_{\phi,p}\left(\widehat{\mu}_n, \mu\right) = 0$, we can choose a subsequence $(n_k)_{k\in\mathbb{N}}$ such that

$$\mathbb{E}\,\mathrm{SOW}^p_{\phi,p}\left(\widehat{\mu}_{n_k}, \mu\right) \le 2^{-2kp}.$$

By Markov's inequality,

$$\mathbb{P}\left(\mathrm{SOW}_{\phi,p}^{p}\left(\widehat{\mu}_{n_k},\mu\right)>2^{-k}\right)\leq 2^{k}\mathbb{E}\,\mathrm{SOW}_{\phi,p}^{p}(\widehat{\mu}_{n_k},\mu)\leq 2^{-kp}.$$

Since $p\geq 1$, $\displaystyle\sum_{k=1}^{\infty}2^{-kp}<+\infty$. Therefore, the Borel–Cantelli lemma gives

$$\lim_{k\to+\infty}\mathrm{SOW}_{\phi,p}^{p}\left(\widehat{\mu}_{n_k},\mu\right)=0\quad\text{almost surely.}\tag{26}$$

We now fix a sample path on which equation 26 holds. Using Proposition 3, for every coordinate vector e_j, we have

$$\mathrm{W}_{\phi}\left(\Pi_{\#}^{e_j}\widehat{\mu}_{n_k},\Pi_{\#}^{e_j}\mu\right)\to 0\quad\text{as } k\to+\infty.\tag{27}$$

Write $X=(X_1,\ldots,X_d)\sim\mu$. For every k, the projected measure $\Pi_{\#}^{e_j}\widehat{\mu}_{n_k}$ has finite support. Combining the argument in Section H.2 with equation 27, we deduce

$$\|\left(|X_j|-R\right)_{+}\|_{\phi}\to 0\quad\text{as } R\to+\infty.$$

Applying Lemma 20, we deduce

$$\mathbb{E}\phi\left(b\,|X_j|\right)<+\infty\quad\text{for every } b>0 \text{ and every } j=1,\ldots,d.$$

Finally, for a fixed $a>0$, the monotonicity and convexity of ϕ give

$$\phi\left(a\|X\|\right)\leq\phi\left(a\sum_{j=1}^{d}|X_j|\right)=\phi\left(\frac{1}{d}\sum_{j=1}^{d}da\,|X_j|\right)\leq\frac{1}{d}\sum_{j=1}^{d}\phi\left(da\,|X_j|\right).$$

By the preceding conclusion, each term on the right has finite expectation. Hence,

$$\mathbb{E}\phi\left(a\|X\|\right)<+\infty.$$

Since $a>0$ is arbitrary, $\mu\in\mathcal{P}_{\phi}^{\mathrm{all}}\left(\mathbb{R}^d\right)$. □

The preceding proposition identifies the tail restriction imposed by Orlicz geometry. However, it does not yet exploit the additional integrability information carried by the p-th power in the expected loss. In fact, the same assumption also forces the target distribution to have a finite p-th moment. Combining these two restrictions yields the following stronger necessary condition

Theorem 5. *Let $p\in[1,+\infty)$ and let $\mu\in\mathcal{P}_{\phi}\left(\mathbb{R}^d\right)$. Let $\widehat{\mu}_n$ denote the empirical measure of n i.i.d. samples drawn from μ. If $\lim_{n\to+\infty}\mathbb{E}\,\mathrm{SOW}_{\phi,p}^{p}\left(\widehat{\mu}_n,\mu\right)=0$, then $\mu\in\mathcal{P}_{\phi}^{\mathrm{all}}\left(\mathbb{R}^d\right)\cap\mathcal{P}_{p}\left(\mathbb{R}^d\right)$.*

Proof. By the preceding proposition, it remains only to prove that $\mu\in\mathcal{P}_{p}\left(\mathbb{R}^d\right)$. Let $X_1,X_2,\ldots,X_n$ be i.i.d. with law μ and define

$$\bar{X}_n\coloneqq\frac{1}{n}\sum_{i=1}^{n}X_i,\quad m\coloneqq\mathbb{E}X_1.$$

The vector m is well defined. Indeed, since $\mu\in\mathcal{P}_{\phi}\left(\mathbb{R}^d\right)$, Lemma 7 gives

$$\mathrm{W}_1\left(\mu,\delta_0\right)\leq\phi^{-1}(1)\,\mathrm{W}_{\phi}\left(\mu,\delta_0\right)<+\infty.$$

Therefore,

$$\mathbb{E}\|X_1\|=\mathrm{W}_1\left(\mu,\delta_0\right)<+\infty.$$

For any $\theta\in\mathbb{S}^{d-1}$, let (U,V) be an arbitrary coupling of $\Pi_{\#}^{\theta}\widehat{\mu}_n$ and $\Pi_{\#}^{\theta}\mu$. Then,

$$\mathbb{E}\,|U-V|\geq|\mathbb{E}U-\mathbb{E}V|=\left|\langle\theta,\bar{X}_n\rangle-\langle\theta,m\rangle\right|=\left|\langle\theta,\bar{X}_n-m\rangle\right|.$$

Taking the infimum over all couplings yields

$$\mathrm{W}_1\left(\Pi_{\#}^{\theta}\widehat{\mu}_n,\Pi_{\#}^{\theta}\mu\right)\geq\left|\langle\theta,\bar{X}_n-m\rangle\right|.$$

By Lemma 7, we have

$$\mathrm{W}_\phi\left(\Pi^\theta_\# \widehat{\mu}_n, \Pi^\theta_\# \mu\right) \geq \frac{\left|\langle \theta, \bar{X}_n - m\rangle\right|}{\phi^{-1}(1)}.$$

Hence,

$$\mathrm{SOW}^p_{\phi,p}\left(\widehat{\mu}_n, \mu\right) \geq \frac{1}{[\phi^{-1}(1)]^p} \int_{\mathbb{S}^{d-1}} \left|\langle \theta, \bar{X}_n - m\rangle\right|^p \mathrm{d}\sigma(\theta) = \frac{\kappa_{p,d}}{[\phi^{-1}(1)]^p} \|\bar{X}_n - m\|^p,$$

where we used equation 22 in the last equality. Therefore,

$$\mathbb{E}\|\bar{X}_n - m\|^p \leq \frac{\left[\phi^{-1}(1)\right]^p}{\kappa_{p,d}} \mathbb{E}\,\mathrm{SOW}^p_{\phi,p}\left(\widehat{\mu}_n, \mu\right) \xrightarrow{n\to+\infty} 0.$$

As a consequence, there exists $n_0 \in \mathbb{N}$ with $n_0 \geq 2$, we have

$$\mathbb{E}\|\bar{X}_{n_0} - m\|^p < +\infty.$$

We set $Z_i := X_i - m$ and $R := \sum_{i=2}^{n_0} Z_i$. Then, Z_1 and R are independent and $Z_1 + R = n_0\left(\bar{X}_{n_0} - m\right)$. So, $\mathbb{E}\|Z_1 + R\|^p < +\infty$. The independence of Z_1 and R, together with Tonelli's theorem gives

$$\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} \|z + r\|^p \mathrm{d}\rho(z) d\eth(r) < +\infty,$$

where ρ and $\eth$ denote the laws of Z_1 and R respectively. As a result,

$$\int_{\mathbb{R}^d} \|z + r\|^p d\rho(z) < +\infty \quad \eth\text{-almost every } r.$$

Fix one such r_0. Since $p \geq 1$, for all $z \in \mathbb{R}^d$ we have

$$\|z\|^p \leq 2^{p-1}\left(\|z + r_0\|^p + \|r_0\|^p\right).$$

Taking the integral with respect to ρ yields

$$\mathbb{E}\|Z_1\|^p < +\infty.$$

Finally, since $X_1 = Z_1 + m$, we have

$$\mathbb{E}\|X_1\|^p \leq 2^{p-1}\left(\mathbb{E}\|Z_1\|^p + \|m\|^p\right) < +\infty.$$

Consequently, $\mu \in \mathcal{P}_p\left(\mathbb{R}^d\right)$. □

The theorem above gives a necessary condition for $\mathbb{E}[\mathrm{SOW}^p_{\phi,p}\left(\widehat{\mu}_n, \mu\right)]$ converges to 0 as n goes to $+\infty$. It shows that any class of unbounded distributions on which the expected empirical loss vanishes must be contained in $\mathcal{P}^{\mathrm{all}}_\phi(\mathbb{R}^d) \cap \mathcal{P}_p\left(\mathbb{R}^d\right)$. Determining whether $\mathcal{P}^{\mathrm{all}}_\phi(\mathbb{R}^d) \cap \mathcal{P}_p\left(\mathbb{R}^d\right)$ is sufficient for vanishing expected empirical loss or whether one must impose a smaller tail class depending on both ϕ and p is left for future work.

I Technical contributions and proof roadmap

In this appendix, we provide a proof roadmap for the main statistical results established in this paper. Our goal is to make explicit where the classical sliced Wasserstein analysis relies on the power-law structure $\phi(t) = t^p$, and to highlight the additional technical ingredients required to extend these arguments to a general Orlicz function

I.1 Empirical convergence of the SOW distance

We first consider the one-sample problem in Theorem 2. Fix a direction $\theta \in \mathbb{S}^{d-1}$. Since both projected measures are supported on an interval of length at most $D > 0$, Lemma 8 relates the projected OW distance to the ordinary W_1 distance

$$\mathrm{W}_\phi\left(\Pi^\theta_\# \widehat{\mu}_n, \Pi^\theta_\# \mu\right) \leq D\omega_\phi\left(\frac{\mathrm{W}_1\left(\Pi^\theta_\# \widehat{\mu}_n, \Pi^\theta_\# \mu\right)}{D}\right).$$

The advantage of this comparison is that W_1 has a simple one-dimensional CDF representation (see, for example, Bobkov & Ledoux (2019, Theorem 2.9)). If $\widehat{F}_{n,\theta}$ and F_θ denote the CDFs of the two projected measures, then

$$\mathrm{W}_1\left(\Pi^\theta_\# \widehat{\mu}_n, \Pi^\theta_\# \mu\right) = \int_{\mathbb{R}} \left|\widehat{F}_{n,\theta}(x) - F_\theta(x)\right| \mathrm{d}x \leq D\|\widehat{F}_{n,\theta} - F_\theta\|_\infty.$$

Consequently, we have

$$\mathrm{W}_\phi\left(\Pi^\theta_\# \widehat{\mu}_n, \Pi^\theta_\# \mu\right) \leq D\omega_\phi\left(\Delta_{n,\theta}\right),$$

where $\Delta_{n,\theta} = \|\widehat{F}_{n,\theta} - F_\theta\|_\infty$. At this point, the estimation problem is reduced to a single quantity $\mathbb{E}\left[\omega_\phi\left(\Delta_{n,\theta}\right)^p\right]$. On the other hand, by the Dvoretzky-Kiefer-Wolfowitz (DKW) inequality, we have

$$\mathbb{P}\left(\Delta_{n,\theta} > u\right) \leq 2\exp\left(-2nu^2\right).$$

In the classical sliced Wasserstein case, $\phi(t) = t^p$ and therefore $\omega_\phi(u)^p = u$. Hence, the remaining quantity becomes simply $\mathbb{E}\left[\Delta_{n,\theta}\right]$, which can be directly controlled by using DKW inequality. Indeed, we have

$$\mathbb{E}\left[\Delta_{n,\theta}\right] = \int_0^{+\infty} \mathbb{P}\left(\Delta_{n,\theta} > u\right) \mathrm{d}u \lesssim n^{-1/2}.$$

For a general Orlicz function, however, the identity $\omega_\phi(u)^p = u$ is no longer available. This is precisely where an additional Orlicz-specific argument is needed. Lemma addresses this gap by decomposing the range of $\Delta_{n,\theta}$ into suitable scales. This decomposition exploits the growth property of ω_ϕ established in Lemma 2, allowing the DKW tail bound to be applied properly and yielding the desired bound.

I.2 Estimation error of the powered SOW functional

We next consider the two-sample estimation problem for the powered functional $\mathrm{SOW}^p_{\phi,p}$ in Theorem 3. Fix a direction $\theta \in \mathbb{S}^{d-1}$, and define the normalized empirical and population quantile-gap profiles by

$$\widehat{h}_\theta(s) = \frac{\left|F^{-1}_{\Pi^\theta_\# \widehat{\mu}_n}(s) - F^{-1}_{\Pi^\theta_\# \widehat{\nu}_m}(s)\right|}{D}, \quad h_\theta(s) = \frac{\left|F^{-1}_{\Pi^\theta_\# \mu}(s) - F^{-1}_{\Pi^\theta_\# \nu}(s)\right|}{D}, \quad s \in (0,1).$$

Both $\widehat{h}_\theta$ and h_θ take values in $[0,1]$. By the one-dimensional quantile representation and the homogeneity of the Luxemburg norm, we have

$$\left|\mathrm{W}^p_\phi\left(\Pi^\theta_\# \widehat{\mu}_n, \Pi^\theta_\# \widehat{\nu}_m\right) - \mathrm{W}^p_\phi\left(\Pi^\theta_\# \mu, \Pi^\theta_\# \nu\right)\right| = D^p \left|\|\widehat{h}_\theta\|^p_\phi - \|h_\theta\|^p_\phi\right|.$$

In the classical sliced Wasserstein setting, a key step in obtaining the parametric rate for the powered functional is to reduce its variation to projected W_1 errors (see, for example, the proof of Proposition 1 in Nietert et al. (2022)). In view of the representation above, obtaining an analogous reduction in the Orlicz setting amounts to controlling $\left|\|\widehat{h}_\theta\|^p_\phi - \|h_\theta\|^p_\phi\right|$ in terms of the L^1 discrepancy $\|\widehat{h}_\theta - h_\theta\|_{L^1}$. When $\phi(t) = t^p$, this control is immediate. Indeed, since $\|\widehat{h}_\theta\|^p_\phi = \int_0^1 |\widehat{h}_\theta(s)|^p \mathrm{d}s, \|h_\theta\|^p_\phi = \int_0^1 |h_\theta(s)|^p \,\mathrm{d}s$ and $0 \leq \widehat{h}_\theta, h_\theta \leq 1$, we have

$$\left|\|\widehat{h}_\theta\|^p_\phi - \|h_\theta\|^p_\phi\right| \leq p\|\widehat{h}_\theta - h_\theta\|_{L^1}.$$

Then, using the triangle inequality and the one-dimensional quantile representation of W_1, we have

$$\|\widehat{h}_\theta - h_\theta\|_{L^1} \leq \frac{1}{D}\left[\mathrm{W}_1\left(\Pi^\theta_\# \widehat{\mu}_n, \Pi^\theta_\# \mu\right) + \mathrm{W}_1\left(\Pi^\theta_\# \widehat{\nu}_m, \Pi^\theta_\# \nu\right)\right],$$

where the expectation of the right-hand side can be controlled by using Bobkov & Ledoux (2019, Theorem 3.2).

For a general Orlicz function, however, the preceding L^1-perturbation argument is no longer available. We therefore introduce the worst-case perturbation modulus

$$\Omega_{\phi,p}(u) = \sup_{\substack{0 \leq f,g \leq 1 \\ \|f-g\|_{L^1(0,1)} \leq u}} \left|\|f\|^p_\phi - \|g\|^p_\phi\right|, \quad u \in [0,1].$$

The main technical step is then to identify the sharp scale of this modulus. Proposition 6 shows that $\Omega_{\phi,p}(u) \asymp_p \psi^{\mathrm{cav}}_{\phi,p}(u)$. Consequently,

$$\left| \|\widehat{h}_\theta\|^p_\phi - \|h_\theta\|^p_\phi \right| \lesssim_p \psi^{\mathrm{cav}}_{\phi,p}(\|\widehat{h}_\theta - h_\theta\|_{L^1(0,1)}).$$

The concavity of $\psi^{\mathrm{cav}}_{\phi,p}$ and Jensen's inequality yield

$$\mathbb{E}\left[\psi^{\mathrm{cav}}_{\phi,p}\left(\|\widehat{h}_\theta - h_\theta\|_{L^1(0,1)}\right)\right] \le \psi^{\mathrm{cav}}_{\phi,p}\left(\mathbb{E}\|\widehat{h}_\theta - h_\theta\|_{L^1(0,1)}\right).$$

Using the monotonicity of $\psi^{\mathrm{cav}}_{\phi,p}$, the right-hand side then can be controlled by the quantile representation of W_1 together with Bobkov & Ledoux (2019, Theorem 3.2). Combining these estimates and integrating over the projection directions yields Theorem 3

I.3 Minimax lower bound of the powered SOW functional

The minimax lower bounds are established using Le Cam's two-point method. For the SOW distance, a direct two-point perturbation at scale $n^{-1/2}$ already yields the matching lower bound, so we focus here on the additional difficulty arising for the powered functional.

For the powered functional, reusing the two-point construction from Section F.3 only gives a separation of order $D^p\psi_{\phi,p}(n^{-1/2})$. However, the upper bound in Theorem 3 is governed by $D^p\psi^{\mathrm{cav}}_{\phi,p}\left(n^{-1/2}\right)$, which can be larger than $D^p\psi_{\phi,p}(n^{-1/2})$. To overcome this problem, Lemma 19 shows that for every sufficiently small h, there exist two nearby mass levels $0 \le q_0 < q_1 \le 1/2$, $q_1 - q_0 \le h$ such that $\psi_{\phi,p}(q_1) - \psi_{\phi,p}(q_0) \gtrsim_p \psi^{\mathrm{cav}}_{\phi,p}(h)$. This allows us to construct two distributions with small Kullback–Leibler divergence, while their powered SOW values are separated at the desired scale. Le Cam's method then gives the matching lower bound.