Duality and Sample Complexity for the Gromov-Wasserstein Distance

Zhengxin Zhang
Center for Applied Mathematics
Cornell University
zz658@cornell.edu

Ziv Goldfeld
School of Electrical and Computer Engineering
Cornell University
goldfeld@cornell.edu

Youssef Mroueh
IBM Research
IBM
mroueh@us.ibm.com

Bharath K. Sriperumbudur
Department of Statistics
Pennsylvania State University
bks18@psu.edu

Abstract

The Gromov-Wasserstein (GW) distance, rooted in optimal transport (OT) theory, quantifies dissimilarity between metric measure spaces and provides a framework for aligning heterogeneous datasets. While computational aspects of the GW problem have been widely studied, a duality theory and fundamental statistical questions concerning empirical convergence rates remained obscure. This work closes these gaps for the quadratic GW distance over Euclidean spaces of different dimensions d_x and d_y. We derive a dual form that represents the GW distance in terms of the well-understood OT problem. This enables employing proof techniques from statistical OT based on regularity analysis of dual potentials and empirical process theory, using which we establish the first GW empirical convergence rates. The derived two-sample rate is $n^{-2/\max\{\min\{d_x,d_y\},4\}}$ (up to a log factor when $\min\{d_x, d_y\} = 4$), which matches the corresponding rates for OT. We also provide matching lower bounds, thus establishing sharpness of the derived rates. Lastly, the duality is leveraged to shed new light on the open problem of the one-dimensional GW distance between uniform distributions on n points, illuminating why the identity and anti-identity permutations may not be optimal. Our results serve as a first step towards a comprehensive statistical theory as well as computational advancements for GW distances, based on the discovered dual formulations.

1 Introduction

The Gromov-Wasserstein (GW) distance, proposed by Mémoli in [30], quantifies discrepancy between probability distributions supported on different metric spaces by aligning them with one another. Given two metric measure (mm) spaces $(\mathcal{X}, \mathsf{d}_\mathcal{X}, \mu)$ and $(\mathcal{Y}, \mathsf{d}_\mathcal{Y}, \nu)$, the (p, q)-GW distance between them is [42]

$$\mathsf{D}_{p,q}(\mu,\nu) \coloneqq \inf_{\pi\in\Pi(\mu,\nu)} \left(\int_{\mathcal{X}\times\mathcal{Y}}\int_{\mathcal{X}\times\mathcal{Y}} \left|\mathsf{d}_\mathcal{X}(x,x')^q - \mathsf{d}_\mathcal{Y}(y,y')^q\right|^p d\pi\otimes\pi(x,y,x',y')\right)^{\frac{1}{p}}, \quad (1)$$

where $\Pi(\mu,\nu)$ is the set of all couplings between μ and ν. The GW distances thus equals the least amount of distance distortion one can achieve between the mm spaces when optimizing over all possible alignments thereof (as modeled by couplings). This approach, which is rooted in optimal transport (OT) theory [46, 36], is an L^p relaxation of the Gromov-Hausdorff distance between metric spaces and enjoys various favorable properties. Among others, the GW distance (i) identifies pairs of

Accepted to the Optimal Transport and Machine Learning Workshop at NeurIPS 2023.

mm spaces between which there exists an measure preserving isometry; (ii) defines a metric on the space of all mm spaces modulo the aforementioned isomorphic relation; and (iii) captures empirical convergence of mm space, i.e., when μ, ν are estimated by their empirical measures $\hat{\mu}_n, \hat{\nu}_n$ based on n samples. As such, the GW framework has been utilized for many applications concerning heterogeneous data, including single-cell genomics [4, 10], alignment of language models [1], shape and graph matching [29, 49, 48, 24], and heterogeneous domain adaptation [50].

While such applications predominantly run on sampled data, a statistical GW theory to guarantee valid estimation and inference has remained elusive. This gap can be attributed, in part, to the quadratic (in π) structure of the GW functional, which prevents directly using well-developed proof techniques from statistical OT. Indeed, the linear OT problem enjoys strong duality, which enables analyzing empirical OT distances via techniques from empirical process theory, such as chaining, entropy integral bounds, and the functional delta method. These approaches have proven central to the development of statistical OT, leading to a comprehensive account of empirical convergence rates [11, 5, 26, 22] and limit distributions of both classical [41, 43, 7, 27, 21, 16] and regularized OT distances [31, 3, 23, 13, 14, 16, 8, 15]. For the GW distance, on the other hand, while we know that $\mathsf{D}_{p,q}(\hat{\mu}_n, \hat{\nu}_n) \to \mathsf{D}_{p,q}(\mu, \nu)$ as $n \to \infty$ [30],[1] the rate at which this convergence happens is an open problem of theoretical and practical importance. This work closes this gap by deriving a dual formulation for the $(2, 2)$-GW distance over Euclidean spaces, and leveraging it to establish the first empirical convergence rates for the GW problem.

1.1 Contribution

Our first main contribution is a duality theory for GW, which linearizes the quadratic functional and ties it to the well-understood problems of OT. This is done by introducing an auxiliary, matrix-valued optimization variable $\mathbf{A} \in \mathbb{R}^{d_x \times d_y}$ that enables linearizing the dependence on the coupling. We then interchange the optimization over $\mathbf{A}$ and π and identify the inner problem as classical OT with respect to (w.r.t.) a cost function $c_{\mathbf{A}}$ that depends on $\mathbf{A}$. Upon verifying that $c_{\mathbf{A}}$ satisfies mild regularity conditions, we invoke OT duality to arrive at a dual formulation for $\mathsf{D}_{2,2}(\mu, \nu)^2$. The dual form involves optimization over $\mathbf{A}$, which we show can be restricted to a hypercube whose side length depends only on the second moments of μ, ν.

The GW dual form enables an analysis of expected empirical convergence rates by drawing upon proof techniques from statistical OT. Namely, we consider the rates at which $\mathbb{E}\big[\big|\mathsf{D}_{2,2}(\mu, \nu)^2 - \mathsf{D}_{2,2}(\hat{\mu}_n, \hat{\nu}_n)^2\big|\big]$ decay sto zero with n, as well as the one-sample case where ν is not estimated. Invoking strong duality we bound the empirical estimation error by the suprema of empirical processes indexed by OT dual potentials w.r.t. the cost $c_{\mathbf{A}}$, supremized over all feasible matrices $\mathbf{A}$. We then study the regularity of optimal potentials, uniformly in $\mathbf{A}$, which is the main technical difference from the corresponding OT analysis. We focus on compactly supported distributions and exploit smoothness and marginal-concavity of the cost $c_{\mathbf{A}}$ to show that optimal potentials are concave and Lipschitz. Following a chaining argument while leveraging the so-called lower complexity adaptation (LCA) principle form [22], we then establish the $n^{-2/\max\{\min\{d_x, d_y\}, 4\}}$ upper bound on the two-sample rate of the quadratic GW distance (up to a log factor when $\min\{d_x, d_y\} = 4$). We then provide matching lower bounds on the one- and two-sample empirical estimation errors, demonstrating that the said rates are sharp. The lower bound proof is constructive and utilizes a novel inequality between the quadratic GW distance and the 2-Wasserstein procrustes [18], which may be of independent interest.

Lastly, we revisit the open problem of the one-dimensional GW distance between uniform distributions on n points and use our duality theory to shed new light on it. We consider the peculiar example from [2], where, contrary to common belief (cf. [45]), the identity and anti-identity permutations were shown to not necessarily be optimal. Our dual form allows representing the GW distance on $\mathbb{R}$ as a sum of concave and convex functions, explaining why the optimum need not be attained at the boundary. We visualize the different regimes of optimal solutions via simple numerical simulations.

[1] [30] established this convergence for compact mm spaces and $q = 1$, but the argument readily extends to any $q \geq 1$ and arbitrary mm space, so long that μ, ν have bounded pq-th moments.

1.2 Literature review

The GW distance was first proposed in [30] as an L^p relaxation of the Gromov-Hausdorff distance between metric spaces. Basic structural properties of the distance were also established in that work, with more advanced aspects concerning topology and curvature addressed in [42]. The existence of Gromov-Monge maps was studied in [12], showing that optimal couplings are induced by a bimap (viz. two-way map) under quite general conditions. Targeting analytic solutions, optimal couplings between Gaussian distributions were explored in [9], but only upper and lower bounds on the GW value were derived. An exact characterization of the optimal coupling and cost is known for the entropic inner product GW distance between Gaussians [25].

As GW distances grew in popularity for applications, computational tractability became increasingly important. However, exact computation of the GW distance is generally a quadratic assignment problem, which is NP-complete [6]. For this reason, significant attention was devoted to variants of the GW problem that circumvent this computational hardness. The sliced GW distance [45] attempts to reduce the computational burden by considering the average of GW distances between one-dimensional projections of the marginals. However, unlike one-dimensional OT, the GW problem does not have a known simple solution even in one dimension [2]. Another approach is to relax the strict marginal constraints to obtain the unbalanced GW distance [39], which lends well for convex/conic relaxations. A variant that directly optimizes over bi-directional Monge maps between the mm space was considered in [51]. While these methods offer certain advantages, it is the approach based on entropic regularization that is most frequently used in practice. This is since entropic GW (EGW) is computable via iterative optimization routines that employ Sinkhorn iterations [40, 33, 37, 35], which allows scalability and parallelization in large-scale applications. Another notable concurrent work is [19], which pointed out that LCA principle can be applied to GW, and also showed that the dimensionality dependence of the convergence rate is $\min\{d_x, d_y\}$ rather than $\max$.

1.3 Notation

Let $\|\cdot\|$ and $\langle\cdot,\cdot\rangle$ denote the Euclidean norm and inner product, respectively. Let $B_d(x,r) \coloneqq \{y \in \mathbb{R}^d : \|y-x\| \le r\}$ denote the closed ball with center x and radius r. We use $\|\cdot\|_{\mathrm{op}}$ and $\|\cdot\|_{\mathrm{F}}$ for the operator and Frobenius norms of matrices, respectively. For a topological space S, $\mathcal{P}(S)$ denotes the class of Borel probability measures on it. For $p \in [1,\infty)$, let $\mathcal{P}_p(\mathbb{R}^d)$ be the space of Borel probability measures with finite p-th absolute moment, i.e., $M_p(\rho) \coloneqq \int_{\mathbb{R}^d} \|x\|^p d\rho(x) < \infty$ for any $\rho \in \mathcal{P}_p(\mathbb{R}^d)$. For a signed Borel measure ρ and a measurable function f, we use the shorthand $\rho f \coloneqq \int f d\rho$, whenever the integral exists. The support of $\rho \in \mathcal{P}(\mathbb{R}^d)$ is $\mathrm{spt}(\rho)$, while its covariance matrix (when exists) is denoted by Σ_ρ. For a sequence of probability measure $(\rho_n)_{n\in\mathbb{N}}$ that weakly converges to ρ, we write $\rho_n \xrightarrow{w} \rho$. Let $C_b(\mathbb{R}^d)$ be the space of bounded continuous functions on $\mathbb{R}^d$ equipped with the L^∞ norm. The Lipschitz seminorm of a function $f : \mathbb{R}^d \to \mathbb{R}$ is $\|f\|_{\mathrm{Lip}} \coloneqq \sup_{x\neq x'} \frac{|f(x)-f(x')|}{\|x-x'\|}$. For $p \in [1,\infty)$ and $\rho \in \mathcal{P}(\mathbb{R}^d)$, let $L^p(\rho)$ be the space of measurable functions f of $\mathbb{R}^d$ such that $\|f\|_{L^p(\rho)} \coloneqq (\int_{\mathbb{R}^d} |f|^p d\rho)^{1/p} < \infty$. with $D^0 f = f$. We write $N(\varepsilon, \mathcal{F}, \mathsf{d})$ for the ε-covering number of a function class $\mathcal{F}$ w.r.t. a metric d, and $N_{[\,]}(\varepsilon, \mathcal{F}, \mathsf{d})$ for the bracketing number. We use $\lesssim_x$ to denote inequalities up to constants that only depend on x; the subscript is dropped when the constant is universal. For $a, b \in \mathbb{R}$, let $a \vee b = \max\{a,b\}$ and $a \wedge b = \min\{a,b\}$.

2 Background and preliminaries

2.1 Classical optimal transport

We briefly review basic definitions and results concerning the classical OT problem, which serve as a building block for our subsequent analysis of the GW distance. For a detailed exposition the reader is referred to [46, 36, 34]. Let $\mathcal{X}, \mathcal{Y}$ be two Polish spaces and consider a lower semicontinuous cost function $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, where note that we allow c to take negative value. The OT problem between $(\mu,\nu) \in \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{Y})$ with cost c is

$$\mathsf{OT}_c(\mu,\nu) \coloneqq \inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{Y}} c\, d\pi, \tag{2}$$

where $\Pi(\mu,\nu)$ is the set of all couplings of μ and ν, i.e., each $\pi \in \Pi(\mu,\nu)$ is a probability distribution on $\mathcal{X}\times\mathcal{Y}$ that has μ and ν as its first and second marginals, respectively. The special case of the p-Wasserstein distance, for $p \in [1,\infty)$, is given by $\mathsf{W}_p(\mu,\nu) \coloneqq \big(\mathsf{OT}_{\|\cdot\|^p}(\mu,\nu)\big)^{1/p}$. W_p is a metric on $\mathcal{P}_p(\mathbb{R}^d)$ which metrizes weak convergence plus convergence of p-th moments, i.e., $\mathsf{W}_p(\hat{\mu}_n,\mu)\to 0$ if and only if $\hat{\mu}_n \xrightarrow{w} \mu$ and $M_p(\hat{\mu}_n)\to M_p(\mu)$.

OT is a linear program and as such it admits strong duality. Suppose that the cost function satisfies $c(x,y)\geq a(x)+b(y)$, for all $(x,y)\in\mathcal{X}\times\mathcal{Y}$, for some upper semicontinuous functions $(a,b)\in L^1(\mu)\times L^1(\nu)$. Then (cf. [46, Theorem 5.10]):

$$\mathsf{OT}_c(\mu,\nu) = \sup_{(\varphi,\psi)\in\Phi_c} \int_{\mathcal{X}} \varphi d\mu + \int_{\mathcal{Y}} \psi d\nu, \tag{3}$$

where $\Phi_c \coloneqq \big\{(\varphi,\psi)\in C_b(\mathcal{X})\times C_b(\mathcal{Y}) : \varphi(x)+\psi(y)\leq c(x,y),\ \forall (x,y)\in\mathcal{X}\times\mathcal{Y}\big\}$. Furthermore, defining the c- and $\bar{c}$-transform of $\varphi\in C_b(\mathcal{X})$ and $\psi\in C_b(\mathcal{Y})$ as $\varphi^c(y)\coloneqq \inf_{x\in\mathcal{X}} c(x,y)-\varphi(x)$ and $\psi^{\bar{c}}(x)\coloneqq \inf_{y\in\mathcal{Y}} c(x,y)-\psi(y)$, respectively, the optimization above can be restricted to pairs (φ,ψ) such that $\psi=\varphi^c$ and $\varphi=\psi^{\bar{c}}$.

2.2 Classical Gromov-Wasserstein distance

The objects of interest in this work is the GW distance. The (p,q)-GW distance quantifies similarity between (complete and separable) mm spaces $(\mathcal{X},\mathsf{d}_\mathcal{X},\mu)$ and $(\mathcal{Y},\mathsf{d}_\mathcal{Y},\nu)$ as [30, 42].

$$\mathsf{D}_{p,q}(\mu,\nu) \coloneqq \inf_{\pi\in\Pi(\mu,\nu)} \big\|\Delta_q^{\mathcal{X},\mathcal{Y}}\big\|_{L^p(\pi\otimes\pi)},$$

where $\Delta_q^{\mathcal{X},\mathcal{Y}}(x,y,x',y') = \big|\mathsf{d}_\mathcal{X}(x,x')^q - \mathsf{d}_\mathcal{Y}(y,y')^q\big|$. This definition is an L^p relaxation of the Gromov-Hausdorff distance between metric spaces,[2] and gives rise to a metric on the collection of all isomorphism classes of mm spaces[3] with finite pq-size, i.e., $\int \mathsf{d}_\mathcal{X}(x,x')^{pq}d\mu\otimes\mu(x,x') < \infty$ and similarly for ν. Like the p-Wasserstein distance, Theorem 5.1 in [30] reveals that $\mathsf{D}_{p,q}$ captures empirical convergence of mm spaces: if $X_1,\dots,X_n$ are samples from $\mu\in\mathcal{P}(\mathcal{X})$ and $\hat{\mu}_n\coloneqq n^{-1}\sum_{i=1}^n \delta_{X_i}$ is their empirical measures, then $\mathsf{D}_{p,q}(\hat{\mu}_n,\mu)\to 0$ a.s. The rate at which this empirical convergence happens is, however, an open problem.

Towards a complete resolution, one of our main contributions is to quantify the empirical convergence rate of the $(2,2)$-GW distance between Euclidean mm spaces $(\mathbb{R}^{d_x},\|\cdot\|,\mu)$ and $(\mathbb{R}^{d_y},\|\cdot\|,\nu)$ of different dimensions. Abbreviating $\Delta_2^{\mathbb{R}^{d_x},\mathbb{R}^{d_y}} = \Delta$, the distance of interest is

$$\begin{aligned}\mathsf{D}(\mu,\nu) &\coloneqq \inf_{\pi\in\Pi(\mu,\nu)} \|\Delta\|_{L^2(\pi\otimes\pi)} \\ &= \inf_{\pi\in\Pi(\mu,\nu)} \left(\int_{\mathbb{R}^{d_x}\times\mathbb{R}^{d_y}}\int_{\mathbb{R}^{d_x}\times\mathbb{R}^{d_y}} \big|\|x-x'\|^2-\|y-y'\|^2\big|^2 d\pi\otimes\pi(x,y,x',y')\right)^{\frac{1}{2}}. \end{aligned}\tag{4}$$

We drop subscripts from our notation because we focus on the $(2,2)$-GW case from here on out. For finiteness we will always assume $\mu\in\mathcal{P}_4(\mathbb{R}^{d_x})$ and $\nu\in\mathcal{P}_4(\mathbb{R}^{d_y})$. Another major gap in GW theory is the lack of dual formulation, without which an empirical convergence rate analysis of GW distances remained obscure. In what follows, we close this gap.

3 Gromov-Wasserstein distance

We now consider the $(2,2)$-GW distance from (4), establish duality, derive its sample complexity, and study its one-dimensional structure.

[2] The Gromov-Hausdorff distance between $(\mathcal{X},\mathsf{d}_\mathcal{X})$ and $(\mathcal{Y},\mathsf{d}_\mathcal{Y})$ is given by $\frac{1}{2}\inf_{R\in\mathcal{R}(\mathcal{X},\mathcal{Y})}\|\Delta_{1,1}^{\mathcal{X},\mathcal{Y}}\|_{L^\infty(R)}$, where $\mathcal{R}(\mathcal{X},\mathcal{Y})$ is the collection of all correspondence sets of $\mathcal{X}$ and $\mathcal{Y}$, i.e., subsets $R\subset\mathcal{X}\times\mathcal{Y}$ such that the coordinate projection maps are surjective when restricted to R. The correspondence set can be thought of as $\mathrm{spt}(\pi)$ in the GW formulation.

[3] The mm spaces $(\mathcal{X},\mathsf{d}_\mathcal{X},\mu)$ and $(\mathcal{Y},\mathsf{d}_\mathcal{Y},\nu)$ are isomorphic if there is an isometry $f:\mathcal{X}\to\mathcal{Y}$ with $f_\sharp\mu=\nu$.

3.1 Duality

We first derive a dual formulation for the GW distance. This duality serves as the key component for our sample complexity analysis of empirical GW in the next subsection. Let $(\mu,\nu)\in\mathcal{P}_4(\mathbb{R}^{d_x})\times\mathcal{P}_4(\mathbb{R}^{d_y})$. Towards the dual form, first observe that D is invariant to isometric operations on the marginal spaces, such as translation and orthonormal rotation. Thus, without loss of generality (w.l.o.g.), we assume that μ and ν are centered, i.e., $\int x d\mu(x)=\int y d\nu(y)=0$.

Next, by expanding the $(2,2)$-GW cost, we split the GW functional into two terms as

$$\mathsf{D}(\mu,\nu)^2=\mathsf{S}^1(\mu,\nu)+\mathsf{S}^2(\mu,\nu), \tag{5}$$

where

$$\mathsf{S}^1(\mu,\nu):=\int\|x-x'\|^4 d\mu\otimes\mu(x,x')+\int\|y-y'\|^4 d\nu\otimes\nu(y,y')-4\int\|x\|^2\|y\|^2 d\mu\otimes\nu(x,y)$$
$$\mathsf{S}^2(\mu,\nu):=\inf_{\pi\in\Pi(\mu,\nu)}\int -4\|x\|^2\|y\|^2 d\pi(x,y)-8\sum_{\substack{1\le i\le d_1\\1\le j\le d_2}}\Big(\int x_i y_j d\pi(x,y)\Big)^2.$$

Evidently, the first term depends only on the marginals μ,ν, while the second captures the dependence on the coupling π. The following theorem (proved in Appendix A.1) establishes duality for $\mathsf{S}^2(\mu,\nu)$, which, in turn, yields a dual form for $\mathsf{D}(\mu,\nu)^2$ via the above decomposition.

Theorem 3.1 (GW duality)**.** *Let $(\mu,\nu)\in\mathcal{P}_4(\mathbb{R}^{d_x})\times\mathcal{P}_4(\mathbb{R}^{d_y})$ and define $M_{\mu,\nu}:=\sqrt{M_2(\mu)M_2(\nu)}$. We have*

$$\mathsf{S}^2(\mu,\nu)=\inf_{\mathbf{A}\in\mathbb{R}^{d_x\times d_y}}32\|\mathbf{A}\|_{\mathrm{F}}^2+\mathsf{OT}_{\mathbf{A}}(\mu,\nu), \tag{6}$$

where $\mathsf{OT}_{\mathbf{A}}$ is the OT problem with cost function $c_{\mathbf{A}}:(x,y)\in\mathbb{R}^{d_x}\times\mathbb{R}^{d_y}\mapsto -4\|x\|^2\|y\|^2-32x^{\intercal}\mathbf{A}y$. Moreover, the infimum is achieved at some $\mathbf{A}^\star\in\mathcal{D}_{M_{\mu,\nu}}:=[-M_{\mu,\nu}/2,M_{\mu,\nu}/2]^{d_x\times d_y}$.

Note that the optimization in $\mathbf{A}$ is unconstrained, and the optimum is guaranteed to exist in a simple bounded domain $\mathcal{D}_{M_{\mu,\nu}}$, which will prove crucial for the proof of our sample complexity. The variational representation above relates the GW to the well understood problem of OT. This enables leveraging knowledge on the latter to make progress in the study of GW. In particular, this representation unlocks our sample complexity analysis, which relies on inserting the OT dual from (3) into the above. Since (6) allows utilizing OT duality for the GW analysis, we synonymously refer to it as the GW dual (even though it is somewhat of a misnomer, since strictly speaking, (6) is not a dual problem for $\mathsf{D}(\mu,\nu)^2$ in the standard optimization theory sense). We note that a similar result was also discovered independently in [44] Theorem 4.2.5, with similar derivation but different focuses. The employed equivalence analysis therein can be generalized to arbitrary concave programming, while our focus is more on the reformulation of GW and analysis of the optimal matrix.

3.2 Sample complexity

Given the dual form for $\mathsf{D}(\mu,\nu)^2$ we proceed with a sample complexity analysis. We focus on compactly supported distributions and refer the reader to Remark 1 ahead for a discussion on extensions to unbounded domains. The following theorem gives a sharp characterization of the one- and two-sample empirical convergence rate of the quadratic GW distance.

Theorem 3.2 (GW sample complexity)**.** *Let $(\mu,\nu)\in\mathcal{P}(\mathcal{X})\times\mathcal{P}(\mathcal{Y})$, where $\mathcal{X}\subset\mathbb{R}^{d_x}$ and $\mathcal{Y}\subset\mathbb{R}^{d_y}$ are compact, and let $R=\mathrm{diam}(\mathcal{X})\vee\mathrm{diam}(\mathcal{Y})$. We have*

$$\mathbb{E}\big[\big|\mathsf{D}(\mu,\nu)^2-\mathsf{D}(\hat{\mu}_n,\nu)^2\big|\big]\lesssim_{d_x,d_y}\frac{R^4}{\sqrt{n}}+\big(1+R^4\big)n^{-\frac{2}{(d_x\wedge d_y)\vee 4}}(\log n)^{\mathbb{1}\{d_x\wedge d_y=4\}}$$
$$\mathbb{E}\big[\big|\mathsf{D}(\mu,\nu)^2-\mathsf{D}(\hat{\mu}_n,\hat{\nu}_n)^2\big|\big]\lesssim_{d_x,d_y}\frac{R^4}{\sqrt{n}}+\big(1+R^4\big)n^{-\frac{2}{(d_x\wedge d_y)\vee 4}}(\log n)^{\mathbb{1}\{d_x\wedge d_y=4\}},$$

and if μ,ν are separated in the $(2,2)$-GW distance, i.e., $\mathsf{D}(\mu,\nu)>0$, then the same rates hold for estimating D itself, without the square.

Furthermore, the above rates are sharp in the sense that for any n large enough, we have

$$\sup_{(\mu,\nu)\in\mathcal{P}(\mathcal{X})\times\mathcal{P}(\mathcal{Y})} \mathbb{E}\big[\big|\mathsf{D}(\mu,\nu)^2 - \mathsf{D}(\hat{\mu}_n,\nu)^2\big|\big] \gtrsim_{d_x,d_y,R} n^{-\frac{2}{(d_x\wedge d_y)\vee 4}}$$

$$\sup_{(\mu,\nu)\in\mathcal{P}(\mathcal{X})\times\mathcal{P}(\mathcal{Y})} \mathbb{E}\big[\big|\mathsf{D}(\mu,\nu)^2 - \mathsf{D}(\hat{\mu}_n,\hat{\nu}_n)^2\big|\big] \gtrsim_{d_x,d_y,R} n^{-\frac{2}{(d_x\wedge d_y)\vee 4}}.$$

Theorem 3.2 is proven in Appendix A.2. The upper bounds leverage the duality from Theorem 3.1 to reduce the empirical estimation analysis of D^2 to that of the OT problem with cost $c_{\mathbf{A}}$. The OT estimation error is then bounded by the suprema of empirical processes indexed by dual OT potentials. To control the corresponding entropy integrals, we exploit smoothness of our cost as well as Lipschitzness and convexity of optimal potentials as c-transforms of each other. The fact that the two-sample convergence rate adapts to the smaller dimension is a consequence of the LCA principle [22, Lemma 2.1], whereby the L^∞ covering number of a function class $\mathcal{F}$ is no less than that of its c-transform $\mathcal{F}^c$. This observation enables adapting the bound to the class of dual potentials over the lower-dimensional space. Still, when the estimated measure(s) are high-dimensional, both the one- and two-sample rates for the GW distance suffer from the curse of dimensionality. This is expected and is in line with empirical convergence rates for OT; see Remark 1 ahead for further discussion on the comparison between the empirical rates for GW and OT.

To prove the lower bound, we present a reduction from GW distance estimation to that of the 2-Wasserstein procrustes $\inf_{\mathbf{U}\in E(d)} \mathsf{W}_2(\mu,\mathbf{U}_\sharp\nu)$, where $E(d)$ is the isometry group on $\mathbb{R}^d$ [18] (see also [38, 17]). This relies on the following lemma, which may be of independent interest. We state two-sided bounds, but only the lower bound is used in the derivation.

Lemma 3.3 (GW vs. W-procrustes)**.** *For any $p,q\in[1,\infty)$ and $\mu,\nu\in\mathcal{P}_{pq}(\mathbb{R}^d)$, we have*

$$\mathsf{D}_{p,q}(\mu,\nu) \le q^p\, 2^{pq+p-1+1/q}\big(M_{pq}(\mu)+M_{pq}(\nu)\big)^{\frac{q-1}{pq}}\,\mathsf{W}_{pq}(\mu,\nu).$$

Furthermore, for $p=q=2$, if μ and ν have covariance matrices $\mathbf{\Sigma}_\mu$ and $\mathbf{\Sigma}_\nu$ with full rank and smallest eigenvalues $\lambda_{\min}(\mathbf{\Sigma}_\mu)$ and $\lambda_{\min}(\mathbf{\Sigma}_\nu)$, respectively, then

$$\Big(32\big(\lambda_{\min}(\mathbf{\Sigma}_\mu)^2+\lambda_{\min}(\mathbf{\Sigma}_\nu)^2\big)\Big)^{\frac{1}{4}} \inf_{\mathbf{U}\in E(d)} \mathsf{W}_2(\mu,\mathbf{U}_\sharp\nu) \le \mathsf{D}(\mu,\nu).$$

If μ and ν are also centered, then it suffices to optimize only over the orthogonal group $O(d)$.

The lemma enables showing that the empirical GW rate, when the population measures are uniform over the unit ball and its scaled version, is at least as large as that of the Wasserstein procrustes. We then develop a new lower bound on the convergence rate of the latter, showing that it is at least $n^{-1/d}$. This, in turn, gives rise to the rates from Theorem 3.2.

Remark 1 (Comparison to OT and unbounded domains). The rates in Theorem 3.2 are inline with those for the classical OT problem with Hölder smooth costs [26] (although our analysis is different from theirs). Over compact domains, this smoothness of the cost enables establishing global Lipschitzness and convexity of OT potentials, which, in turn, leads to the quadratic improvement from the standard $n^{-1/d}$ empirical convergence rate to $n^{-2/d}$, when $d>4$. Evidently, a similar phenomenon happens in the GW case. Unbounded domains are treated in Theorem 13 of [26], but this result relies on restrictive assumptions on the population distributions and the cost. Namely, the distributions must satisfy certain high-level concentration and anti-concentration conditions, while the cost must be locally Hölder smooth and be lower and upper bounded by a polynomial of appropriate degree. Our cost $c_{\mathbf{A}}$ does not immediately adhere to these assumptions. While we believe that the argument can be adapted, we leave this extension as a question for future work.

3.3 One-dimensional case study

We leverage our duality theory to shed new light on the one-dimensional GW distance. The solution to the GW problem between distributions on $\mathbb{R}$ is currently unknown and remains one of the most basic open questions in that space. While the standard p-Wasserstein distance between distributions on $\mathbb{R}$ is given by the $L^p([0,1])$ distance between their quantile functions,[4] there is no known simple solution

[4]For $p=1$, the formula further simplifies to the $L^1(\mathbb{R})$ distance between the cumulative distribution functions.

for the one-dimensional GW problem. Even for uniform distributions over n distinct points, for which it was previously believed that the optimal GW coupling is always induced by the identity or anti-identity permutations [45], it was recently shown that this is not true in general [2] (see also discussion in [12]). Indeed, [2] produced an example of discrete distributions, defined up to a tuning parameter ξ, for which the identity or anti-identity become suboptimal once ξ surpasses a certain threshold. We revisit this example and attempt to better understand it using our dual formulation. Consider two uniform distributions on n distinct points, i.e., $\mu = n^{-1}\sum_{i=1}^{n}\delta_{x_i}$ and $\nu = n^{-1}\sum_{i=1}\delta_{y_i}$, where $(x_i)_{i=1}^n, (y_i)_{i=1}^n \subset \mathbb{R}$ with $x_1 < x_2 < \ldots < x_n$ and $y_1 < y_2 < \ldots < y_n$. To compute $\mathsf{D}(\mu,\nu)$ it suffices to optimize over couplings induced by permutations [45, Theorem 9.2] (see also [28]), i.e.,

$$\mathsf{D}(\mu,\nu)^2 = \frac{1}{n^2}\min_{\sigma\in S_n}\sum_{i=1}^{n}\sum_{j=1}^{n}\left||x_i - x_j|^2 - |y_{\sigma(i)} - y_{\sigma(j)}|^2\right|^2, \tag{7}$$

where S_n is the symmetric group over n elements. For $\xi \in (0, 2/(n-3))$ and $n > 6$, define the point sets $x^\xi = (x_i^\xi)_{i=1}^n$ and $y^\xi = (y_i^\xi)_{i=1}^n$ as

$$x_i^\xi := \begin{cases} -1, & i = 1 \\ \frac{2i-n-1}{2}\xi, & 2 \le i \le n-1 \\ 1, & i = n \end{cases} \quad \text{and} \quad y_i^\xi := \begin{cases} -1, & i = 1 \\ -1+\xi, & i = 2 \\ (i-2)\xi, & 3 \le i \le n \end{cases}. \tag{8}$$

Note that each of these sets indeed has ascending ordered, pairwise distinct components. The proof of Proposition 1 in [2] shows that there exists $\xi^\star \in (0, 2/(n-3))$, such that the cyclic permutation $\sigma_{\mathrm{cyc}}(i) = i + 1 \mod n$ between $x^{\xi^\star}$ and $y^{\xi^\star}$ achieves a strictly smaller cost in (7) than both the identity $\mathrm{id}(i) = i$ and the anti-identity $\overline{\mathrm{id}}(i) = n - i + 1$ permutations.

To better understand the reason for the existence of strict optimizers outside the boundary, we recall that $\mathsf{D}(\mu,\nu)^2 = \mathsf{S}^1(\mu,\nu) + \mathsf{S}^2(\mu,\nu)$ and henceforth focus on $\mathsf{S}^2(\mu,\nu)$, which is the term that depends on the coupling. As mentioned before, this decomposition requires μ and ν to be centered, but we may assume this w.l.o.g. due the translation invariance of the GW-distance and of optimal permutations. By Theorem 3.1 we have the following representation:

$$\mathsf{S}^2(\mu,\nu) = \inf_{\mathbf{A}\in\mathcal{D}_M} 32\|\mathbf{A}\|_{\mathrm{F}}^2 + \inf_{\pi\in\Pi(\mu,\nu)}\int c_{\mathbf{A}}(x,y)d\pi(x,y).$$

Specializing to the one-dimensional case, we further obtain

$$\mathsf{S}^2(\mu,\nu) = \inf_{a\in[0.5W_-,0.5W_+]} 32a^2 + \inf_{\pi\in\Pi(\mu,\nu)}\int \big(-4x^2y^2 - 32axy\big)d\pi(x,y), \tag{9}$$

where $W_- := \inf_{\pi\in\Pi(\mu,\nu)}\int xyd\pi(x,y)$ and $W_+ := \sup_{\pi\in\Pi(\mu,\nu)}\int xyd\pi(x,y)$. Here, we have used the fact that, switching the infima order, for each $\pi \in \Pi(\mu,\nu)$, optimality is attained at $a^\star(\pi) = \frac{1}{2}\int xyd\pi(x,y)$. The notation W_- and W_+ reflects the relation to the 2-Wasserstein distance: indeed, $2W_+ = M_2(\mu) + M_2(\nu) - \mathsf{W}_2^2(\mu,\nu)$, while W_- is OT with product cost.

Once we identify the optimal $a^\star$ in (9), the GW problem is reduced to an OT problem. Hence, we investigate the optimization in a. Define $f(a) := 32a^2$ and $g(a) := \inf_{\pi\in\Pi(\mu,\nu)}\int\big(-4x^2y^2 - 32axy\big)d\pi(x,y)$, and note that g is concave (as the infimum of affine functions). We see that the optimization over a in (9), which is rewritten as $\inf_{a\in[0.5W_-,0.5W_+]}(f+g)(a)$, minimizes the sum of a convex and a concave function. The next proposition identifies a correspondence between the boundary values of a and optimal permutations in (7); see Appendix D for the proof.

Proposition 1 (Boundary values and optimal permutations). Consider the GW problem from (7) between uniform distributions over n distinct points and its representation as $\mathsf{D}(\mu,\nu)^2 = \mathsf{S}^1(\mu,\nu) + \mathsf{S}^2(\mu,\nu)$, where $\mathsf{S}^2(\mu,\nu)$ is given in (9). Let $\mathcal{S}^\star \subset S_n$ and $\mathcal{A}^\star \subset [0.5W_-, 0.5W_+]$ be the argmin sets for (7) and (9), respectively. Then $\mathcal{A}^\star \subset \{0.5W_-, 0.5W_+\}$ if and only if $\mathcal{S}^\star \subset \{\mathrm{id}, \overline{\mathrm{id}}\}$.

Proposition 1 thus implies that the identity and anti-identity can only optimize the GW distance when (9) achieves its minimum on the boundary. However, as f is convex and g is concave, it is not necessarily the case that $\mathcal{A}^\star$ contains only boundary points, as other values may be optimal. To visualize this behavior, Fig. 1 plots the two datasets x^ξ and y^ξ from (8) and the corresponding f, g, and $f+g$ functions for different ξ values. While the infimum is achieved at the boundaries for $\xi = 0.06$ and $\xi = 0.03$, when $\xi = 0.01$ the optimizing $a^\star \approx 0$ and, by Proposition 1, the optimal permutation

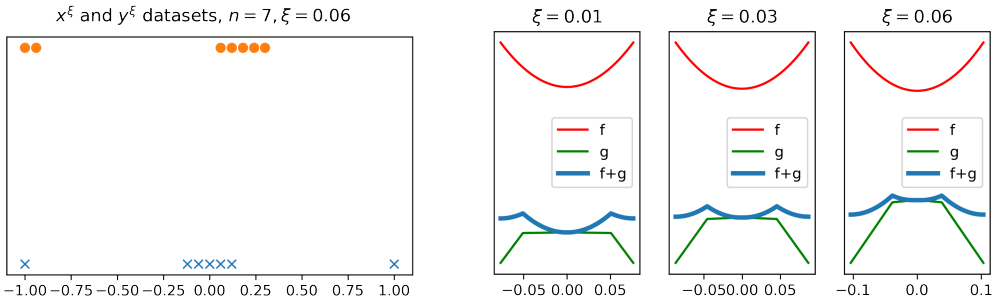

Figure 1: (Left) The datasets x^ξ and y^ξ from (8), for $n = 7$ and $\xi = 0.06$; (Right) The functions f, g, and $f + g$ on the interval $a \in [0.5W_-, 0.5W_+]$, for $\xi = 0.01,\ 0.03,\ 0.06$. When $\xi = 0.01$, the minimizer of $f + g$ is attained outside the boundary and thus the corresponding optimal permutation is neither the identity nor the anti-identity.

is different from id and $\overline{\mathrm{id}}$. The structure of the corresponding optimal coupling is not trivial, as already seen from the proof of Proposition 1 from [2]. Better understanding the relation between optimal a values and their corresponding couplings is an interesting research avenue. Nevertheless, the above clarifies the optimization structure of the one-dimensional GW problem and provides a visual argument for the suboptimality of id and $\overline{\mathrm{id}}$ in the example above.

4 Outlook and concluding remarks

This paper established a dual formulation for the $(2, 2)$-GW distance, between distributions supported on Euclidean spaces of different dimensions d_x and d_y. The dual forms represented GW as infima of a class of OT problems, indexed by a $d_x \times d_y$ auxiliary matrix with bounded entries, which specified the associated cost function. This connection to the well-understood OT problem enabled lifting analysis techniques from statistical OT to establish, for the first time, sharp empirical convergences rates for GW. The derived two-sample rate is $n^{-2/((d_x \wedge d_y) \vee 4)}$ (up to a log factor when $d_x \wedge d_y = 4$) for GW. The GW result accounts for compactly supported distributions, and provides matching upper and lower rate bound. These results are in line with the empirical convergence rates of OT [26, 22].

Lastly, we reexamined the open problem of the one-dimensional GW distance between discrete distributions on n points. Leveraging our duality, we shed new light on the peculiar example from [2], that showed that the identity and anti-identity permutations are not necessarily optimal. Specifically, the dual form allows representing the GW distance as a sum of concave and convex functions, illuminating that, in certain regimes, the optimum is not necessarily attained on the boundary.

Future research directions stemming from this work are aplenty. Due to the central role of duality for statistical and algorithmic advancements, a first key objective is to extend our duality theory beyond the $(2, 2)$-cost and to non-Euclidean mm spaces. While our techniques are rather specialized for the $(2, 2)$-cost and treating arbitrary (p, q) values may require new ideas, we comment here on one relatively direct extension. Consider the GW distance of order $(p, q) = (2, 2k)$, for some $k \in \mathbb{N}$, between distributions $(\mu, \nu) \in \mathcal{P}_{4k}(\mathbb{R}^{d_x}) \times \mathcal{P}_{4k}(\mathbb{R}^{d_y})$ (in fact, we can treat any even p parameter as well, but restrict to $p = 2$ for simplicity). Following a decomposition along the lines of (11), in Appendix E we show that

$$\mathsf{D}_{2,2k}(\mu, \nu)^2 = 4 \sup_{a \in \mathbb{R}^\ell} \inf_{b \in \mathbb{R}^{m-\ell}} \left\{ -\|a\|^2 + \|b\|^2 + \inf_{\pi \in \Pi(\mu,\nu)} \int c_{a,b}(x, y) d\pi(x, y) \right\}, \tag{10}$$

where notice that the inner optimization over π specifies an OT problem with cost $c_{a,b}$ where

$$c_{a,b} : (x, y) \mapsto -\|x\|^{2k}\|y\|^{2k} + \sum_{i=1}^{\ell} a_i g_i(x, y) - \sum_{i=\ell+1}^{m} b_{i-\ell}\, g_i(x, y),$$

$g_1, \ldots, g_m$ are polynomials of degree at most $4k$, m corresponds to the number of polynomials emerging from the quadratic expansion of the $(2, 2k)$-cost, and $\ell \leq m$ is determined by a certain

diagonalization argument (see Appendix E for the specifics). One may further show that $\int g_i d\pi$ are uniformly bounded for all $i = 1, \dots, m$ and $\pi \in \Pi(\mu, \nu)$, and so we may restrict optimization over a, b to bounded domains. In the appendix, we also show how the above dual reduces to the one from Theorem 3.1 once we set $k = 1$ and assume that μ, ν are centered. Also notice now that $c_{a,b}$ smooth (indeed, a polynomial) but not necessarily convex in x or y. For the standard $(2, 2k)$-GW distance between compactly supported distributions, an argument similar to the proof of Theorem 3.2, would result in a two-sample convergence rate of $O\big(n^{-1/(d_x \wedge d_y)}\big)$. This rate stems from the fact that the corresponding dual potentials are Lipschitz continuous, but it is unclear whether they posses further convexity/concavity properties. In sum, while a duality theory for general (p, q) remains an open question, our results for the quadratic GW distance can be extended to cover any even q value.

As mentioned above, extending our duality to cover non-Euclidean mm spaces is of great interest, as this would enable accounting for graph and manifold data modalities. We also believe that our dual can be used to derive new and efficient algorithms for computing the GW and EGW distances. Lastly, we mention the avenue of generalizing the GW empirical convergence results to distributions with unbounded supports. Identifying sufficient conditions for deriving explicit rates seems non-trivial and may require assumptions along the lines of Theorem 13 from [26], where empirical convergence of OT on unbounded domains was treated.

Acknowledgments and Disclosure of Funding

Z. Goldfeld is partially supported by the NSF CAREER award under Grant CCF-2046018 and NSF Grant DMS-2210368. B. K. Sriperumbudur is partially supported by the NSF CAREER award under Grant DMS-1945396 and NSF Grant DMS-19453.

References

[1] David Alvarez-Melis and Tommi Jaakkola. Gromov-wasserstein alignment of word embedding spaces. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1881–1890. Association for Computational Linguistics, 2018. URL `http://aclweb.org/anthology/D18-1214`.

[2] Robert Beinert, Cosmas Heiss, and Gabriele Steidl. On assignment problems related to Gromov-Wasserstein distances on the real line. *arXiv preprint arXiv:2205.09006*, 2022.

[3] Jérémie Bigot, Elsa Cazelles, and Nicolas Papadakis. Central limit theorems for entropy-regularized optimal transport on finite spaces and statistical applications. *Electronic Journal of Statistics*, 13(2):5120–5150, 2019.

[4] Andrew J Blumberg, Mathieu Carriere, Michael A Mandell, Raul Rabadan, and Soledad Villar. MREC: a fast and versatile framework for aligning and matching point clouds with applications to single cell molecular data. *arXiv preprint arXiv:2001.01666*, 2020.

[5] Lenaic Chizat, Pierre Roussillon, Flavien Léger, François-Xavier Vialard, and Gabriel Peyré. Faster Wasserstein distance estimation with the Sinkhorn divergence. *Advances in Neural Information Processing Systems*, 33:2257–2269, 2020.

[6] Clayton W. Commander. A survey of the quadratic assignment problem, with applications. *Morehead Electronic Journal of Applicable Mathematics*, 4:MATH–2005–01, 2005.

[7] Eustasio del Barrio and Jean-Michel Loubes. Central limit theorems for empirical transportation cost in general dimension. *The Annals of Probability*, 47(2):926–951, 2019.

[8] Eustasio del Barrio, Alberto Gonzalez-Sanz, Jean-Michel Loubes, and Jonathan Niles-Weed. An improved central limit theorem and fast convergence rates for entropic transportation costs. *arXiv preprint arXiv:2204.09105*, 2022.

[9] Julie Delon, Agnes Desolneux, and Antoine Salmona. Gromov-Wasserstein distances between Gaussian distributions. *Journal of Applied Probability*, pages 1–21, 2022.

[10] Pinar Demetci, Rebecca Santorella, Björn Sandstede, William Stafford Noble, and Ritambhara Singh. SCOT: single-cell multi-omics alignment with optimal transport. *Journal of Computational Biology*, 29(1):3–18, 2022.

[11] Richard Mansfield Dudley. The speed of mean Glivenko-Cantelli convergence. *The Annals of Mathematical Statistics*, 40(1):40–50, 1969.

[12] Theo Dumont, Théo Lacombe, and François-Xavier Vialard. On the existence of Monge maps for the Gromov-Wasserstein distance. *arXiv preprint arXiv:2210.11945*, 2022.

[13] Ziv Goldfeld and Kristjan Greenewald. Gaussian-smoothed optimal transport: Metric structure and statistical efficiency. In *International Conference on Artificial Intelligence and Statistics*, pages 3327–3337. PMLR, 2020.

[14] Ziv Goldfeld, Kengo Kato, Sloan Nietert, and Gabriel Rioux. Limit distribution theory for smooth p-Wasserstein distances. *arXiv preprint arXiv:2203.00159*, 2022.

[15] Ziv Goldfeld, Kengo Kato, Gabriel Rioux, and Ritwik Sadhu. Limit theorems for entropic optimal transport maps and the Sinkhorn divergence. *arXiv preprint arXiv:2207.08683*, 2022.

[16] Ziv Goldfeld, Kengo Kato, Gabriel Rioux, and Ritwik Sadhu. Statistical inference with regularized optimal transport. *arXiv preprint arXiv:2205.04283*, 2022.

[17] Colin Goodall. Procrustes methods in the statistical analysis of shape. *Journal of the Royal Statistical Society: Series B (Methodological)*, 53(2):285–321, 1991.

[18] Edouard Grave, Armand Joulin, and Quentin Berthet. Unsupervised alignment of embeddings with Wasserstein procrustes. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1880–1890. PMLR, 2019.

[19] Michel Groppe and Shayan Hundrieser. Lower complexity adaptation for empirical entropic optimal transport. *arXiv preprint arXiv:2306.13580*, 2023.

[20] Adityanand Guntuboyina and Bodhisattva Sen. l_1 covering numbers for uniformly bounded convex functions. In *Conference on Learning Theory*, pages 12–1. JMLR Workshop and Conference Proceedings, 2012.

[21] Shayan Hundrieser, Marcel Klatt, Thomas Staudt, and Axel Munk. A unifying approach to distributional limits for empirical optimal transport. *arXiv preprint: arXiv 2202.12790*, 2022.

[22] Shayan Hundrieser, Thomas Staudt, and Axel Munk. Empirical optimal transport between different measures adapts to lower complexity. *arXiv preprint arXiv:2202.10434*, 2022.

[23] Marcel Klatt, Carla Tameling, and Axel Munk. Empirical regularized optimal transport: Statistical theory and applications. *SIAM Journal on Mathematics of Data Science*, 2(2): 419–443, 2020.

[24] Patrice Koehl, Marc Delarue, and Henri Orland. Computing the Gromov-Wasserstein distance between two surface meshes using optimal transport. *Algorithms*, 16(3):131, 2023.

[25] Khang Le, Dung Q Le, Huy Nguyen, Dat Do, Tung Pham, and Nhat Ho. Entropic Gromov-Wasserstein between Gaussian distributions. In *International Conference on Machine Learning*, pages 12164–12203. PMLR, 2022.

[26] Tudor Manole and Jonathan Niles-Weed. Sharp convergence rates for empirical optimal transport with smooth costs. *arXiv preprint arXiv:2106.13181*, 2021.

[27] Tudor Manole, Sivaraman Balakrishnan, Jonathan Niles-Weed, and Larry Wasserman. Plugin estimation of smooth optimal transport maps. *arXiv preprint arXiv:2107.12364*, 2021.

[28] Haggai Maron and Yaron Lipman. (probably) concave graph matching. *Advances in Neural Information Processing Systems*, 31, 2018.

[29] Facundo Mémoli. Spectral Gromov-Wasserstein distances for shape matching. In *2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops*, pages 256–263. IEEE, 2009.

[30] Facundo Mémoli. Gromov-Wasserstein distances and the metric approach to object matching. *Found. Comput. Math.*, 11(4):417–487, 2011. URL `http://dblp.uni-trier.de/db/journals/focm/focm11.html#Memoli11`.

[31] Gonzalo Mena and Jonathan Niles-Weed. Statistical bounds for entropic optimal transport: sample complexity and the central limit theorem. *Advances in Neural Information Processing Systems*, 32, 2019.

[32] Jonathan Niles-Weed and Philippe Rigollet. Estimation of wasserstein distances in the spiked transport model. *Bernoulli*, 28(4):2663–2688, 2022.

[33] Gabriel Peyré, Marco Cuturi, and Justin Solomon. Gromov-Wasserstein averaging of kernel and distance matrices. In *International Conference on Machine Learning*, pages 2664–2672. PMLR, 2016.

[34] 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.

[35] Gabriel Rioux, Ziv Goldfeld, and Kengo Kato. Entropic gromov-wasserstein distances: Stability, algorithms, and distributional limits. *arXiv preprint arXiv:2306.00182*, 2023.

[36] Filippo Santambrogio. Optimal transport for applied mathematicians. *Birkäuser, NY*, 55(58-63): 94, 2015.

[37] Meyer Scetbon, Gabriel Peyré, and Marco Cuturi. Linear-time Gromov- Wasserstein distances using low rank couplings and costs. In *International Conference on Machine Learning*, pages 19347–19365. PMLR, 2022.

[38] Peter H Schönemann. A generalized solution of the orthogonal procrustes problem. *Psychometrika*, 31(1):1–10, 1966.

[39] Thibault Séjourné, François-Xavier Vialard, and Gabriel Peyré. The unbalanced gromov wasserstein distance: Conic formulation and relaxation. *Advances in Neural Information Processing Systems*, 34:8766–8779, 2021.

[40] Justin Solomon, Gabriel Peyré, Vladimir G Kim, and Suvrit Sra. Entropic metric alignment for correspondence problems. *ACM Transactions on Graphics (ToG)*, 35(4):1–13, 2016.

[41] Max Sommerfeld and Axel Munk. Inference for empirical Wasserstein distances on finite spaces. *Journal of Royal Statistical Society: Series B (Statistical Methodology)*, 80:219–238, 2018.

[42] Karl-Theodor Sturm. The space of spaces: curvature bounds and gradient flows on the space of metric measure spaces. *arXiv preprint arXiv:1208.0434*, 2012.

[43] Carla Tameling, Max Sommerfeld, and Axel Munk. Empirical optimal transport on countable metric spaces: Distributional limits and statistical applications. *The Annals of Applied Probability*, 29:2744–2781, 2019.

[44] Titouan Vayer. A contribution to optimal transport on incomparable spaces. *arXiv preprint arXiv:2011.04447*, 2020.

[45] Titouan Vayer, Rémi Flamary, Romain Tavenard, Laetitia Chapel, and Nicolas Courty. Sliced Gromov-Wasserstein, 2020.

[46] Cédric Villani. *Optimal Transport: old and new*, volume 338. Springer, 2009.

[47] Jonathan Weed and Francis Bach. Sharp asymptotic and finite-sample rates of convergence of empirical measures in wasserstein distance. *Bernoulli*, 25(4 A):2620–2648, 2019.

[48] Hongteng Xu, Dixin Luo, and Lawrence Carin. Scalable Gromov-Wasserstein learning for graph partitioning and matching, 2019.

[49] Hongteng Xu, Dixin Luo, Hongyuan Zha, and Lawrence Carin Duke. Gromov-Wasserstein learning for graph matching and node embedding. In *International conference on machine learning*, pages 6932–6941. PMLR, 2019.

[50] Yuguang Yan, Wen Li, Hanrui Wu, Huaqing Min, Mingkui Tan, and Qingyao Wu. Semi-supervised optimal transport for heterogeneous domain adaptation. In *IJCAI*, volume 7, pages 2969–2975, 2018.

[51] Zhengxin Zhang, Youssef Mroueh, Ziv Goldfeld, and Bharath Sriperumbudur. Cycle consistent probability divergences across different spaces. In *International Conference on Artificial Intelligence and Statistics*, pages 7257–7285. PMLR, 2022.

A Proofs of Main Theorem

A.1 Proof of Theorem 3.1

For completeness we first show the decomposition of $\mathsf{D}(\mu,\nu)$ for centered μ,ν, given in (5). Expanding the $(2,2)$-GW cost we have

$$\begin{aligned}\mathsf{D}(\mu,\nu)^2 = &\int \|x-x'\|^4 d\mu\otimes\mu(x,x') + \int \|y-y'\|^4 d\nu\otimes\nu(y,y') - 4\int \|x\|^2\|y\|^2 d\mu\otimes\nu(x,y)\\ &+ \inf_{\pi\in\Pi(\mu,\nu)}\Bigg\{ -4\int \|x\|^2\|y\|^2 d\pi(x,y) - 8\int \langle x,x'\rangle\langle y,y'\rangle d\pi\otimes\pi(x,y,x',y')\\ &\qquad + 8\int \big(\langle x,x'\rangle\|y\|^2 + \|x\|^2\langle y,y'\rangle\big) d\pi\otimes\pi(x,y,x',y')\Bigg\}. \end{aligned} \tag{11}$$

By the centering assumption, the term in the last line nullifies, while the first and second lines on the RHS correspond to $\mathsf{S}_1(\mu,\nu)$ and $\mathsf{S}^2(\mu,\nu)$, respectively.

We now move to derive the dual form for S^2. Recall that $M_{\mu,\nu} \coloneqq \sqrt{M_2(\mu)M_2(\nu)}$, $\mathcal{D}_{M_{\mu,\nu}} \coloneqq [-M_{\mu,\nu}/2, M_{\mu,\nu}/2]^{d_x\times d_y}$. Consider:

$$\begin{aligned}\mathsf{S}^2_\varepsilon(\mu,\nu) &= \inf_{\pi\in\Pi(\mu,\nu)} \int -4\|x\|^2\|y\|^2 d\pi(x,y) - 8\sum_{\substack{1\le i\le d_x\\ 1\le j\le d_y}} \Big(\int x_i y_j d\pi(x,y)\Big)^2\\ &= \inf_{\pi\in\Pi(\mu,\nu)} \int -4\|x\|^2\|y\|^2 d\pi(x,y) + \sum_{\substack{1\le i\le d_x\\ 1\le j\le d_y}} \inf_{|a_{ij}|\le \frac{M_{\mu,\nu}}{2}} 32\left(a_{ij}^2 - \int a_{ij}x_i y_j d\pi(x,y)\right)\\ &= \inf_{\mathbf{A}\in\mathcal{D}_{M_{\mu,\nu}}}\inf_{\pi\in\Pi(\mu,\nu)} \int -4\|x\|^2\|y\|^2 d\pi(x,y) + \sum_{\substack{1\le i\le d_x\\ 1\le j\le d_y}} 32\left(a_{ij}^2 - \int a_{ij}x_i y_j d\pi(x,y)\right)\\ &= \inf_{\mathbf{A}\in\mathcal{D}_{M_{\mu,\nu}}} 32\|\mathbf{A}\|_{\mathrm{F}}^2 + \inf_{\pi\in\Pi(\mu,\nu)} \int c_{\mathbf{A}}(x,y) d\pi(x,y)\end{aligned}$$

where in the second step we introduced a_{ij} whose optimum is achieved at $\frac{1}{2}\int x_i y_j d\pi(x,y)$. This means we may restrict the optimization to $\mathcal{D}_{M_{\mu,\nu}}$ without affecting the value since $\int x_i y_j d\pi(x,y) \le M_{\mu,\nu}$ by the Cauchy–Schwarz inequality. We also switched the order of the two $\inf$ and claimed that the optimums are achieved, which follows from the lower semicontinuity in π and $\mathbf{A}$. We conclude by identifying the OT problem $\mathsf{OT}_{\mathbf{A}}$ in the last line.

A.2 Proof of Theorem 3.2

A.2.1 Upper bounds

We maintain our convention of suppressing the subscript $\mathbf{A}$ from our notation for optimal dual potentials for the OT problem with cost $c_{\mathbf{A}}$, simply writing (φ,ψ). For simplicity we only prove the two-sample case. The one-sample result follows similarly. Derivations of technical lemmas stated throughout this proof are deferred to Appendix C.

Assume w.l.o.g. that μ,ν are centered and recall that we have the decomposition $\mathsf{D}(\mu,\nu)^2 = \mathsf{S}^1(\mu,\nu) + \mathsf{S}^2(\mu,\nu)$. To split our sample complexity analysis into those of S^1 and S^2, we need to account for the fact that empirical measures are generally not centered. Let $\tilde{\mu}_n$ and $\tilde{\nu}_n$ be centered versions of the empirical measures $\hat{\mu}_n$ and $\hat{\nu}_n$, respectively.

This decomposition is convenient for analysis as it allows separately treating the marginals- and the coupling-dependents terms. However, while the EGW distance S_ε is translation invariant and we may assume μ,ν are both centered w.l.o.g., the empirical measures $\hat{\mu}_n,\hat{\nu}_n$ are generally not centered and the decomposition into S^1 and S^2 may not hold. To amend this, we center $\hat{\mu}_n,\hat{\nu}_n$ and quantify the bias that this incurs on D. This is stated in the following lemma, which is proven in Appendix C.1

Lemma A.1 (Centering bias). *If μ, ν are centered, then*

$$\mathbb{E}\big[\big|\mathsf{D}(\mu,\nu)^2-\mathsf{D}(\hat{\mu}_n,\hat{\nu}_n)^2\big|\big] \le \mathbb{E}\big[\big|\mathsf{S}^1(\mu,\nu)-\mathsf{S}^1(\hat{\mu}_n,\hat{\nu}_n)\big|\big]+\mathbb{E}\big[\big|\mathsf{S}^2(\mu,\nu)-\mathsf{S}^2(\hat{\mu}_n,\hat{\nu}_n)\big|\big]+\frac{R^4}{\sqrt{n}}, \quad (12)$$

Given this decomposition, we proceed to separately treat the empirical errors of S^1 and S^2. The analysis of S^1 reduces to estimating moments of μ, ν, with parametric convergence rate for the error. The following lemma is proven in Appendix C.2.

Lemma A.2 (S^1 parametric rate). *If μ, ν satisfy the conditions of Theorem 3.2, then*

$$\mathbb{E}\big[\big|\mathsf{S}^1(\mu,\nu) - \mathsf{S}^1(\hat{\mu}_n,\hat{\nu}_n)\big|\big] \lesssim \frac{R^4}{\sqrt{n}}.$$

To treat S^2, we start from the variational representation from Theorem 3.1 and choose $M = R^2 \ge M_{\mu,\nu}$, which is evidently feasible. Invoking this result, we obtain

$$\big|\mathsf{S}^2(\mu,\nu) - \mathsf{S}^2(\hat{\mu}_n,\hat{\nu}_n)\big| \le \sup_{\mathbf{A}\in\mathcal{D}_{R^2}} \big|\mathsf{OT}_{\mathbf{A}}(\mu,\nu) - \mathsf{OT}_{\mathbf{A}}(\hat{\mu}_n,\hat{\nu}_n)\big|, \quad (13)$$

and proceed to show that for any $\mathbf{A} \in \mathcal{D}_{R^2}$, corresponding optimal dual potentials can be restricted to concave Lipschitz functions and their c-transforms (w.r.t. the cost function $c_{\mathbf{A}}$).

(i) Smoothness of OT potentials. Let

$$\mathcal{F}_R := \left\{ \varphi : B_{d_x}(0,R) \to \mathbb{R} : \begin{array}{c} \varphi \text{ concave, } \|\varphi\|_\infty \le 1 + 10\big(1+4\sqrt{d_x d_y}\big)R^4, \\ \|\varphi\|_{\mathrm{Lip}} \le 8\big(1+2\sqrt{d_x d_y}\big)R^3 \end{array} \right\}$$

and define $\mathcal{G}_R$ analogously over $B_{d_y}(0,R)$. Recall that the c-transform of $\varphi : \mathbb{R}^{d_x} \to \mathbb{R}$ w.r.t. $c_{\mathbf{A}}$ is a new function $\varphi^c : \mathbb{R}^{d_y} \to \mathbb{R}$, given by $\varphi^c = \inf_{x\in\mathcal{X}} c_{\mathbf{A}}(x,\cdot) - \varphi(x)$. The next lemma allows restricting the set of optimal dual potentials for $\mathsf{OT}_{\mathbf{A}}(\mu,\nu)$ to pairs $(\varphi,\varphi^c) \in \mathcal{F}_R \times \mathcal{G}_R$.

Lemma A.3 (Uniform regularity of OT potentials). *Fix $R > 0$ and suppose that $(\mu,\nu) \in \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{Y})$, with $\mathcal{X} \subset B_{d_x}(0,R)$ and $\mathcal{Y} \subset B_{d_y}(0,R)$. Then, for any $\mathbf{A} \in \mathcal{D}_{R^2}$, there exist $\varphi \in \mathcal{F}_R$ with $\varphi^c \in \mathcal{G}_R$, such that (φ,φ^c) is a pair of optimal dual potentials for $\mathsf{OT}_{\mathbf{A}}(\mu,\nu)$.*

The proof, which is given in Appendix C.3, arrives at the above properties by exploiting concavity of $c_{\mathbf{A}}$ and the c-transform representation of optimal dual pairs.

(ii) Sample complexity analysis. Equipped with Lemma A.3, we are ready to conduct the sample complexity analysis. Suppose w.l.o.g. that $d_x \le d_y$; otherwise, flip their roles in the derivation below. For each $\mathbf{A} \in \mathcal{D}_{R^2}$, let $\Phi_{\mathbf{A}}$ be the class of of optimal dual potential pairs for $\mathsf{OT}_{\mathbf{A}}(\mu,\nu)$ (see (3)). Define $\mathcal{F}_{\mathbf{A}} := \mathrm{proj}_{\mathcal{F}_R}\big(\Phi_{\mathbf{A}} \cap (\mathcal{F}_R \times \mathcal{G}_R)\big)$ and let $\mathcal{F}_{\mathbf{A}}^c$ be its c-transform w.r.t. $c_{\mathbf{A}}$. We may now further upper bound the RHS of (13), to arrive at

$$\mathbb{E}\Big[\big|\mathsf{S}^2(\mu,\nu) - \mathsf{S}^2(\hat{\mu}_n,\hat{\nu}_n)\big|\Big] \le \mathbb{E}\left[\sup_{\varphi\in\cup_{\mathbf{A}}\mathcal{F}_{\mathbf{A}}} \big|(\mu-\hat{\mu}_n)\varphi\big|\right] + \mathbb{E}\left[\sup_{\psi\in\cup_{\mathbf{A}}\mathcal{F}_{\mathbf{A}}^c} \big|(\nu-\hat{\nu}_n)\psi\big|\right]. \quad (14)$$

As Lemma A.3 implies that $\cup_{\mathbf{A}}\mathcal{F}_{\mathbf{A}} \subset \mathcal{F}_R$, the first term above is controlled by the expected supremum of an empirical process indexed by $\mathcal{F}_R$. Dudley's entropy integral formula yields

$$\mathbb{E}\left[\sup_{\varphi\in\mathcal{F}_R} \big|(\mu-\hat{\mu}_n)\varphi\big|\right] \lesssim \inf_{\alpha>0} \alpha + \frac{1}{\sqrt{n}} \int_\alpha^{2\sup_{\varphi\in\mathcal{F}_R}\|\varphi\|_\infty} \sqrt{\log N(\xi,\mathcal{F}_R,\|\cdot\|_\infty)}\, d\xi.$$

Theorem 1 from [20] provides a bound on the metric entropy of bounded, convex, Lipschitz functions, whereby if $\widetilde{\mathcal{F}}_d := \{f : B_d(0,1) \to \mathbb{R} : f \text{ convex}, \|f\|_\infty \vee \|f\|_{\mathrm{Lip}} \le 1\}$, then $\log N(\xi,\widetilde{\mathcal{F}}_d,\|\cdot\|_\infty) \le C_d\,\xi^{-\frac{d}{2}}$. For any $\varphi : \mathbb{R}^d \to \mathbb{R}$, define its rescaled version[5] $(S\varphi)(z) := \varphi(Rz)/(1 + C_{d_x,d_y}R^4)$,

[5]With some abuse of notation, we apply this re-scaling transform to functions defined on spaces of possibly different dimensions without explicitly reflecting this in the notation.

where $C_{d_x,d_y} = 10\big(1+4\sqrt{d_x d_y}\big)$, and note that $S\varphi \in \widetilde{\mathcal{F}}_{d_x}$, for any $\varphi \in \mathcal{F}_R$. We also define the map $s : x \mapsto x/R$. Combining the above, for $d_x \geq 4$, we have

$$\begin{aligned}\mathbb{E}\left[\sup_{\varphi\in\mathcal{F}_R}\left|(\mu-\hat{\mu}_n)\varphi\right|\right] &\lesssim_{d_x,d_y} (1+R^4)\mathbb{E}\left[\sup_{\varphi\in\mathcal{F}_R}\left|\big(s_\sharp\mu - s_\sharp\hat{\mu}_n\big)(S\varphi)\right|\right]\\ &\lesssim_{d_x,d_y} (1+R^4)\left(\inf_{\alpha>0}\alpha + \frac{1}{\sqrt{n}}\int_\alpha^2 \xi^{-\frac{d_x}{4}}d\xi\right)\\ &\lesssim_{d_x,d_y} (1+R^4)n^{-\frac{2}{d_x}}(\log n)^{\mathbb{1}\{d_x=4\}}.\end{aligned}$$

When $d_x < 4$, the entropy integral is finite and we may pick $\alpha = 0$. Hence, in this case, $\mathcal{F}_R$ is a Donsker class and the resulting convergence rate is parametric $n^{-1/2}$. Altogether, we have

$$\mathbb{E}\left[\sup_{\varphi\in\cup_{\mathbf{A}}\mathcal{F}_{\mathbf{A}}}\left|(\mu-\hat{\mu}_n)\varphi\right|\right] \lesssim_{d_x,d_y} \big(1+R^4\big)n^{-\frac{2}{d_x\vee 4}}(\log n)^{\mathbb{1}\{d_x=4\}}. \tag{15}$$

We now move to treat the second term on the RHS of (14). First, observe that one may control it by the expected supremum of an empirical process indexed by $\mathcal{G}_R$, which is bounded by $\big(1+R^4\big)n^{-2/(d_y\vee 4)}(\log n)^{\mathbb{1}\{d_y=4\}}$ via similar steps as above. Together with (15), this would yield a two-sample empirical convergence rate bound of $n^{-2/(d_x\vee d_y\vee 4)}(\log n)^{\mathbb{1}\{d_x\vee d_y=4\}}$ for the squared $(2,2)$-GW distance. However, we aim to arrive at an upper bound that depends on the smaller dimension $d_x \wedge d_y$, as opposed to the larger one. As pointed out in Remark 5.6 of [19], this is possible by employing the LCA principle from [22, Lemma 2.1], which states that for any cost function c and function class $\mathcal{F}$, we have $N(\xi,\mathcal{F}^c,\|\cdot\|_\infty) \leq N(\xi,\mathcal{F},\|\cdot\|_\infty)$. Starting from a rescaling step as before, we obtain

$$\mathbb{E}\left[\sup_{\psi\in\cup_{\mathbf{A}}\mathcal{F}^c_{\mathbf{A}}}\left|(\nu-\hat{\nu}_n)\psi\right|\right] \lesssim_{d_x,d_y} (1+R^4)\mathbb{E}\left[\sup_{\psi\in\cup_{\mathbf{A}}\mathcal{F}^c_{\mathbf{A}}}\left|\big(s_\sharp\nu - s_\sharp\hat{\nu}_n\big)(S\psi)\right|\right]. \tag{16}$$

Using the LCA principle, we have the following bound on the covering number of the union of rescaled c-transformed classes.

Lemma A.4. *For any $\xi > 0$, we have the covering bound*

$$N\left(\xi,\cup_{\mathbf{A}\in\mathcal{D}_{R^2}}S(\mathcal{F}^c_{\mathbf{A}}),\|\cdot\|_\infty\right) \leq N\left(\frac{\xi}{64R^2},\mathcal{D}_{R^2},\|\cdot\|_{\mathrm{op}}\right)N\left(\frac{\xi}{2},\widetilde{\mathcal{F}}_{d_x},\|\cdot\|_\infty\right).$$

Armed with the lemma, we proceed from (16) and, for $d_x \geq 4$, obtain

$$\begin{aligned}\mathbb{E}\left[\sup_{\psi\in\cup_{\mathbf{A}}\mathcal{F}^c_{\mathbf{A}}}\left|(\nu-\hat{\nu}_n)\psi\right|\right] &\lesssim_{d_x,d_y} (1+R^4)\left(\inf_{\alpha>0}\alpha + \frac{1}{\sqrt{n}}\int_\alpha^2 \xi^{-\frac{d_x}{4}} + \log\frac{R^4}{\xi}d\xi\right)\\ &\lesssim_{d_x,d_y} (1+R^4)n^{-\frac{2}{d_x}}(\log n)^{\mathbb{1}\{d_x=4\}}.\end{aligned}$$

As before, when $d_x < 4$, a parametric rate bound holds instead. Inserting the above along with (14) into (15) concludes the proof of the two-sample upper bound for the squared distance.

Lastly, observe that if $\mathsf{D}(\mu,\nu) > 0$, then the two-sample rate for $\mathsf{D}(\mu,\nu)^2$ readily extends to $\mathsf{D}(\mu,\nu)$, since $\mathbb{E}\big[\big|\mathsf{D}(\mu,\nu) - \mathsf{D}(\hat{\mu}_n,\hat{\nu}_n)\big|\big] \leq \mathsf{D}(\mu,\nu)^{-1}\mathbb{E}\big[\big|\mathsf{D}(\mu,\nu)^2 - \mathsf{D}(\hat{\mu}_n,\hat{\nu}_n)^2\big|\big]$, and similarly for the one-sample case. We note, however, that unlike the bounds for D^2, this bound is not uniform over pairs of distributions with compact supports.

A.2.2 Lower bounds

We now move to establish the lower bounds. As the parametric lower bound of $n^{-1/2}$ trivially holds for our problem, we assume w.l.o.g. that $4 < d_x \leq d_y$ and $R = 4$.[6] Denoting $d := d_x$, we shall construct compactly supported distributions $\mu,\nu \in \mathbb{R}^d$ with the desired $n^{-2/d}$ empirical convergence rate lower bound. This is sufficient since lower-dimensional distributions can be

[6] To treat general R, one only needs to include a factor of $R^4/256$ in front of the one- and two-sample errors.

canonically embedded into higher dimensions without changing the value of D. As the lower bound holds for n sufficiently large, we occasionally absorb terms of order $O(1/n)$, $O(1/\sqrt{n})$ and $O(\sqrt{\log(n)/n})$ into the $n^{-2/d}$ convergence rate. Consider the uniform distributions $\mu = \mathrm{Unif}\big(B_d(0,1)\big)$ and $\nu = \mathrm{Unif}\big(B_d(0,2)\big)$.

We start from the one-sample case and establish $\mathbb{E}\big[\big|\mathsf{D}(\mu,\nu)^2 - \mathsf{D}(\hat{\mu}_n,\nu)^2\big|\big] \geq n^{-2/d}$. Theorem 9.21 of [42] implies that $T : x \mapsto 2x$ is an optimal Gromov-Monge map from μ and ν, and thus $\mathsf{D}(\mu,\nu)^2 = \int_{\mathcal{X}\times\mathcal{X}} \big|\|x-x'\|^2 - \|2x-2x'\|^2\big|^2 d\mu\otimes\mu(x,x')$. Let $\pi_n \in \Pi(\hat{\mu}_n,\nu)$ be an optimal coupling for $\mathsf{D}(\hat{\mu}_n,\nu)$ and notice that $\pi_n' = (\mathrm{id},\cdot/2)_\sharp\pi_n \in \Pi(\hat{\mu}_n,\mu)$ is optimal for $\mathsf{D}(\hat{\mu}_n,\mu)$. By completing the square, we then have

$$\begin{aligned}
\mathsf{D}(\hat{\mu}_n,\nu)^2 &= \int \big|\|y-y'\|^2 - \|z-z'\|^2\big|^2 d\pi_n\otimes\pi_n(y,z,y',z') \\
&= \int \big|\|y-y'\|^2 - \|2x-2x'\|^2\big|^2 d\pi_n'\otimes\pi_n'(y,x,y',x') \\
&= 4\mathsf{D}(\hat{\mu}_n,\mu)^2 - 3\int \|y-y'\|^4 d\hat{\mu}_n\otimes\hat{\mu}_n(y,y') + 12\int \|x-x'\|^4 d\mu\otimes\mu(x,x').
\end{aligned} \tag{17}$$

Combining this with the above expression for $\mathsf{D}(\mu,\nu)^2$, we obtain

$$\begin{aligned}
&\mathbb{E}\big[\big|\mathsf{D}(\mu,\nu)^2 - \mathsf{D}(\hat{\mu}_n,\nu)^2\big|\big] \\
&\geq 4\mathbb{E}\big[\mathsf{D}(\hat{\mu}_n,\mu)^2\big] + 3\mathbb{E}\left[\int \|x-x'\|^4 d\mu\otimes\mu(x,x') - \int \|y-y'\|^4 d\hat{\mu}_n\otimes\hat{\mu}_n(y,y')\right].
\end{aligned}$$

Evidently, the second term decays as n^{-1} since

$$\mathbb{E}\left[\int \|y-y'\|^4 d\hat{\mu}_n\otimes\hat{\mu}_n(y,y')\right] - \int \|x-x'\|^4 d\mu\otimes\mu(x,x') = \frac{1}{n}\int \|x-x'\|^4 d\mu\otimes\mu(x,x').$$

For the first term, let $\tilde{\mu}_n$ be the centered version of $\hat{\mu}_n$ and invoke Lemma 3.3 to obtain

$$\begin{aligned}
\mathbb{E}\big[\mathsf{D}^2(\hat{\mu}_n,\mu)\big] &= \mathbb{E}\big[\mathsf{D}^2(\tilde{\mu}_n,\mu)\big] \\
&\gtrsim \lambda_{\min}(\boldsymbol{\Sigma}_\mu) \inf_{\mathbf{U}\in O(d)} \mathsf{W}_2(\tilde{\mu}_n,\mathbf{U}_\sharp\mu)^2 \\
&= \lambda_{\min}(\boldsymbol{\Sigma}_\mu)\,\mathbb{E}\big[\mathsf{W}_2(\tilde{\mu}_n,\mu)^2\big] \\
&\geq \lambda_{\min}(\boldsymbol{\Sigma}_\mu)\left(\mathbb{E}\big[\mathsf{W}_1(\hat{\mu}_n,\mu) - \mathsf{W}_1(\hat{\mu}_n,\tilde{\mu}_n)\big]\right)^2,
\end{aligned}$$

where the equality uses the rotational invariance of μ, while the last step is by monotonicity of $p \mapsto \mathsf{W}_p$ and Jensen's inequality. Observe that $\mathbb{E}[\mathsf{W}_1(\hat{\mu}_n,\tilde{\mu}_n)] \leq \mathbb{E}[\|\bar{x}_n\|] \leq \sqrt{M_2(\mu)/n}$, where $\bar{x}_n := \int x\hat{\mu}_n(x)$ is the sample mean. Combining this with the fact that $\mathbb{E}[\mathsf{W}_1(\hat{\mu}_n,\mu)] \gtrsim n^{-1/d}$ [11], produces the desired lower bound on the one-sample GW convergence rate.

We proceed with the two-sample lower bound, which requires more work. Given the empirical measures $\hat{\mu}_n,\hat{\nu}_n$, define $\hat{\mu}_n' := (\cdot/2)_\sharp\hat{\nu}_n$ and note that it forms an empirical distribution of μ that is independent of $\hat{\mu}_n$. Write $X_1',\dots,X_n'$ for the samples comprising $\hat{\mu}_n'$. Let $\pi_n \in \Pi(\hat{\mu}_n,\hat{\nu}_n)$ be an optimal GW coupling for $\mathsf{D}(\hat{\mu}_n,\hat{\nu}_n)$ and set $\pi_n' := (\mathrm{id},\cdot/2)_\sharp\pi_n \in \Pi(\hat{\mu}_n,\hat{\mu}_n')$, which is optimal for $\mathsf{D}(\hat{\mu}_n,\hat{\mu}_n')$. Repeating the steps in (17), with $\hat{\nu}_n,\hat{\mu}_n'$ in place of ν,μ yields

$$\mathsf{D}(\hat{\mu}_n,\hat{\nu}_n)^2 = 4\mathsf{D}(\hat{\mu}_n,\hat{\mu}_n')^2 - 3\int \|y-y'\|^4 d\hat{\mu}_n\otimes\hat{\mu}_n(y,y') + 12\int \|y-y'\|^4 d\hat{\mu}_n'\otimes\hat{\mu}_n'(y,y').$$

Consequently, we represent the two-sample error as

$$\begin{aligned}
\mathsf{D}(\hat{\mu}_n,\hat{\nu}_n)^2 - \mathsf{D}(\mu,\nu)^2 &= 4\mathsf{D}(\hat{\mu}_n,\hat{\mu}_n')^2 - 3\int \|y-y'\|^4 d\hat{\mu}_n\otimes\hat{\mu}_n(y,y') \\
&\quad + 12\int \|y-y'\|^4 d\hat{\mu}_n'\otimes\hat{\mu}_n'(y,y') - 9\int \|y-y'\|^4 d\mu\otimes\mu(y,y').
\end{aligned} \tag{18}$$

As before, we have $\mathbb{E}\left[\int \|y-y'\|^4 d\hat{\mu}_n\otimes\hat{\mu}_n(y,y')\right] = \frac{n-1}{n}\int \|y-y'\|^4 d\mu\otimes\mu(y,y')$ and similarly for $\mathbb{E}\left[\int \|y-y'\|^4 d\hat{\mu}_n'\otimes\hat{\mu}_n'(y,y')\right]$, and the problem reduces to lower bounding $\mathbb{E}[\mathsf{D}(\hat{\mu}_n,\hat{\mu}_n')^2]$. We have the technical lemma below, which is proven in Appendix C.5

Lemma A.5 (Intermediate lower bound)**.** *The following bound holds*

$$\mathbb{E}[\mathsf{D}(\hat{\mu}_n, \hat{\mu}_n')^2] \gtrsim \mathbb{E}\left[\lambda_{\min}(\mathbf{\Sigma}_{\hat{\mu}_n})\mathbb{E}\left[\inf_{\mathbf{U}\in O(d)} \mathsf{W}_1(\hat{\mu}_n, \mathbf{U}_\sharp\hat{\mu}_n')^2 \middle| X_1, \dots, X_n\right]\right] - 2\sqrt{\frac{M_2(\mu)}{n}}. \quad (19)$$

To treat the inner (conditional) expectation on the RHS of (19), we make use of the next lemma; see Appendix C.6 for the proof.

Lemma A.6. *For any $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$ with $\mathrm{spt}(\mu), \mathrm{spt}(\nu) \subset B_d(0,1)$, we have*

$$\mathbb{E}\left[\inf_{\mathbf{U}\in O(d)} \mathsf{W}_1(\hat{\mu}_n, \mathbf{U}_\sharp\nu)\right] \geq \inf_{\mathbf{U}\in O(d)} \mathbb{E}\big[\mathsf{W}_1(\hat{\mu}_n, \mathbf{U}_\sharp\nu)\big] - C_d\sqrt{\frac{\log n}{n}},$$

where C_d depends only on the dimension d.

Applying the lemma, we obtain

$$\mathbb{E}\left[\inf_{\mathbf{U}\in O(d)} \mathsf{W}_1(\hat{\mu}_n, \mathbf{U}_\sharp\hat{\mu}_n') \middle| X_1, \cdots, X_n\right] \geq \inf_{\mathbf{U}\in O(d)} \mathbb{E}\big[\mathsf{W}_1(\hat{\mu}_n, \mathbf{U}_\sharp\hat{\mu}_n') \big| X_1, \cdots, X_n\big] - C_d\sqrt{\frac{\log n}{n}}$$

Note that for any $\mathbf{U} \in O(d)$, we have $\mathbb{E}\big[\mathsf{W}_1(\hat{\mu}_n, \mathbf{U}_\sharp\hat{\mu}_n') \big| X_1, \cdots, X_n\big] \geq \mathsf{W}_1(\mu, \mathbf{U}_\sharp\hat{\mu}_n') = \mathsf{W}_1(\mu, \hat{\mu}_n')$, where the first inequality follows because $\mathbb{E}[\mathsf{W}_1(\hat{\mu}_n, \nu)] \geq \mathsf{W}_1(\mu, \nu)$ for any μ, ν (due to convexity), while the second equality uses the fact that $\mathsf{W}_p(\mu, \nu) = \mathsf{W}_p(f_\sharp\mu, f_\sharp\nu)$ for any isometry f and the rotational invariance of μ. Inserting this back into (19), yields

$$\mathbb{E}[\mathsf{D}(\hat{\mu}_n, \hat{\mu}_n')^2] \gtrsim \mathbb{E}\left[\lambda_{\min}(\mathbf{\Sigma}_{\hat{\mu}_n}) \inf_{\mathbf{U}\in O(d)} \mathsf{W}_2(\hat{\mu}_n, \mathbf{U}_\sharp\hat{\mu}_n')^2\right] \geq \mathbb{E}\big[\lambda_{\min}(\mathbf{\Sigma}_{\hat{\mu}_n})\mathsf{W}_1(\hat{\mu}_n, \mu)^2\big].$$

To lower bound the expectation on the RHS, recall that by Proposition 2.1 in [11] (see also [47, Proposition 6]), for n sufficiently large, we have $\mathsf{W}_1(\alpha, \beta_n) \gtrsim_d n^{-1/d}$ for any distributions $\alpha, \beta_n \in \mathcal{P}(\mathbb{R}^d)$, such that α has a Lebesgue density and β_n is supported on n points. In particular, we conclude that there exists $n_0 \in \mathbb{N}$ and $c_d > 0$, such that for all $n > n_0$, we have $\mathsf{W}_1(\mu, \hat{\mu}_n') \geq c_d n^{-1/d}$ a.s. Inserting this into the bound above gives

$$\mathbb{E}[\mathsf{D}(\hat{\mu}_n, \hat{\mu}_n')^2] \gtrsim_d \mathbb{E}\big[\lambda_{\min}(\mathbf{\Sigma}_{\hat{\mu}_n})\big] \cdot n^{-2/d}, \quad (20)$$

and the problem reduces to lower bounding the expected smallest eigenvalue.

Write $\mathbb{E}\big[\lambda_{\min}(\mathbf{\Sigma}_{\hat{\mu}_n})\big] = \mathbb{E}\big[\inf_{\|v\|=1} \hat{\mu}_n|v\cdot x|^2\big]$. We again control this quantity via bounds on an empirical processes indexed by the Donsker class $\{|v\cdot x|^2 : \|v\| = 1\}$. Specifically, there is an $n_1 \in \mathbb{N}$ that depends only on d, such that for any $n > n_1$, we have $\mathbb{E}\big[\sup_{\|v\|=1} \big|(\hat{\mu}_n - \mu)|v\cdot x|^2\big|\big] \leq \lambda_{\min}(\mathbf{\Sigma}_\mu)/2$. Consequently

$$\begin{aligned}
\mathbb{E}\left[\inf_{\|v\|=1} \hat{\mu}_n|v\cdot x|^2\right] &= \mathbb{E}\left[\inf_{\|v\|=1} \hat{\mu}_n|v\cdot x|^2 - \inf_{\|v\|=1} \mathbb{E}\left[\hat{\mu}_n|v\cdot x|^2\right]\right] + \inf_{\|v\|=1} \mathbb{E}\left[\hat{\mu}_n|v\cdot x|^2\right] \\
&\geq \mathbb{E}\left[\inf_{\|v\|=1} \hat{\mu}_n|v\cdot x|^2 - \mathbb{E}\left[\hat{\mu}_n|v\cdot x|^2\right]\right] + \inf_{\|v\|=1} \mu|v\cdot x|^2 \\
&= \mathbb{E}\left[\inf_{\|v\|=1} (\hat{\mu}_n - \mu)|v\cdot x|^2\right] + \inf_{\|v\|=1} \mu|v\cdot x| \\
&\geq \inf_{\|v\|=1} \mu|v\cdot x|^2 - \mathbb{E}\left[\sup_{\|v\|=1} |(\hat{\mu}_n - \mu)|v\cdot x|^2|\right] \\
&\geq \frac{\lambda_{\min}(\mathbf{\Sigma}_\mu)}{2}.
\end{aligned}$$

Inserting this back into (20) and recalling the decomposition of the empirical estimation error from (18) concludes the proof of the two-sample lower bound. □

Remark 2 (Wasserstein Procrustes empirical convergence rate). Our two-sample analysis above essentially establishes an $n^{-1/d}$ lower bound on the Wasserstein Procrustes empirical convergence rate, whenever $d \geq 3$. Since the Procrustes is trivially upper bounded by standard W_2 and is a pseudometric, it inherits the $n^{-1/d}$ upper bound on the rate from it as well. Together, these show that the $n^{-1/d}$ empirical convergence rate is sharp in general. Our argument is readily adjusted to cover both the one- and two-sample settings and can be extended to any order $p \geq 1$.

B Proof of Lemma 3.3

Throughout this proof, we omit the dummy variables from the probability measure in our notation for integrals, writing $\int f(x,y,x',y')d\pi \otimes \pi$ instead of $\int f(x,y,x',y')d\pi \otimes \pi(x,y,x',y')$. For the first inequality, we now have

$$
\begin{aligned}
&\mathsf{D}_{p,q}(\mu,\nu)^p \\
&= \int \big| \|x-x'\|^q - \|y-y'\|^q \big|^p d\pi \otimes \pi \\
&\leq q^p \int \left(\|x-x'\|^{q-1} + \|y-y'\|^{q-1}\right)^p \left(\|x-y\| + \|x'-y'\|\right)^p d\pi \otimes \pi \\
&\leq q^p \left(\int \left(\|x-x'\|^{q-1} + \|y-y'\|^{q-1}\right)^{\frac{pq}{q-1}} d\pi \otimes \pi \right)^{\frac{q-1}{q}} \left(\int \left(\|x-y\| + \|x'-y'\|\right)^{qp} d\pi \otimes \pi \right)^{\frac{1}{q}} \\
&\leq q^p 2^{2p-1} \left(\int \|x-x'\|^{pq} + \|y-y'\|^{pq} d\pi \otimes \pi \right)^{\frac{q-1}{q}} \left(\int \|x-y\|^{qp} + \|x'-y'\|^{qp} d\pi \otimes \pi \right)^{\frac{1}{q}} \\
&\leq q^p 2^{pq+p-1+1/q} \big(M_{pq}(\mu) + M_{pq}(\nu) \big)^{\frac{q-1}{q}} \left(\int \|x-y\|^{qp} d\pi \right)^{\frac{1}{q}},
\end{aligned}
$$

where the second line follows by mean value theorem for the function $x \mapsto x^q$, while the third line uses by Hölder's inequality.

For the second inequality, suppose first that μ, ν are centered. We may now expand

$$
\begin{aligned}
&\mathsf{D}(\mu,\nu)^2 \\
&\quad = \inf_{\pi \in \Pi(\mu,\nu)} 2\big(M_2(\mu) - M_2(\nu)\big)^2 + 2\int \big(\|x\|^2 - \|y\|^2\big)^2 d\pi + 4\int \big(\langle x,x'\rangle - \langle y,y'\rangle\big)^2 d\pi \otimes \pi \\
&\quad = 2\big(M_2(\mu) - M_2(\nu)\big)^2 + \inf_{\pi \in \Pi(\mu,\nu)} 2\int \big(\|x\|^2 - \|y\|^2\big)^2 d\pi + 4\big(\|\mathbf{\Sigma}_\mu\|_{\mathrm{F}}^2 + \|\mathbf{\Sigma}_\nu\|_{\mathrm{F}}^2 - 2\|\mathbf{\Gamma}_\pi\|_{\mathrm{F}}^2\big),
\end{aligned}
$$

where $\mathbf{\Gamma}_\pi = \int xy^\intercal d\pi$ is the cross-covariance of $(X,Y) \sim \pi$.

As the bound trivializes when $\mathsf{D}(\mu,\nu) = 0$, suppose that $\mathsf{D}(\mu,\nu)^2 = \iota > 0$ and let π be the corresponding optimal coupling. This implies $4\big(\|\mathbf{\Sigma}_\mu\|_{\mathrm{F}}^2 + \|\mathbf{\Sigma}_\nu\|_{\mathrm{F}}^2 - 2\|\mathbf{\Gamma}_\pi\|_{\mathrm{F}}^2\big) \leq \iota$. Consider the singular value decomposition $\mathbf{\Gamma}_\pi = \mathbf{P}\mathbf{\Lambda}\mathbf{Q}^\intercal$, where $\mathbf{P}, \mathbf{Q} \in O(d)$, and $\mathbf{\Gamma}$ is diagonal. By invariance of the GW distance to rotations and since $\tilde{\pi} = (\mathbf{P}^\intercal, \mathbf{Q}^\intercal)_\sharp \pi$ is optimal for $\mathsf{GW}(\mathbf{P}_\sharp^\intercal \mu, \mathbf{Q}_\sharp^\intercal \nu)$, we similarly obtain $4\big(\|\mathbf{P}^\intercal \mathbf{\Sigma}_\mu \mathbf{P}\|_{\mathrm{F}}^2 + \|\mathbf{Q}^\intercal \mathbf{\Sigma}_\nu \mathbf{Q}\|_{\mathrm{F}}^2 - 2\|\mathbf{\Lambda}\|_{\mathrm{F}}^2\big) \leq \iota$. Denote the singular values of a matrix $\mathbf{A} \in \mathbb{R}^{d \times d}$ by $\sigma_1(\mathbf{A}), \ldots, \sigma_d(\mathbf{A})$. Also denote the diagonal entries of $\mathbf{P}\mathbf{\Sigma}_\mu\mathbf{P}^\intercal, \mathbf{Q}\mathbf{\Sigma}_\nu\mathbf{Q}^\intercal$ as $a_1, \cdots, a_d$ and $b_1, \cdots, b_d$, respectively. We thus obtain

$$
\sum_{i=1}^d \left(a_i^2 + b_i^2 - 2\sigma_i(\mathbf{\Gamma}_\pi)^2\right) \leq \iota.
$$

Observing that $a_i + b_i - 2\sigma_i(\mathbf{\Gamma}_\pi) \geq 0$, as $\int x_i^2 + y_i^2 - 2x_i y_i d\tilde{\pi} \geq 0$, we further have

$$
\sqrt{\frac{a_i^2 + b_i^2}{2}} \geq \frac{a_i + b_i}{2} \geq \sigma_i(\mathbf{\Gamma}_\pi), \quad \forall i = 1, \ldots, d,
$$

which implies

$$\begin{aligned}\iota &\geq 8\sum_{i=1}^{d}\left(\sqrt{\frac{a_i^2+b_i^2}{2}}+\sigma_i(\mathbf{\Gamma}_\pi)\right)\left(\sqrt{\frac{a_i^2+b_i^2}{2}}-\sigma_i(\mathbf{\Gamma}_\pi)\right)\\ &\geq 8\min_{i=1,\dots,d}\sqrt{\frac{a_i^2+b_i^2}{2}}\cdot\sum_{i=1}^{d}\left(\sqrt{\frac{a_i^2+b_i^2}{2}}-\sigma_i(\mathbf{\Gamma}_\pi)\right).\end{aligned}$$

Having that, we compute

$$\begin{aligned}\mathsf{W}_2\big(\mu,(\mathbf{P}\mathbf{Q}^\intercal)_\sharp\nu\big)^2 &= \mathsf{W}_2\big(\mathbf{P}_\sharp^\intercal\mu,\mathbf{Q}_\sharp^\intercal\nu\big)^2\\ &\leq \int\|x-y\|^2 d\tilde{\pi}\\ &= \sum_{i=1}^{d}\big(a_i+b_i-2\sigma_i(\mathbf{\Gamma}_\pi)\big)\\ &\leq \sum_{i=1}^{d}\left(\sqrt{\frac{a_i^2+b_i^2}{2}}-\sigma_i(\mathbf{\Gamma}_\pi)\right)\\ &\leq \frac{\iota}{8\min_i\sqrt{\frac{a_i^2+b_i^2}{2}}}.\end{aligned}$$

Notice that $\lambda_{\min}(\mathbf{\Sigma}_\mu)\leq a_i$ and $\lambda_{\min}(\mathbf{\Sigma}_\nu)\leq b_i$, for all $i=1,\dots,d$, and use the fact that $\mathbf{P}\mathbf{Q}^\intercal\in O(d)$ to conclude that

$$\Big(32\big(\lambda_{\min}(\mathbf{\Sigma}_\mu)^2+\lambda_{\min}(\mathbf{\Sigma}_\nu)^2\big)\Big)^{\frac{1}{4}}\inf_{\mathbf{U}\in O(d)}\mathsf{W}_2(\mu,\mathbf{U}_\sharp\nu)\leq\mathsf{D}(\mu,\nu),$$

whenever μ,ν are centered. To remove the centering assumption one only has to replace $O(d)$ above with the isometry group $E(d)$, which contains translations in addition to rotations. □

C Proofs of Lemmas for Theorem 3.2

C.1 Proof of Lemma A.1

Let $\bar{x}_n\coloneqq\int x\hat{\mu}_n$, $\bar{y}_n\coloneqq\int y\hat{\nu}_n$ denote the sample means and define $\tilde{\mu}_n,\tilde{\nu}_n$ as the centered versions of the empirical distributions, i.e., $\tilde{\mu}_n=(\cdot-\bar{x}_n)_\sharp\hat{\mu}_n$ and similarly for $\tilde{\nu}$. Note that $\mathsf{S}_\varepsilon(\hat{\mu}_n,\hat{\nu}_n)=\mathsf{S}_\varepsilon(\tilde{\mu}_n,\tilde{\nu}_n)=\mathsf{S}^1(\tilde{\mu}_n,\tilde{\nu}_n)+\mathsf{S}^2(\tilde{\mu}_n,\tilde{\nu}_n)$ and so

$$\begin{aligned}\mathbb{E}\big[\big|\mathsf{D}(\mu,\nu)-\mathsf{D}(\hat{\mu}_n,\hat{\nu}_n)\big|\big] &\leq \mathbb{E}\big[\big|\mathsf{S}^1(\mu,\nu)-\mathsf{S}^1(\hat{\mu}_n,\hat{\nu}_n)\big|\big]+\mathbb{E}\big[\big|\mathsf{S}^2(\mu,\nu)-\mathsf{S}^2(\hat{\mu}_n,\hat{\nu}_n)\big|\big]\\ &\quad+\mathbb{E}\big[\big|\mathsf{S}^1(\hat{\mu}_n,\hat{\nu}_n)-\mathsf{S}^1(\tilde{\mu}_n,\tilde{\nu}_n)\big|\big]+\mathbb{E}\big[\big|\mathsf{S}^2(\hat{\mu}_n,\hat{\nu}_n)-\mathsf{S}^2(\tilde{\mu}_n,\tilde{\nu}_n)\big|\big]\end{aligned}$$

We proceed by bounding the terms in the second line. For the first one, observe

$$\begin{aligned}\mathbb{E}\big[\big|\mathsf{S}^1(\hat{\mu}_n,\hat{\nu}_n)-\mathsf{S}^1(\tilde{\mu}_n,\tilde{\nu}_n)\big|\big] &\lesssim \mathbb{E}\left[\left|\int\Big(\|x-\bar{x}_n\|^2\|y-\bar{y}_n\|^2-\|x\|^2\|y\|^2\Big)d\hat{\mu}_n\otimes\hat{\nu}_n(x,y)\right|\right]\\ &= \mathbb{E}\left[\left|\|\bar{x}_n\|^2\|\bar{y}_n\|^2-\|\bar{x}_n\|^2\int\|y\|^2d\hat{\nu}_n(y)-\|\bar{y}_n\|^2\int\|x\|^2d\hat{\mu}_n(x)\right|\right]\\ &\lesssim \frac{R^4}{n}.\end{aligned}\tag{21}$$

In the last step above we have used the following bound on the 4th absolute moment of the sample mean. Write $\bar{x}_n=\frac{1}{n}\sum_{i=1}^n X_i$, where $X_1,\dots,X_n$ are the i.i.d. samples defining the empirical

measure $\hat{\mu}_n$. Consider:

$$\begin{aligned}
\mathbb{E}[\|\bar{x}_n\|^4] &= \frac{1}{n^4}\mathbb{E}\left[\left(\sum\nolimits_{i,j}\langle X_i, X_j\rangle\right)^2\right] \\
&= \frac{1}{n^4}\mathbb{E}\left[2\sum\nolimits_{i\neq j}\langle X_i, X_j\rangle^2 + \sum\nolimits_{i,j}\langle X_i, X_i\rangle\langle X_j, X_j\rangle\right] \\
&= \frac{1}{n^4}\mathbb{E}\left[2\sum\nolimits_{i\neq j}\langle X_i, X_j\rangle^2 + \sum\nolimits_{i\neq j}\langle X_i, X_i\rangle\langle X_j, X_j\rangle + \sum\nolimits_{i}\langle X_i, X_i\rangle^2\right] \\
&= \frac{1}{n^4}\left(2n(n-1)\|\Sigma_\mu\|_{\mathrm{F}}^2 + n(n-1)M_2(\mu)^2 + nM_4(\mu)\right) \\
&\leq \frac{3n^2 M_4(\mu)}{n^4} \\
&\leq \frac{3R^4}{n^2}
\end{aligned}$$

where the two last steps bound $\|\Sigma_\mu\|_{\mathrm{F}}^2 \leq M_2(\mu)^2 \leq M_4(\mu)$ and $M_4(\mu) \leq R^4$.

It remains to analyze the centering bias of S^2. Consider

$$\begin{aligned}
&\mathbb{E}\left[\left|\mathsf{S}^2(\hat{\mu}_n, \hat{\nu}_n) - \mathsf{S}^2(\tilde{\mu}_n, \tilde{\nu}_n)\right|\right] \\
&\quad\lesssim \mathbb{E}\left[\sup_{\pi\in\Pi(\hat{\mu}_n,\hat{\nu}_n)}\left|\int\left(\|x-\bar{x}_n\|^2\|y-\bar{y}_n\|^2 - \|x\|^2\|y\|^2\right)d\pi(x,y)\right|\right] \\
&\quad+ \mathbb{E}\left[\sup_{\pi\in\Pi(\hat{\mu}_n,\hat{\nu}_n)}\left|\sum_{\substack{1\leq i\leq d_x\\1\leq j\leq d_y}}\left(\int x_i y_j d\pi(x,y)\right)^2 - \left(\int (x_i-\bar{x}_{n,i})(y_j-\bar{y}_{n,j})d\pi(x,y)\right)^2\right|\right].
\end{aligned} \tag{22}$$

For the first term above, we have

$$\begin{aligned}
&\mathbb{E}\left[\sup_{\pi\in\Pi(\hat{\mu}_n,\hat{\nu}_n)}\left|\int\left(\|x-\bar{x}_n\|^2\|y-\bar{y}_n\|^2 - \|x\|^2\|y\|^2\right)d\pi(x,y)\right|\right] \\
&\quad= \mathbb{E}\Bigg[\sup_{\pi\in\Pi(\hat{\mu}_n,\hat{\nu}_n)}\Bigg|2\int\left(2\langle x,\bar{x}_n\rangle\langle y,\bar{y}_n\rangle - \langle x,\bar{x}_n\rangle\|y\|^2 - \langle y,\bar{y}_n\rangle\|x\|^2\right)d\pi(x,y) \\
&\qquad\qquad - 3\|\bar{x}_n\|^2\|\bar{y}_n\|^2 + \|\bar{x}_n\|^2\int\|y\|^2 d\hat{\nu}_n(y) + \|\bar{y}_n\|^2\int\|x\|^2 d\hat{\mu}_n(x)\Bigg|\Bigg] \\
&\quad\lesssim \frac{R^4}{\sqrt{n}},
\end{aligned}$$

using the same fourth moment expansion of $\bar{x}_n$ as above. For the second term, we have

$$\begin{aligned}
&\mathbb{E}\left[\sup_{\pi\in\Pi(\hat{\mu}_n,\hat{\nu}_n)}\left|\sum_{\substack{1\leq i\leq d_x\\1\leq j\leq d_y}}\Big(\int x_i y_j d\pi(x,y)\Big)^2 - \Big(\int (x_i-\bar{x}_{n,i})(y_j-\bar{y}_{n,j})d\pi(x,y)\Big)^2\right|\right] \\
&\quad= \mathbb{E}\left[\sup_{\pi\in\Pi(\hat{\mu}_n,\hat{\nu}_n)}\left|\sum_{\substack{1\leq i\leq d_x\\1\leq j\leq d_y}}\left(\int x_i y_j d\pi(x,y) - \int (x_i-\bar{x}_{n,i})(y_j-\bar{y}_{n,j})d\pi(x,y)\right)\right.\right. \\
&\qquad\qquad\left.\left.\times\left(\int x_i y_j d\pi(x,y) + \int (x_i-\bar{x}_{n,i})(y_j-\bar{y}_{n,j})d\pi(x,y)\right)\right|\right]
\end{aligned}$$

with

$$\begin{aligned}
&\mathbb{E}\left[\sup_{\pi\in\Pi(\hat{\mu}_n,\hat{\nu}_n)}\sum_{\substack{1\le i\le d_x\\1\le j\le d_y}}\left(\int x_i y_j d\pi(x,y)-\int(x_i-\bar{x}_{n,i})(y_j-\bar{y}_{n,j})d\pi(x,y)\right)^2\right]\\
&\qquad=\mathbb{E}\left[\sup_{\pi\in\Pi(\hat{\mu}_n,\hat{\nu}_n)}\sum_{\substack{1\le i\le d_x\\1\le j\le d_y}}\left(\int x_i\bar{y}_{n,j}+\bar{x}_{n,i}y_j-\bar{x}_{n,i}\bar{y}_{n,j}d\pi(x,y)\right)^2\right]\\
&\qquad\lesssim\sum_{\substack{1\le i\le d_x\\1\le j\le d_y}}\mathbb{E}\left[\int(x_i\bar{y}_{n,j})^2d\hat{\mu}_n(x)+\int(\bar{x}_{n,i}y_j)^2d\hat{\nu}_n(x)+(\bar{x}_{n,i}\bar{y}_{n,j})^2\right]\\
&\qquad=\mathbb{E}\left[\|\bar{y}_n\|^2\int\|x\|^2d\hat{\mu}_n(x)+\|\bar{x}_n\|^2\int\|y\|^2d\hat{\nu}_n(y)+\|\bar{x}_n\|^2\|\bar{y}_n\|^2\right]\\
&\qquad\lesssim\frac{R^4}{n}
\end{aligned}$$

and

$$\begin{aligned}
&\mathbb{E}\left[\sup_{\pi\in\Pi(\hat{\mu}_n,\hat{\nu}_n)}\sum_{\substack{1\le i\le d_x\\1\le j\le d_y}}\left(\int x_i y_j d\pi(x,y)+\int(x_i-\bar{x}_{n,i})(y_j-\bar{y}_{n,j})d\pi(x,y)\right)^2\right]\\
&=\mathbb{E}\left[\sup_{\pi\in\Pi(\hat{\mu}_n,\hat{\nu}_n)}\sum_{\substack{1\le i\le d_x\\1\le j\le d_y}}\left(\int 2x_i y_j-x_i\bar{y}_{n,j}-\bar{x}_{n,i}y_j+\bar{x}_{n,i}\bar{y}_{n,j}d\pi(x,y)\right)^2\right]\\
&\lesssim\sum_{\substack{1\le i\le d_x\\1\le j\le d_y}}\mathbb{E}\Big[\int(x_i\bar{y}_{n,j})^2d\hat{\mu}_n(x)+\int(\bar{x}_{n,i}y_j)^2d\hat{\nu}_n(y)+(\bar{x}_{n,i}\bar{y}_{n,j})^2+\int x_i^2d\hat{\mu}_n(x)\int y_j^2d\hat{\nu}_n(y)\Big]\\
&=\mathbb{E}\left[\|\bar{y}_n\|^2\int\|x\|^2d\hat{\mu}_n(x)+\|\bar{x}_n\|^2\int\|y\|^2d\hat{\nu}_n(y)+\|\bar{x}_n\|^2\|\bar{y}_n\|^2\right]+\mathbb{E}\big[M_2(\hat{\mu}_n)M_2(\hat{\nu}_n)\big]\\
&\lesssim R^4.
\end{aligned}$$

Combine the pieces, we obtain $\mathbb{E}\big[\big|\mathsf{S}_\varepsilon^2(\hat{\mu}_n,\hat{\nu}_n)-\mathsf{S}_\varepsilon^2(\tilde{\mu}_n,\tilde{\nu}_n)\big|\big]\lesssim R^4n^{-1/2}$, which together with (21) concludes the proof. □

C.2 Proof of Lemma A.2

First, rewrite

$$\begin{aligned}
&\mathsf{S}^1(\mu,\nu)\\
&=\int\|x-x'\|^4d\mu\otimes\mu(x,x')+\int\|y-y'\|^4d\nu\otimes\nu(y,y')-4\int\|x\|^2\|y\|^2d\mu\otimes\nu(x,y)\\
&=2\big(M_4(\mu)+M_4(\nu)\big)+2\big(M_2(\mu)^2+M_2(\nu)^2\big)+4\big(\|\Sigma_\mu\|_{\mathrm{F}}^2+\|\Sigma_\nu\|_{\mathrm{F}}^2\big)-4M_2(\mu)M_2(\nu).
\end{aligned}$$

With this expansion, the empirical estimation error of S^1 can be bounded as

$$\begin{aligned}
&\mathbb{E}\big[\big|\mathsf{S}^1(\mu,\nu)-\mathsf{S}^1(\hat{\mu}_n,\hat{\nu}_n)\big|\big]\\
&\lesssim \mathbb{E}\big[\big|M_4(\mu)-M_4(\hat{\mu}_n)\big|\big]+\mathbb{E}\big[\big|M_4(\nu)-M_4(\hat{\nu}_n)\big|\big]+\mathbb{E}\big[\big|M_2(\mu)M_2(\nu)-M_2(\hat{\mu}_n)M_2(\hat{\nu}_n)\big|\big]\\
&+\sqrt{\mathbb{E}\left[\|\Sigma_\mu-\Sigma_{\hat{\mu}_n}\|_{\mathrm{F}}^2\right]\mathbb{E}\left[\|\Sigma_\mu+\Sigma_{\hat{\mu}_n}\|_{\mathrm{F}}^2\right]}+\sqrt{\mathbb{E}\left[\|\Sigma_\nu-\Sigma_{\hat{\nu}_n}\|_{\mathrm{F}}^2\right]\mathbb{E}\left[\|\Sigma_\nu+\Sigma_{\hat{\nu}_n}\|_{\mathrm{F}}^2\right]}\\
&+\sqrt{\mathbb{E}\Big[\big(M_2(\mu)+M_2(\hat{\mu}_n)\big)^2\Big]\mathbb{E}\Big[\big(M_2(\mu)-M_2(\hat{\mu}_n)\big)^2\Big]}\\
&\qquad\qquad+\sqrt{\mathbb{E}\Big[\big(M_2(\nu)+M_2(\hat{\nu}_n)\big)^2\Big]\mathbb{E}\Big[\big(M_2(\nu)-M_2(\hat{\nu}_n)\big)^2\Big]}. \quad (23)
\end{aligned}$$

Note that all terms above are moment estimations. Simple computation yields

$$\mathbb{E}\big[\big|\mathsf{S}^1(\mu,\nu)-\mathsf{S}^1(\hat{\mu}_n,\hat{\nu}_n)\big|\big]\lesssim\frac{R^4}{\sqrt{n}}.$$

C.3 Proof of Lemma A.3

With some abuse of notation, let $\mathcal{X}=B_{d_x}(0,R)$ and $\mathcal{Y}=B_{d_y}(0,R)$ be the ambient spaces. Recall from Section 2.1 that, for any $\mathbf{A}\in\mathcal{D}_{R^2}$, we have $\Phi_{\mathbf{A}}\subset C_b(\mathcal{X})\times C_b(\mathcal{Y})$ and we may further restrict to pairs of potentials that can be written as $(\varphi^{c\bar{c}},\varphi^c)$, for some $\varphi\in C_b(\mathcal{X})$. Since $\varphi^{c\bar{c}c}=\varphi^c$, the potentials are c- and $\bar{c}$-transforms of each other, i.e., we may only consider pairs (φ,ψ) with

$$\varphi(x)=\inf_{y\in\mathcal{Y}}c_{\mathbf{A}}(x,y)-\psi(y)\qquad\text{and}\qquad\psi(y)=\inf_{x\in\mathcal{X}}c_{\mathbf{A}}(x,y)-\varphi(x).$$

Observing that $c_{\mathbf{A}}$ is concave in both arguments, we see that φ and ψ are both concave. Indeed, one readily verifies that the epigraphs of $-\varphi$ and $-\psi$ are convex sets, since for any $\alpha\in[0,1]$ and $x_1,x_2\in\mathcal{X}$, we have

$$\varphi\big(\alpha x_1+(1-\alpha)x_2\big)\geq\inf_{y\in\mathcal{Y}}\alpha c_{\mathbf{A}}(x_1,y)+(1-\alpha)c_{\mathbf{A}}(x_1,y)-\psi(y)\geq\alpha\varphi(x_1)+(1-\alpha)\varphi(x_2)$$

and similarly for the other dual potential.

To bound the sup-norm of the augmented potentials, observe that the functional value is invariant to translations (i.e., $(\varphi-a,\psi+a)$ for some constant a). Since

$$\mathsf{OT}_{c_{\mathbf{A}}}(\mu,\nu)\geq-\|c_{\mathbf{A}}\|_\infty\geq-4\big(1+4\sqrt{d_xd_y}\big)R^4,$$

we further restrict to a class of functions with $\int\varphi d\mu+\int\psi d\nu\geq-2\Big(1+2\big(1+4\sqrt{d_xd_y}\big)R^4\Big)$. For such functions there must exist a point (x_0,y_0), for which

$$\varphi(x_0)+\psi(y_0)\geq-2\Big(1+2\big(1+4\sqrt{d_xd_y}\big)R^4\Big),$$

and by shifting the potentials to coincide on (x_0,y_0), i.e., $\varphi(x_0)=\psi(y_0)$, we obtain

$$\varphi(x_0)\geq-1-2\big(1+4\sqrt{d_xd_y}\big)R^4\qquad\text{and}\qquad\psi(y_0)\geq-1-2\big(1+4\sqrt{d_xd_y}\big)R^4.$$

By the constraint, we then have

$$\begin{aligned}
\varphi(x)&\leq c_{\mathbf{A}}(x,y_0)-\psi(y_0)\leq1+6\big(1+4\sqrt{d_xd_y}\big)R^4\\
\psi(y)&\leq c_{\mathbf{A}}(x_0,y)-\varphi(x_0)\leq1+6\big(1+4\sqrt{d_xd_y}\big)R^4.
\end{aligned}$$

From the above, we also deduce

$$\begin{aligned}
-\varphi(x)&\leq\|c_{\mathbf{A}}\|_\infty+\|\psi\|_\infty\leq1+10\big(1+4\sqrt{d_xd_y}\big)R^4\\
-\psi(y)&\leq\|c_{\mathbf{A}}\|_\infty+\|\varphi\|_\infty\leq1+10\big(1+4\sqrt{d_xd_y}\big)R^4,
\end{aligned}$$

which concludes the boundedness.

For Lipschitzness of optimal potentials note that for any $x \in \mathbb{R}^{d_x}$ we can find a sequence $\{y_k\}_{k\in\mathbb{N}} \subset \mathbb{R}^{d_y}$, such that $\varphi(x) \le c_{\mathbf{A}}(x, y_k) - \psi(y_k) \le \varphi(x) + 1/k$. So for any $x' \neq x$,

$$\begin{aligned}\varphi(x') - \varphi(x) &\le c_{\mathbf{A}}(x', y_k) - \psi(y_k) - \big(c_{\mathbf{A}}(x, y_k) - \psi(y_k)\big) + \frac{1}{k} \\ &= \frac{1}{k} + 4\|y_k\|^2\big(\|x\|^2 - \|x'\|^2\big) + 32(x - x')^\intercal \mathbf{A} y \\ &\le \frac{1}{k} + 8\big(1 + 2\sqrt{d_x d_y}\big)R^3\|x - x'\|.\end{aligned}$$

Now take $k \to \infty$ and interchange x, x' to conclude that the φ is Lipschitz. Applying the same argument for ψ concludes the proof of the lemma. □

C.4 Proof of Lemma A.4

We aim to prove the covering bound

$$N\left(\xi, \cup_{\mathbf{A}\in\mathcal{D}_{R^2}} S(\mathcal{F}^c_{\mathbf{A}}), \|\cdot\|_\infty\right) \le N\left(\frac{\xi}{64R^2}, \mathcal{D}_{R^2}, \|\cdot\|_{\mathrm{op}}\right) N\left(\frac{\xi}{2}, \widetilde{\mathcal{F}}_{d_x}, \|\cdot\|_\infty\right).$$

First, note that by Lemma A.3, we have

$$N\left(\xi, \cup_{\mathbf{A}\in\mathcal{D}_{R^2}} S(\mathcal{F}^c_{\mathbf{A}}), \|\cdot\|_\infty\right) \le N\left(\xi, \cup_{\mathbf{A}\in\mathcal{D}_{R^2}} S(\mathcal{F}^c_{R}), \|\cdot\|_\infty\right). \tag{24}$$

Set $\xi_1 = \frac{\xi}{64R^2}$ and $\xi_2 = \frac{\xi}{2}$, and take a ξ_1-net $\{\mathbf{A}_i\}_{i=1}^{N_1}$ of $\mathcal{D}_{R^2}$ and a ξ_2-net $\{\varphi_i\}_{i=1}^{N_2}$ of $\widetilde{\mathcal{F}}_{d_x}$. For $i = 1, \dots, N_1$ and $j = 1, \dots, N_2$, define the functions $g_{i,j} : \mathbb{R}^{d_y} \to \mathbb{R}$ by

$$g_{i,j}(y) = S\Big[\inf_x \big(c_{\mathbf{A}_j}(x, y) - (S^{-1}\varphi_i)(x)\big)\Big],$$

where $(S\varphi)(z) \coloneqq \varphi(Rz)/(1 + C_{d_x,d_y}R^4)$ is the rescaling operator defined after Eq. (14). We will show that $\{g_{i,j}\}_{i,j=(1,1)}^{(N_1,N_2)}$ forms a ξ-net of $\cup_{\mathbf{A}\in\mathcal{D}_{R^2}} S(\mathcal{F}^c_R)$, which together with the covering bound from (24) yields the result. Indeed, for any $\varphi \in \mathcal{F}_R$, we have

$$\begin{aligned}\Big\|S\Big[\inf_x \big(c_{\mathbf{A}}(x, \cdot) - \varphi(x)\big)\Big] - g_{i,j}\Big\|_\infty &\le \sup_{x,y} \frac{|32x^\intercal(\mathbf{A} - \mathbf{A}_i)y|}{1 + C_{d_x,d_y}R^4} + \|\varphi - \varphi_j\|_\infty \\ &\le \frac{32R^2}{1 + C_{d_x,d_y}R^4}\xi_1 + \xi_2 \\ &\le \xi,\end{aligned}$$

which concludes the proof. □

C.5 Proof of Lemma A.5

Using Lemma 3.3 along with the centering step from the proof of the one-sample lower bound, we have

$$\begin{aligned}\mathbb{E}[\mathsf{D}(\hat{\mu}_n, \hat{\mu}'_n)^2] &= \mathbb{E}[\mathsf{D}(\tilde{\mu}_n, \tilde{\mu}'_n)^2] \\ &\gtrsim \mathbb{E}\left[\lambda_{\min}(\mathbf{\Sigma}_{\tilde{\mu}_n}) \inf_{\mathbf{U}\in O(d)} \mathsf{W}_2(\tilde{\mu}_n, \mathbf{U}_\sharp\tilde{\mu}'_n)^2\right] \\ &\gtrsim \mathbb{E}\left[\lambda_{\min}(\mathbf{\Sigma}_{\hat{\mu}_n}) \inf_{\mathbf{U}\in O(d)} \mathsf{W}_2(\hat{\mu}_n, \mathbf{U}_\sharp\hat{\mu}'_n)^2\right] - 2\mathbb{E}\left[\|\bar{x}_n\|\right] \\ &\ge \mathbb{E}\left[\lambda_{\min}(\mathbf{\Sigma}_{\hat{\mu}_n})\mathbb{E}\left[\inf_{\mathbf{U}\in O(d)} \mathsf{W}_1(\hat{\mu}_n, \mathbf{U}_\sharp\hat{\mu}'_n)^2 \middle| X_1, \dots, X_n\right]\right] - 2\sqrt{\frac{M_2(\mu)}{n}}.\end{aligned}$$

To justify the third step above, observe that

$$\big|\lambda_{\min}(\mathbf{\Sigma}_{\hat{\mu}_n}) - \lambda_{\min}(\mathbf{\Sigma}_{\tilde{\mu}_n})\big| \le \sup_{\|v\|=1} \hat{\mu}_n\big||v \cdot x|^2 - |v \cdot (x - \bar{x}'_n)|^2\big| \le 3\|\bar{x}'_n\|,$$

and

$$\begin{aligned}
&\Big|\inf_{\mathbf{U}\in O(d)} \mathsf{W}_2(\tilde{\mu}_n, \mathbf{U}_\sharp\tilde{\mu}'_n)^2 - \inf_{\mathbf{U}\in O(d)} \mathsf{W}_2(\hat{\mu}_n, \mathbf{U}_\sharp\hat{\mu}'_n)^2\Big| \\
&\quad\le \sup_{\mathbf{U}\in O(d)} \big|\mathsf{W}_2(\tilde{\mu}_n, \mathbf{U}_\sharp\tilde{\mu}'_n)^2 - \mathsf{W}_2(\hat{\mu}_n, \mathbf{U}_\sharp\hat{\mu}'_n)^2\big| \\
&\quad= \sup_{\mathbf{U}\in O(d)} \big(\mathsf{W}_2(\tilde{\mu}_n, \mathbf{U}_\sharp\tilde{\mu}'_n) + \mathsf{W}_2(\hat{\mu}_n, \mathbf{U}_\sharp\hat{\mu}'_n)\big)\big|\mathsf{W}_2(\tilde{\mu}_n, \mathbf{U}_\sharp\tilde{\mu}'_n) - \mathsf{W}_2(\hat{\mu}_n, \mathbf{U}_\sharp\hat{\mu}'_n)\big| \\
&\quad\le \sup_{\mathbf{U}\in O(d)} \big(\mathsf{W}_2(\tilde{\mu}_n, \mathbf{U}_\sharp\tilde{\mu}'_n) + \mathsf{W}_2(\hat{\mu}_n, \mathbf{U}_\sharp\hat{\mu}'_n)\big)\big(\mathsf{W}_2(\hat{\mu}_n, \tilde{\mu}_n) + \mathsf{W}_2(\mathbf{U}_\sharp\hat{\mu}'_n, \mathbf{U}_\sharp\tilde{\mu}'_n)\big) \\
&\quad\le 6\big(\mathsf{W}_2(\hat{\mu}_n, \tilde{\mu}_n) + \mathsf{W}_2(\hat{\mu}'_n, \tilde{\mu}'_n)\big) \\
&\quad\le 6(\|\bar{x}_n\| + \|\bar{x}'_n\|).
\end{aligned}$$

Together, these imply the desired bound as $\inf_{\mathbf{U}\in O(d)} \mathsf{W}_2(\hat{\mu}_n, \mathbf{U}_\sharp\hat{\mu}'_n)^2 \le 4$, $\lambda_{\min}(\mathbf{\Sigma}_{\tilde{\mu}_n}) \le 1$, and

$$\begin{aligned}
&\Big|\lambda_{\min}(\mathbf{\Sigma}_{\tilde{\mu}_n}) \inf_{\mathbf{U}\in O(d)} \mathsf{W}_2(\tilde{\mu}_n, \mathbf{U}_\sharp\tilde{\mu}'_n)^2 - \lambda_{\min}(\mathbf{\Sigma}_{\hat{\mu}_n}) \inf_{\mathbf{U}\in O(d)} \mathsf{W}_2(\hat{\mu}_n, \mathbf{U}_\sharp\hat{\mu}'_n)^2\Big| \\
&\qquad\le \big|\lambda_{\min}(\mathbf{\Sigma}_{\tilde{\mu}_n}) - \lambda_{\min}(\mathbf{\Sigma}_{\hat{\mu}_n})\big| \inf_{\mathbf{U}\in O(d)} \mathsf{W}_2(\hat{\mu}_n, \mathbf{U}_\sharp\hat{\mu}'_n)^2 \\
&\qquad\quad + \lambda_{\min}(\mathbf{\Sigma}_{\tilde{\mu}_n})\Big|\inf_{\mathbf{U}\in O(d)} \mathsf{W}_2(\tilde{\mu}_n, \mathbf{U}_\sharp\tilde{\mu}'_n)^2 - \inf_{\mathbf{U}\in O(d)} \mathsf{W}_2(\hat{\mu}_n, \mathbf{U}_\sharp\hat{\mu}'_n)^2\Big|,
\end{aligned}$$

which validates (19).

C.6 Proof of Lemma A.6

Consider the following decomposition

$$\begin{aligned}
&\mathbb{E}\left[\inf_{\mathbf{U}\in O(d)} \mathsf{W}_1(\hat{\mu}_n, \mathbf{U}_\sharp\nu)\right] \\
&\quad= \mathbb{E}\left[\inf_{\mathbf{U}\in O(d)} \mathsf{W}_1(\hat{\mu}_n, \mathbf{U}_\sharp\nu) - \inf_{\mathbf{U}\in O(d)} \mathbb{E}\big[\mathsf{W}_1(\hat{\mu}_n, \mathbf{U}_\sharp\nu)\big]\right] + \inf_{\mathbf{U}\in O(d)} \mathbb{E}\big[\mathsf{W}_1(\hat{\mu}_n, \mathbf{U}_\sharp\nu)\big] \\
&\quad\ge \mathbb{E}\left[\inf_{\mathbf{U}\in O(d)} \Big(\mathsf{W}_1(\hat{\mu}_n, \mathbf{U}_\sharp\nu) - \mathbb{E}\big[\mathsf{W}_1(\hat{\mu}_n, \mathbf{U}_\sharp\nu)\big]\Big)\right] + \inf_{\mathbf{U}\in O(d)} \mathbb{E}\big[\mathsf{W}_1(\hat{\mu}_n, \mathbf{U}_\sharp\nu)\big] \\
&\quad= -\mathbb{E}\left[\sup_{\mathbf{U}\in O(d)} \Big(\mathbb{E}\big[\mathsf{W}_1(\hat{\mu}_n, \mathbf{U}_\sharp\nu)\big] - \mathsf{W}_1(\hat{\mu}_n, \mathbf{U}_\sharp\nu)\Big)\right] + \inf_{\mathbf{U}\in O(d)} \mathbb{E}\big[\mathsf{W}_1(\hat{\mu}_n, \mathbf{U}_\sharp\nu)\big].
\end{aligned}$$

Denoting $R_{\mathbf{U}} := \mathbb{E}\big[\mathsf{W}_1(\hat{\mu}_n, \mathbf{U}_\sharp\nu)\big] - \mathsf{W}_1(\hat{\mu}_n, \mathbf{U}_\sharp\nu)$, we proceed to upper bound $\mathbb{E}[\sup_{\mathbf{U}} R_{\mathbf{U}}]$. Note that $|R_{\mathbf{U}} - R_{\mathbf{V}}| \le 2\mathsf{W}_1(\mathbf{U}_\sharp\nu, \mathbf{V}_\sharp\nu) \le 2\|\mathbf{U} - \mathbf{V}\|_{\mathrm{op}}$, and thus the process $\{R_{\mathbf{U}}\}_{\mathbf{U}\in O(d)}$ is Lipschitz in $\mathbf{U}$. We further claim that $\{R_{\mathbf{U}}\}_{\mathbf{U}\in O(d)}$ is a sub-Gaussian process. To see, for fixed $\mathbf{U} \in O(d)$, define the function $w_{\mathbf{U}} : (x_1, \cdots, x_n) \in B_d(0,1)^n \mapsto \mathsf{W}_1\big(n^{-1}\sum_{i=1}^n \delta_{x_i}, \mathbf{U}\nu\big)$ and note that it has bounded differences:

$$\begin{aligned}
&\sup_{x_i, x'_i} \big|w_{\mathbf{U}}(x_1, \cdots, x_{i-1}, x_i, x_{i+1}, \cdots, x_n) - w_{\mathbf{U}}(x_1, \cdots, x_{i-1}, x'_i, x_{i+1}, \cdots, x_n)\big| \\
&\qquad\le \frac{\|x_i - x'_i\|}{n} \le \frac{2}{n},
\end{aligned}$$

For $X_1, \ldots, X_n$ i.i.d. from μ (for which $\mathrm{spt}(\mu) \subset B_d(0,1)$ by assumption), McDiarmid's inequality now yields

$$\mathbb{P}\big(\big|w_{\mathbf{U}}(X_1, \ldots, X_n) - \mathbb{E}[w_{\mathbf{U}}(X_1, \ldots, X_n)]\big| \ge t\big) \le 2e^{-nt^2/2}.$$

Observing that $R_{\mathbf{U}} = \mathbb{E}[w_{\mathbf{U}}(X_1, \ldots, X_n)] - w_{\mathbf{U}}(X_1, \ldots, X_n)$, by equivalence between definitions of sub-Gaussianity, we further obtain $\mathbb{E}[e^{sR_{\mathbf{U}}}] \le e^{s^2\sigma^2/2}$, for all s, where $\sigma = \frac{3\sqrt{2}}{\sqrt{n}}$. Thus, $\{R_{\mathbf{U}}\}_{\mathbf{U}\in O(d)}$ is indeed sub-Gaussian.

Combining Lipschitzness and sub-Gaussianity, we deploy a standard ϵ-net argument. Let $\{\mathbf{U}_i\}_{i=1}^N$ is an ϵ-net of $O(d)$ w.r.t. the operator norm. We have

$$\begin{aligned}\mathbb{E}\big[\sup_{\mathbf{U}} R_{\mathbf{U}}\big] &\leq \inf_{\epsilon>0} 2\epsilon + \mathbb{E}\left[\max_{i=1,\dots,N} R_{\mathbf{U}_i}\right] \\ &\leq \inf_{\epsilon>0} 2\epsilon + \frac{3\sqrt{2}}{\sqrt{n}}\sqrt{\log\big(N(O(d),\epsilon,\|\cdot\|_{\mathrm{op}})\big)} \\ &\lesssim_d \sqrt{\log(n)}/\sqrt{n},\end{aligned}$$

where the last step uses the fact that $\log\big(N(O(d),\epsilon,\|\cdot\|_{\mathrm{op}})\big) \leq (c\sqrt{d}\epsilon^{-1})^{d^2}$, for a universal constant c; cf. Lemma 4 from [32]. This concludes the proof. □

D Proof of Proposition 1

As mentioned before, for any $\sigma^\star \in \mathcal{S}^\star$, we have

$$a^\star = \frac{1}{2}\int xyd\pi^\star(x,y) \in \mathcal{A}^\star,$$

where $\pi^\star$ is the coupling induced by $\sigma^\star$, and consequently $(a^\star,\pi^\star)$ jointly minimize (9). For the first direction, suppose that $\mathcal{A}^\star \subset \{0.5W_-, 0.5W_+\}$ but that there exists $\tilde{\sigma} \notin \{\mathrm{id}, \overline{\mathrm{id}}\}$ that optimizes (7). Denoting the corresponding coupling by $\tilde{\pi}$, the above implies that $\tilde{a} = \frac{1}{2}\int xyd\tilde{\pi}(x,y) \in \mathcal{A}^*$. However, by the rearrangement inequality $0.5W_- < \tilde{a} < 0.5W_+$, which is a contradiction. Since the minimum in (7) is achieved, we conclude that $\mathcal{S}^\star \subset \{\mathrm{id}, \overline{\mathrm{id}}\}$.

For the other direction, suppose that $\mathcal{S}^\star \subset \{\mathrm{id}, \overline{\mathrm{id}}\}$ but that there exists $\tilde{a} \in (0.5W_-, 0.5W_+)$ with $\tilde{a} \in \mathcal{A}^\star$. We first argue the $f+g$ is differentiable at $\tilde{a}$. This follows because g is piecewise linear and concave, and so for any non-differentiability point a_0, the left and right derivatives satisfy $g'_-(a_0) > g'_+(a_0)$. Since f is smooth, we further obtain $(f+g)'_-(a_0) > (f+g)'_+(a_0)$, so a_0 cannot be a local minimum. We conclude that $f+g$ is differentiable at $\tilde{a}$ with $(f+g)'(\tilde{a}) = 0$.

Having that, let $\Pi_{\tilde{a}} \subset \Pi(\mu,\nu)$ be the argmin set for $g(\tilde{a})$ and fix $\tilde{\pi} \in \Pi_{\tilde{a}}$. Since $(f+g)'(\tilde{a}) = 0$, computing the derivative, we obtain

$$64\tilde{a} - 32\int xyd\tilde{\pi}(x,y) = 0.$$

Thus, $\int xyd\tilde{\pi}(x,y) = 2\tilde{a}$, for every $\pi \in \Pi_{\tilde{a}}$. Now, Since $(\tilde{a},\tilde{\pi})$ minimize (9), consider

$$\begin{aligned}\mathsf{S}^2(\mu,\nu) &= 32\tilde{a}^2 + \inf_{\pi\in\Pi_{\tilde{a}}}\int \big(-4x^2y^2 - 32\tilde{a}xy\big)d\pi(x,y) \\ &= \inf_{\pi\in\Pi_{\tilde{a}}} 32\tilde{a}^2 + \int \big(-4x^2y^2 - 32\tilde{a}xy\big)d\pi(x,y) \\ &= \inf_{\pi\in\Pi_{\tilde{a}}} 8\left(\int xyd\pi(x,y)\right)^2 + \int -4x^2y^2d\pi(x,y) - 16\left(\int xyd\pi(x,y)\right)^2 \\ &= \inf_{\pi\in\Pi_{\tilde{a}}} \int -4x^2y^2d\pi(x,y) - 8\left(\int xyd\pi(x,y)\right)^2.\end{aligned}$$

Recalling that

$$\mathsf{S}^2(\mu,\nu) = \inf_{\pi\in\Pi(\mu,\nu)}\int -4x^2y^2d\pi(x,y) - 8\left(\int xyd\pi(x,y)\right)^2$$

by definition (see (5)), we conclude that all elements of $\Pi_{\tilde{a}}$ are minimizers for S^2, and hence also minimizers of $\mathsf{D}(\mu,\nu)$.

To get a contradiction, recall that $\int xyd\tilde{\pi}(x,y) = 2\tilde{a}$, for all $\tilde{\pi} \in \Pi_{\tilde{a}}$. Since $\tilde{a} \in (0.5W_-, 0.5W_+)$, one readily verifies $W_- < \int xyd\tilde{\pi}(x,y) < W_+$. However, the couplings induced by id and $\overline{\mathrm{id}}$ achieve exactly W_+ and W_- for the said integral, and thus they are not contained in $\Pi_{\tilde{a}}$. Since by assumption the argmin is $\mathcal{S}^\star = \{\mathrm{id}, \overline{\mathrm{id}}\}$, we again have a contradiction and $\tilde{a}$ cannot be optimal for (9). The infimum must therefore be achieved on the boundary, i.e., $\mathcal{A}^\star \subset \{0.5W_-, 0.5W_+\}$, which concludes the proof. □

E Generalized Duality

We derive here the generalized dual representation for the GW distance of order $(p,q) = (2,2k)$, where $k \in \mathbb{N}$. The approach naturally extends to any even p value, but the cost of tedious technical details, which we prefer to avoid for presentation. Like in Appendix B, we omit dummy variables from our integral notation, writing $\int f(x,y,x',y')d\pi \otimes \pi$ instead of $\int f(x,y,x',y')d\pi \otimes \pi(x,y,x',y')$. Let $(\mu,\nu) \in \mathcal{P}_{4k}(\mathbb{R}^{d_x}) \times \mathcal{P}_{4k}(\mathbb{R}^{d_y})$, and expand the distortion cost to obtain

$$\mathsf{D}_{2,2k}(\mu,\nu)^2 = \inf_{\pi\in\Pi(\mu,\nu)} \int \left(\left| (\|x\|^2 - 2x\cdot x' + \|x'\|^2)^k - (\|y\|^2 - 2y\cdot y' + \|y'\|^2)^k \right|^2 \right) d\pi \otimes \pi.$$

Collecting terms that depend only on the marginals into $S^1(\mu,\nu)$ as before and omitting them for now, we seek a dual for the optimization problem

$$\inf_{\pi\in\Pi(\mu,\nu)} \int \left(-2(\|x\|^2 - 2x\cdot x' + \|x'\|^2)^k (\|y\|^2 - 2y\cdot y' + \|y'\|^2)^k\right) d\pi \otimes \pi.$$

The integrand is a homogeneous polynomial that is symmetric in (x,y) and (x',y'). Consequently, there exist polynomials $f_1,\dots,f_m$ of degree at most $4k$, and a symmetric matrix $\mathbf{C} \in \mathbb{R}^{m\times m}$ (whose entries are denoted by C_{ij}, for $i,j = 1,\dots,m$), such that

$$\begin{aligned}
&\inf_{\pi\in\Pi(\mu,\nu)} \int \left(-2(\|x\|^2 - 2x\cdot x' + \|x'\|^2)^k (\|y\|^2 - 2y\cdot y' + \|y'\|^2)^k\right) d\pi \otimes \pi \\
&= \inf_{\pi\in\Pi(\mu,\nu)} \int \Big(-2\|x\|^{2k}\|y\|^{2k} - 2\|x'\|^{2k}\|y'\|^{2k} + \sum_{1\le i,j\le m} C_{ij} f_i(x,y) f_j(x',y') \Big) d\pi \otimes \pi \\
&= \inf_{\pi\in\Pi(\mu,\nu)} \int -4\|x\|^{2k}\|y\|^{2k} d\pi + \sum_{1\le i,j\le m} C_{ij} \int f_i(x',y')d\pi \int f_j(x,y)d\pi.
\end{aligned}$$

Note that m is bounded by the number of monomials of degree at most $4k$ that can constructed from entries of x,y, i.e., $m = O((d_x+d_y)^{4k})$.[7] By diagonalizing C, we rewrite

$$\sum_{1\le i,j\le m} C_{ij} \int f_i(x,y)d\pi \int f_j(x',y')d\pi = \sum_{i=1}^{\ell} \left(\int g_i(x,y)d\pi \right)^2 - \sum_{i=\ell+1}^{m} \left(\int g_i(x,y)d\pi \right)^2, \tag{25}$$

where g_i are linear combinations of f_i, and ℓ is the number of positive eigenvalues of $\mathbf{C}$. Notice that the sum of squares on the RHS above can have positive or negative coefficient, which differs form the $(2,2)$ case where only a negative coefficient is present.

Armed with (25), we proceed with the same linearization step from the proof of Theorem 3.1 by introducing the new auxiliary optimization variables $a \in \mathbb{R}^{\ell}$ and $b \in \mathbb{R}^{m-\ell}$, as follows

$$\begin{aligned}
&\inf_{\pi\in\Pi(\mu,\nu)} \int \left(-2(\|x\|^2 - 2x\cdot x' + \|x'\|^2)^k (\|y\|^2 - 2y\cdot y' + \|y'\|^2)^k\right) d\pi \otimes \pi \\
&= \inf_{\pi\in\Pi(\mu,\nu)} \int -4\|x\|^{2k}\|y\|^{2k} d\pi + \sum_{i=1}^{\ell} \left(\int g_i(x,y)d\pi \right)^2 - \sum_{i=\ell+1}^{m} \left(\int g_i(x,y)d\pi \right)^2 \\
&= \inf_{\pi\in\Pi(\mu,\nu)} \int -4\|x\|^{2k}\|y\|^{2k} d\pi + \sup_{a\in\mathbb{R}^{\ell}} \sum_{i=1}^{\ell} \left(4a_i \int g_i(x,y)d\pi - 4a_i^2 \right) \\
&\qquad\qquad + \inf_{b\in\mathbb{R}^{m-\ell}} \sum_{i=\ell+1}^{m} \left(4b_{i-\ell}^2 - 4b_{i-\ell} \int g_i(x,y)d\pi \right) \\
&= 4 \sup_{a\in\mathbb{R}^{\ell}} \inf_{b\in\mathbb{R}^{m-\ell}} -\|a\|^2 + \|b\|^2 \\
&\qquad\qquad + \inf_{\pi\in\Pi(\mu,\nu)} \int \left(-\|x\|^{2k}\|y\|^{2k} + \sum_{i=1}^{\ell} a_i g_i(x,y) - \sum_{i=\ell+1}^{m} b_{i-\ell} g_i(x,y) \right) d\pi,
\end{aligned}$$

[7] In practice, m is often be much smaller, as seen from the example below.

where the last step follows from Sion's minimax theorem. The RHS above is the desired dual representation from (10). Further observe that as $\int g_i d\pi$ are uniformly bounded for all i and π, we may restrict optimization domains for a and b to compact sets. We identify the inner optimization over π as an OT problem with cost $c_{a,b} : (x, y) \mapsto -\|x\|^{2k}\|y\|^{2k} + \sum_{i=1}^{\ell} a_i g_i(x, y) - \sum_{i=\ell+1}^{m} b_{i-\ell} g_i(x, y)$, which is smooth (indeed, a polynomial) but not necessarily convex in x or y. Considering compactly supported distributions, one may invoke OT duality and establish Lipschitzness of the optimal potential, although convexity seems challenging to obtain in general. As explain in Section 4, by following the steps in the proof of Theorem 3.2, this leads to a two-sample empirical convergence rate of $O(n^{-1/(d_x \wedge d_y)})$. We leave further refinements of this rate as well as proofs of lower bounds for future work.

To illustrate the above procedure, we consider the special case of $p = q = 2$. This will also show how the duality formula from (10) reduces back to that from Theorem 3.1, after assuming that the populations are centered. As above, we start by expanding the $(2, 2)$-cost and omitting terms that depend only on the marginals (cf. (11)), to arrive at

$$\begin{aligned}
&\inf_{\pi \in \Pi(\mu,\nu)} \int -4\|x\|^2\|y\|^2 d\pi \\
&\qquad + 4 \int \Big(\langle x, x' \rangle (\|y\|^2 + \|y'\|^2) + (\|x\|^2 + \|x'\|^2)\langle y, y' \rangle - 2\langle x, x' \rangle \langle y, y' \rangle \Big) d\pi \otimes \pi \\
&= \inf_{\pi \in \Pi(\mu,\nu)} \int \left(-4 \sum_{1 \le i \le d_x, 1 \le j \le d_y} x_i^2 y_j^2 \right) d\pi \\
&\qquad + 4 \int \left(\sum_{1 \le i \le d_x, 1 \le j \le d_y} x_i x_i' (y_j^2 + y_j'^2) + (x_i^2 + x_i'^2) y_j y_j' - 2 x_i x_i' y_j y_j' \right) d\pi \otimes \pi \quad (26)
\end{aligned}$$

To diagonalize the second term, consider the set of linearly independent monomials $\{x_i, y_j, x_i y_j^2, x_i^2 y_j, x_i y_j\}_{1 \le i \le d_x, 1 \le j \le d_y}$, of which there are $d_x + d_y + 3d_x d_y$ in total (these are denoted by f_i in the general derivation above). For concreteness and simplicity, we henceforth assume $d_x = d_y = 1$. Define the vector

$$v(\pi) = \left(\int x \, \mathrm{d}\pi, \int y \, \mathrm{d}\pi, \int xy^2 \, \mathrm{d}\pi, \int x^2 y \, \mathrm{d}\pi, \int xy \, \mathrm{d}\pi \right)^{\intercal},$$

and construct coefficient matrix

$$\mathbf{C} = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -2 \end{bmatrix}.$$

For instance, we set $C_{1,3} = 1$ since the term $\int x d\pi \int xy^2 d\pi$, which is the product of $v_1(\pi)$ and $v_3(\pi)$, appears inside the functional from (26). We may now express

$$\begin{aligned}
&\inf_{\pi \in \Pi(\mu,\nu)} -4 \int x^2 y^2 d\pi + 4 \int \left(xx'y^2 + x'xy'^2 + x^2 yy' + x'^2 y'y - 2xx'yy' \right) d\pi \otimes \pi \\
&\qquad = \inf_{\pi \in \Pi(\mu,\nu)} -4 \int x^2 y^2 d\pi + 4 v(\pi)^{\intercal} \mathbf{C} v(\pi).
\end{aligned}$$

Diagonalizing $\mathbf{C}$, further yields

$$\begin{aligned}
v(\pi)^{\intercal} \mathbf{C} v(\pi) = &\left(\int \frac{\sqrt{2}}{2} x + \frac{\sqrt{2}}{2} xy^2 d\pi \right)^2 + \left(\int \frac{\sqrt{2}}{2} y + \frac{\sqrt{2}}{2} x^2 y d\pi \right)^2 \\
&- \left(\int -\frac{\sqrt{2}}{2} x + \frac{\sqrt{2}}{2} xy^2 d\pi \right)^2 - \left(\int -\frac{\sqrt{2}}{2} y + \frac{\sqrt{2}}{2} x^2 y d\pi \right)^2 - \left(\int \sqrt{2} xy d\pi \right)^2.
\end{aligned}$$

We proceed by introducing $a \in \mathbb{R}^2$ and $b \in \mathbb{R}^3$, as follows

$$
\begin{aligned}
&\inf_{\pi \in \Pi(\mu,\nu)} -4 \int x^2 y^2 d\pi + 4 \int \left(x x' y^2 + x' x y'^2 + x^2 y y' + x'^2 y' y - 2 x x' y y' \right) d\pi \otimes \pi \\
&= \inf_{\pi \in \Pi(\mu,\nu)} -4 \int x^2 y^2 d\pi + \left(\int \sqrt{2} x + \sqrt{2} x y^2 d\pi \right)^2 + \left(\int \sqrt{2} y + \sqrt{2} x^2 y d\pi \right)^2 \\
&\qquad - \left(\int -\sqrt{2} x + \sqrt{2} x y^2 d\pi \right)^2 - \left(\int -\sqrt{2} y + \sqrt{2} x^2 y d\pi \right)^2 - \left(\int 2\sqrt{2} x y d\pi \right)^2 \\
&= \inf_{\pi \in \Pi(\mu,\nu)} -4 \int x^2 y^2 d\pi \\
&\quad + \sup_{a \in \mathbb{R}^2} 4a_1 \int \left(\sqrt{2} x + \sqrt{2} x y^2 \right) d\pi - 4a_1^2 + 4a_2 \int \left(\sqrt{2} y + \sqrt{2} x^2 y \right) d\pi - 4a_2^2 \\
&\quad + \inf_{b \in \mathbb{R}^3} 4b_1^2 - 4b_1 \int \left(-\sqrt{2} x + \sqrt{2} x y^2 \right) d\pi + 4b_2^2 - 4b_2 \int \left(-\sqrt{2} y + \sqrt{2} x^2 y \right) d\pi \\
&\qquad\qquad + 4b_3^2 - 4b_3 \int 2\sqrt{2} x y d\pi \\
&= 4 \sup_{a \in \mathbb{R}^2} \inf_{b \in \mathbb{R}^3} -\|a\|^2 + \|b\|^2 + \inf_{\pi \in \Pi(\mu,\nu)} \int c_{a,b}(x,y) d\pi,
\end{aligned}
$$

where the cost function is

$$
\begin{aligned}
c_{a,b}(x,y) = &-x^2 y^2 + \sqrt{2} a_1 x + \sqrt{2} a_1 x y^2 + \sqrt{2} a_2 y + \sqrt{2} a_2 x^2 y + \sqrt{2} b_1 x - \sqrt{2} b_1 x y^2 \\
&+ \sqrt{2} b_2 y - \sqrt{2} b_2 x^2 y - 2\sqrt{2} b_3 x y.
\end{aligned}
$$

Lastly, notice that if μ, ν are centered, then

$$
v(\pi)^\intercal \mathbf{C} v(\pi) = -2 \left(\int x y d\pi \right)^2,
$$

which immediately recovers the dual form from Theorem 3.1, where the OT cost function is $c_{\mathbf{A}}(x,y) = -4x^2y^2 - 32\mathbf{A}_{1,1}xy$. The cost $c_{a,b}$ that arises from the general derivation is evidently more complex and comprises additional mixed terms (of order 3). This makes it harder to analyze, e.g., it is unclear whether $c_{a,b}$ is marginally convex/concave in each argument. Consequently, this approach may not lead to the same regularity profile for dual potentials as we have in Lemma A.3, which, in turn, may result in suboptimal empirical convergence rates.