The Annals of Statistics
2024, Vol. 52, No. 4, 1616–1645
https://doi.org/10.1214/24-AOS2406

GROMOV–WASSERSTEIN DISTANCES: ENTROPIC REGULARIZATION, DUALITY AND SAMPLE COMPLEXITY

BY ZHENGXIN ZHANG[1,a], ZIV GOLDFELD[2,b], YOUSSEF MROUEH[3,c] AND BHARATH K. SRIPERUMBUDUR[4,d]

[1]*Center for Applied Mathematics, Cornell University,* [a]*zz658@cornell.edu*
[2]*School of Electrical and Computer Engineering, Cornell University,* [b]*goldfeld@cornell.edu*
[3]*IBM Research, IBM,* [c]*mroueh@us.ibm.com*
[4]*Department of Statistics, Pennsylvania State University,* [d]*bks18@psu.edu*

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 treat both the standard and the entropically regularized GW distance, and derive dual forms that represent them in terms of the well-understood OT and entropic OT (EOT) problems, respectively. 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 rates are $n^{-2/\max\{\min\{d_x,d_y\},4\}}$ (up to a log factor when $\min\{d_x, d_y\} = 4$) for standard GW and $n^{-1/2}$ for entropic GW (EGW), which matches the corresponding rates for standard and entropic OT. The parametric rate for EGW is evidently optimal, while for standard GW we provide matching lower bounds, which establish sharpness of the derived rates. We also study stability of EGW in the entropic regularization parameter and prove approximation and continuity results for the cost and optimal couplings. 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 [45], 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 [61]

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

where $\Pi(\mu, \nu)$ is the set of all couplings between μ and ν. The GW distance 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 [53, 65], is an L^p relaxation of the Gromov–Hausdorff distance between metric spaces and enjoys various favorable properties. Among

Received January 2023; revised May 2024.
MSC2020 subject classifications. Primary 62G99; secondary 62H99.
Key words and phrases. Empirical convergence, entropic regularization, Gromov–Wasserstein distance, sample complexity, strong duality.

others, the GW distance (i) identifies pairs of mm spaces between which there exists a measure preserving isometry; (ii) defines a metric on the space of all mm spaces modulo the aforementioned isomorphic relation; and (iii) captures the empirical convergence of mm space, that is, when μ, ν are replaced with 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 [5, 16], alignment of language models [1], shape and graph matching [39, 44, 67, 68], heterogeneous domain adaptation [69] and generative modeling [7].

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 [10, 18, 37, 42] and limit distributions of both classical [14, 27, 36, 41, 59, 62] and regularized OT distances [4, 13, 25–28, 38, 46]; cf. Remarks 2.1 and 2.2 ahead for a detailed discussion about the utility of duality for the statistical analysis of standard and regularized OT, respectively. 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 $n \to \infty$ [45],[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 (standard and entropic) (2, 2)-GW distance over Euclidean spaces, and leveraging it to establish the first empirical convergence rates for the GW problem.

1.1. *Contribution.* For probability distributions μ and ν supported in $\mathbb{R}^{d_x}$ and $\mathbb{R}^{d_y}$, respectively, we study both the standard (2, 2)-GW distances from (1) and its entropically regularized version [58]

$$\mathsf{S}_\varepsilon(\mu, \nu) := \inf_{\pi \in \Pi(\mu,\nu)} \iint \big| \|x - x'\|^2 - \|y - y'\|^2 \big|^2 \, d\pi \otimes \pi(x, y, x', y') + \varepsilon \mathsf{D}_{\mathsf{KL}}(\pi \| \mu \otimes \nu),$$

where $\mathsf{D}_{\mathsf{KL}}(\cdot\|\cdot)$ is the Kullback–Leibler (KL) divergence. The interest in entropic GW (EGW) stems from its computational tractability [49, 51, 54, 58], which makes it a popular approach in practice. Our first main contribution is a duality theory for GW and EGW, which linearizes these quadratic functionals and ties them, respectively, to the well understood problems of OT and EOT. 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 or entropic OT (EOT) 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 or EOT duality to arrive at a dual formulation for $\mathsf{D}_{2,2}(\mu, \nu)^2$ and $\mathsf{S}_\varepsilon(\mu, \nu)$. The dual form involves optimization over $\mathbf{A}$, which we show can be restricted to a hypercube whose side length depends only on the 2nd moments of μ, ν.

The GW and EGW dual forms enable an analysis of expected empirical convergence rates by drawing upon proof techniques from statistical OT. Namely, we consider the rates at which $\mathbb{E}[|\mathsf{D}_{2,2}(\mu, \nu)^2 - \mathsf{D}_{2,2}(\hat{\mu}_n, \hat{\nu}_n)^2|]$ and $\mathbb{E}[|\mathsf{S}_\varepsilon(\mu, \nu) - \mathsf{S}_\varepsilon(\hat{\mu}_n, \hat{\nu}_n)|]$ decay to 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 or EOT dual

[1] [45] 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 pqth moments.

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 and EOT analyses. For EGW, we show that the potentials are Hölder smooth to an arbitrary order and provide bounds on the growth rate of their Hölder norm. Combining the regularity theory with a chaining argument and entropy integral bounds for Hölder classes, we arrive at $n^{-1/2}$ as the empirical convergence rate for EGW. This parametric rate holds in any ambient dimensions d_x, d_y and is inline with EOT empirical rates [21, 46]. For the unregularized GW problem, 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 similar analysis to the entropic case while leveraging a variant of the lower complexity adaptation (LCA) principle from [37] leads to an $n^{-2/\max\{\min\{d_x,d_y\},4\}}$ 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 establish 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 [32], which may be of independent interest.

We also address basic structural properties of the GW and EGW distances. First, we study stability of the entropic variant in the regularization parameter ε and establish an $O(\varepsilon \log(1/\varepsilon))$ bound on the gap between (squared) GW and EGW. This bound matches the entropic approximation error in the standard OT case [21]. However, unlike the result from [21], that accounts only for compactly supported distributions, our derivation relies on maximum entropy inequalities and holds for arbitrary distributions. After treating the entropic approximation of the GW cost, we prove that optimal entropic couplings weakly converge towards an optimal GW coupling as $\varepsilon \to 0$ by leveraging the notion of Γ-convergence. 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. [63]), 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 verify and visualize the different regimes of optimal solutions via simple numerical simulations.

1.2. *Literature review.* The GW distance was first proposed in [45] 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 [61]. The existence of Gromov–Monge maps was studied in [19], 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 [15], 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 [40].

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 [11]. For this reason, significant attention was devoted to variants of the GW problem that circumvent this computational hardness. The sliced GW distance [63] 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 [56], which lends well for convex/conic relaxations. A variant that directly optimizes over bi-directional Monge maps between the mm space was considered in [71]. While these methods offer certain advantages, it is the approach based on entropic regularization that is most frequently used in practice. This is since EGW is computable via iterative optimization routines that employ Sinkhorn iterations [49, 51, 54, 58], which allows scalability and parallelization in large-scale applications.

Follow-up works. We address two follow-up works that appeared on arXiv several months after the original submission of this work (and its upload to arXiv on March 2). The paper [51], by one of the authors of the current work and other collaborators, leveraged the duality theory proposed herein to study algorithms, limit theorems and resampling methods for the EGW distance. Their approach relied on a stability analysis of the dual formulation in **A**, based on which L-smoothness and sufficient conditions for convexity of the objective function were derived. These, in turn, were used to propose the first algorithms for computing EGW in $O(n^2)$ time (where n is the number of support points of the two marginals) subject to formal convergence guarantees in both the convex and nonconvex regimes. That work also considered stability of the dual in the marginals μ, ν, which led to a limit distribution theory for the empirical EGW distance and, under additional conditions, asymptotic normality, bootstrap consistency and semiparametric efficiency. Our results along with those from [51] now provide the statistical and computational foundations for valid estimation and inference for the EGW distance, with efficient implementations via the aforementioned algorithms.

Another notable follow-up work is [33], which appeared online two months after our paper was posted to arXiv and submitted to the journal. That work studied the LCA principle from [37] under the EOT setting. In particular, they observed that the dependence on dimension in our empirical convergence rate bounds can be relaxed from $\max\{d_x, d_y\}$ to $\min\{d_x, d_y\}$, provided that the populations are compactly supported. For EGW, as the rate is parametric and dimension-free, this observation only serves to improve the constant. Furthermore, our EGW bounds hold for distributions with unbounded supports, which are beyond the scope of [33]. For the standard GW distance, our original submission proved an $n^{-2/\max\{d_x,d_y,4\}}$ upper bound on the two-sample rate, but Remark 5.6 of [33] observed that it can be improved to $n^{-2/\max\{\min\{d_x,d_y\},4\}}$ and provided high-level proof outline. Herein we provide a full proof of the two-sample upper bound with the dependence on the smaller dimension, and also establish new lower bounds that demonstrate the sharpness of the derived one- and two-sample empirical convergence rates. The reader is referred to Remarks 3.6 and 4.3 for a detailed discussion and comparison to [33].

1.3. *Organization.* The rest of this paper is organized as follows. In Section 2, we collect background material on the OT, EOT, GW, and EGW problems. Section 3 treats the EGW distance, covering stability in the regularization parameter, duality and sample complexity. In Section 4, we extend the duality and the statistical treatment to the (unregularized) GW distance itself. Section 5 contains proofs for Sections 3 and 4. Section 6 leaves concluding remarks and discusses future directions. The Appendix contains proofs of technical results that are omitted from the main text.

1.4. *Notation.* Let $\|\cdot\|$ and $\langle\cdot,\cdot\rangle$ denote the Euclidean norm and inner product, respectively. Let $B_d(x,r) := \{y \in \mathbb{R}^d : \|y-x\| \le r\}$ 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 pth absolute moment, that is, $M_p(\rho) := \int_{\mathbb{R}^d} \|x\|^p \, d\rho(x) < \infty$ 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 $\rho f := \int f \, d\rho$, 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 measures $(\rho_n)_{n\in\mathbb{N}}$ that weakly converges to ρ, we write $\rho_n \xrightarrow{w} \rho$. A probability distribution $\rho \in \mathcal{P}(\mathbb{R}^d)$ is called β-sub-Weibull with parameter σ^2 for $\sigma \geq 0$ if $\int \exp(\|x\|^\beta/2\sigma^2)\,d\rho(x) \leq 2$. In particular, ρ is sub-Gaussian if it is 2-sub-Weibull. Notice that $X \sim \rho$ being 4-sub-Weibull is equivalent to $\|X\|^2$ being sub-Gaussian, in which case $\int e^{t\|x\|^2}\,d\rho(x) \leq 2e^{t^2\sigma^2/2}$. The latter bound is repeatedly used in our derivations.

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 $\|f\|_{\mathrm{Lip}} := \sup_{x\neq x'} \frac{|f(x)-f(x')|}{\|x-x'\|}$. 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 $\|f\|_{L^p(\rho)} := (\int_{\mathbb{R}^d} |f|^p\,d\rho)^{1/p} < \infty$. For any multi-index $k = (k_1,\dots,k_d) \in \mathbb{N}_0^d$ with $|k| = \sum_{j=1}^d k_j$ ($\mathbb{N}_0 = \mathbb{N} \cup \{0\}$), define the differential operator $D^k = \frac{\partial^{|k|}}{\partial x_1^{k_1}\cdots\partial x_d^{k_d}}$ 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 and entropic optimal transport.* We briefly review basic definitions and results concerning the classical and entropic OT problems, which serve as building blocks for our subsequent analysis of the GW distance. For a detailed exposition, the reader is referred to [48, 53, 65]. 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.

2.1.1. *Optimal transport.* The OT problem between $(\mu, \nu) \in \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{Y})$ with cost c is

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

where $\Pi(\mu, \nu)$ is the set of all couplings of μ and ν, that is, 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 $\mathsf{W}_p(\mu,\nu) := (\mathsf{OT}_{\|\cdot\|^p}(\mu,\nu))^{1/p}$. W_p is a metric on $\mathcal{P}_p(\mathbb{R}^d)$ which metrizes weak convergence plus convergence of pth moments, that is, $\mathsf{W}_p(\hat{\mu}_n, \mu) \to 0$ 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. [65], 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 $\Phi_c := \{(\varphi,\psi) \in C_b(\mathcal{X}) \times C_b(\mathcal{Y}) : \varphi(x) + \psi(y) \leq c(x,y), \forall (x,y) \in \mathcal{X}\times\mathcal{Y}\}$. Furthermore, defining the c- and $\bar{c}$-transform of $\varphi \in C_b(\mathcal{X})$ and $\psi \in C_b(\mathcal{Y})$ as $\varphi^c(y) := \inf_{x\in\mathcal{X}} c(x,y) - \varphi(x)$ and $\psi^{\bar{c}}(x) := \inf_{y\in\mathcal{Y}} c(x,y) - \psi(y)$, respectively, the optimization above can be restricted to pairs (φ, ψ) such that $\psi = \varphi^c$ and $\varphi = \psi^{\bar{c}}$.

REMARK 2.1 (Duality for statistical OT). The dual form in (3) is key for the statistical analysis of OT, encompassing empirical convergence rates [10, 18, 37, 42] and limit distribution theorems [14, 27, 36, 41, 59, 62]. For instance, if optimal dual potentials in (3) lie, respectively, in functional classes $\mathcal{F}$ and $\mathcal{G}$, one can bound the two-sample error as

$$\mathbb{E}\big[\big|\mathsf{OT}_c(\mu,\nu)-\mathsf{OT}_c(\hat{\mu}_n,\hat{\nu}_n)\big|\big]\lesssim\mathbb{E}\Big[\sup_{\varphi\in\mathcal{F}_c}(\mu-\hat{\mu}_n)\varphi\Big]+\mathbb{E}\Big[\sup_{\psi\in\mathcal{G}_c}(\nu-\hat{\nu}_n)\psi\Big].$$

This reduces the error analysis to that of the expected suprema of two empirical processes indexed by the classes $\mathcal{F}_c$ and $\mathcal{G}_c$. One may then use techniques from empirical process theory [64] to obtain the desired convergence rate. This requires studying regularity of optimal dual potentials to obtain bounds on the covering numbers of the corresponding function classes. Given bounds of the form $N(\varepsilon,\mathcal{F}_c,L^\infty)\vee N(\varepsilon,\mathcal{G}_c,L^\infty)\lesssim\varepsilon^{-k}$, a convergence rate of $n^{-1/k}$ immediately follows by standard chaining arguments and entropy integral bounds.

This argument can be readily used to prove an $n^{-1/d}$ rate for empirical estimation of the 1-Wasserstein distance (relying on the fact that the covering number of the Lipschitz class scales like $\varepsilon^{-1/d}$), although Dudley, who originally established this rate [18], used a different approach that chained the space of Lipschitz functions using a dyadic procedure. More recently, [42] refined this argument to establish a two-sample rate of $n^{-2/d}$, under smoothness and convexity assumptions on the cost, by observing that dual potentials are not only Lipschitz but also convex in that case. This yielded a covering number bound of $\varepsilon^{-d/2}$ and the rate follows. The primal form was used in [6, 17, 20, 66] to derive sharp empirical convergence rates for the p-Wasserstein distance via a block (dyadic) partitioning argument for the optimal coupling. More recently, however, duality enabled an even finer statistical analysis for deriving limit distribution theorems of empirical OT, for example, via linearization arguments [14, 41] or by proving weak convergence of the underlying empirical processes and invoking the functional delta method [52, 57], as done in [27, 36, 59, 62].

2.1.2. *Entropic optimal transport.* EOT is a convexification of the classical OT problem by means of an entropic penalty. For a regularization parameter $\varepsilon>0$, EOT is given by

$$\mathsf{OT}_{c,\varepsilon}(\mu,\nu):=\inf_{\pi\in\Pi(\mu,\nu)}\int c\,d\pi+\varepsilon\mathsf{D}_{\mathsf{KL}}(\pi\|\mu\otimes\nu),\tag{4}$$

where the KL divergence is given by $\mathsf{D}_{\mathsf{KL}}(\alpha\|\beta):=\int\log(d\alpha/d\beta)\,d\alpha$ if $\alpha\ll\beta$ and equals $+\infty$ otherwise. The optimization objective in (4) is strongly convex in π and thus admits a unique solution $\pi^\star$. The entropic cost $\mathsf{OT}_{c,\varepsilon}$ and the optimal solutions $\pi^\star$ are known to converge towards the classical OT cost [21] and a corresponding optimal plan [8] as $\varepsilon\to0$.[2] In particular, Theorem 1 from [21] shows that for smooth costs and compact spaces $\mathcal{X}$, $\mathcal{Y}$, the entropic approximation gap is $|\mathsf{OT}_\varepsilon(\mu,\nu)-\mathsf{OT}(\mu,\nu)|\lesssim\varepsilon\log(1/\varepsilon)$.

EOT satisfies duality and can be rewritten as (cf. [47]):

$$\begin{aligned}\mathsf{OT}_{c,\varepsilon}(\mu,\nu)=&\sup_{(\varphi,\psi)\in L^1(\mu)\times L^1(\nu)}\int\varphi\,d\mu+\int\psi\,d\nu\\&-\varepsilon\int e^{\frac{\varphi(x)+\psi(y)-c(x,y)}{\varepsilon}}\,d\mu\otimes\nu(x,y)+\varepsilon.\end{aligned}\tag{5}$$

There exist functions $(\varphi,\psi)\in L^1(\mu)\times L^1(\nu)$ that achieve the supremum in (5), which we call *EOT potentials*. EOT potentials are almost surely (a.s.) unique up to additive constants in the sense that if $(\tilde{\varphi},\tilde{\psi})$ is another pair of EOT potentials, then there exists a constant $a\in\mathbb{R}$ such that $\tilde{\varphi}=\varphi+a$ μ-a.s. and $\tilde{\psi}=\psi-a$ ν-a.s. Furthermore, a pair of functions $(\varphi,\psi)\in L^1(\mu)\times L^1(\nu)$ are EOT potentials if and only if they satisfy the so-called *Schrödinger system*

$$\int_{\mathcal{X}}e^{\frac{\varphi(x)+\psi(\cdot)-c(x,\cdot)}{\varepsilon}}\,d\mu(x)=1\quad\nu\text{-}a.s.\quad\text{and}\quad\int_{\mathcal{Y}}e^{\frac{\varphi(\cdot)+\psi(y)-c(\cdot,y)}{\varepsilon}}\,d\nu(y)=1\quad\mu\text{-}a.s.\tag{6}$$

[2]For the plan, convergence happens in the weak topology and possibly along a subsequence.

Given EOT potentials (φ, ψ), the unique EOT plan can be expressed in their terms as $d\pi^\star(x, y) = e^{\frac{\varphi(x)+\psi(y)-c(x,y)}{\varepsilon}} d\mu \otimes \nu(x, y)$.

REMARK 2.2 (Duality for statistical EOT). Akin to the utility of duality for the statistical analysis of OT, as discussed in Remark 2.1, the EOT dual served a pivotal role in the development of a statistical theory under entropic regularization. Thanks for the Schrödinger system in (6), smoothness of the cost c implies the existence of EOT potentials that reside in a Hölder space of arbitrarily large smoothness; cf., for example, [28], Lemma 1. This, in turn, enables establishing parametric $n^{-1/2}$ convergence rates for EOT [21, 33, 46, 50] under quite general conditions and a rich limit distribution theory for the EOT cost, plan, dual potentials and the barycentric projection [4, 13, 28–30, 34, 38, 46]. Recently, statistical results for EOT were extended beyond smooth cost functions. In [50], parametric estimation rates of the EOT cost, optimal coupling and the induced map (sometimes termed, the barycentric projection) were derived assuming only boundedness of the cost function. Their argument employed duality, but circumvented the need to control the suprema of an empirical process indexed by dual potential, thereby ridding of the cost smoothness assumption. Limit theorems for the EOT with a bounded cost function were later obtained in [29], while a method to estimate intrinsic dimension via EOT with Lipschitz cost was proposed in [60].

2.2. *Classical and entropic Gromov–Wasserstein distance.* The objects of interest in this work are the GW distance and its entropic version. 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 [45, 61].

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

where $\Delta_q^{\mathcal{X},\mathcal{Y}}(x, y, x', y') = |\mathsf{d}_\mathcal{X}(x, x')^q - \mathsf{d}_\mathcal{Y}(y, y')^q|$. This definition is an L^p relaxation of the Gromov–Hausdorff distance between metric spaces,[3] and gives rise to a metric on the collection of all isomorphism classes of mm spaces[4] with finite pq-size, that is, $\int \mathsf{d}_\mathcal{X}(x, x')^{pq} d\mu \otimes \mu(x, x') < \infty$ and similarly for ν. Like the p-Wasserstein distance, Theorem 5.1 in [45] 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 $\hat{\mu}_n := n^{-1}\sum_{i=1}^n \delta_{X_i}$ 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

$$\begin{aligned}\mathsf{D}(\mu, \nu) &:= \inf_{\pi \in \Pi(\mu,\nu)} \|\Delta\|_{L^2(\pi\otimes\pi)} \\ &= \inf_{\pi \in \Pi(\mu,\nu)} \left(\int_{\mathbb{R}^{d_x}\times\mathbb{R}^{d_y}} \int_{\mathbb{R}^{d_x}\times\mathbb{R}^{d_y}} \big|\|x - x'\|^2 - \|y - y'\|^2\big|^2 \, d\pi \otimes \pi(x, y, x', y') \right)^{\frac{1}{2}}.\end{aligned} \tag{7}$$

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 $\nu \in \mathcal{P}_4(\mathbb{R}^{d_y})$. We also treat the

[3]The Gromov–Hausdorff distance between $(\mathcal{X}, \mathsf{d}_\mathcal{X})$ and $(\mathcal{Y}, \mathsf{d}_\mathcal{Y})$ is $\frac{1}{2}\inf_{R\in\mathcal{R}(\mathcal{X},\mathcal{Y})} \|\Delta_{1,1}^{\mathcal{X},\mathcal{Y}}\|_{L^\infty(R)}$, where $\mathcal{R}(\mathcal{X}, \mathcal{Y})$ is the collection of all correspondence sets of $\mathcal{X}$ and $\mathcal{Y}$, that is, 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.

[4]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 $f_\sharp\mu = \nu$.

GW distance with entropic regularization, which, for $\varepsilon > 0$, is defined as

$$\mathsf{S}_\varepsilon(\mu,\nu) := \inf_{\pi\in\Pi(\mu,\nu)} \|\Delta\|^2_{L^2(\pi\otimes\pi)} + \varepsilon \mathsf{D}_{\mathsf{KL}}(\pi\|\mu\otimes\nu).$$

The motivation for EGW stems from its computational tractability [49, 51, 54, 58], which makes it a popular approach in practice.[5] With the setup above, we have $\mathsf{S}_0(\mu,\nu) = \mathsf{D}(\mu,\nu)^2$ but approximation bounds that account for the gap $|\mathsf{S}_\varepsilon(\mu,\nu) - \mathsf{D}(\mu,\nu)^2|$ are currently unknown (nor is there a proof of weak convergence for the corresponding optimal couplings). Another major gap in GW and EGW theory is the lack of dual formulations, without which an empirical convergence rate analysis of standard and entropic GW distances remained obscure. In what follows, we close these gaps.

3. Entropic Gromov–Wasserstein distance.

3.1. *Continuity in regularization parameter.* We study continuity of the EGW cost and optimal coupling in ε. Our first result quantifies the gap between the GW and EGW costs.

PROPOSITION 3.1 (Cost approximation gap). *For any $\varepsilon \in (0,1]$ and $(\mu,\nu) \in \mathcal{P}_4(\mathbb{R}^{d_x}) \times \mathcal{P}_4(\mathbb{R}^{d_y})$, we have*

$$\left|\mathsf{S}_\varepsilon(\mu,\nu) - \mathsf{D}(\mu,\nu)^2\right| \lesssim_{d_x,d_y,M_4(\mu),M_4(\nu)} \varepsilon \log\frac{1}{\varepsilon}.$$

The proof of Proposition 3.1, which is given in Section A of the Supplementary Material [70], relies on a block approximation of optimal GW couplings and the Gaussian maximum entropy inequality. Specifically, we decompose the space into cubes with side length ℓ and construct a new coupling π^ℓ that is piecewise uniform on these cubes (the block approximation idea for the EOT coupling originally dates back to [8]). The error of the entropic approximation is then quantified in terms of ℓ, with the KL divergence term being bounded using the differential entropy of the Gaussian distribution with a matched covariance matrix. We then optimize the bound over ℓ to arrive at the desired dependence on ε.

REMARK 3.2 (Comparison to EOT approximation results). A similar bound of order $O(\varepsilon\log(1/\varepsilon))$ was derived in Theorem 1 of [21] for the entropic approximation gap of the OT problem on compact domains with Lipschitz cost. Our proof of Proposition 3.1, which relies on a block approximation of optimal GW couplings, is inspired by their derivation but with several key differences. Specifically, by leveraging the Gaussian maximum entropy inequality, we allow for arbitrary distributions with bounded 4th moments (which is always required for finiteness of D) and costs that grow at most polynomially. Our proof technique can directly be used to relax the assumptions of [21], Theorem 1, to match those of Proposition 3.1. Another related entropic approximation result appeared in Theorem 1 of [10], providing an $O(\varepsilon^2)$ bound on the gap between the squared 2-Wasserstein distance and the Sinkhorn divergence (which is a centered version of EOT). Their derivation utilizes a dynamical formulation of the Sinkhorn divergence [9, 12, 22, 23], which allows tying it to the Benamou–Brenier formula for W_2^2 [3]. No dynamical form for the GW distance is currently known.

[5] As discussed in Section 1.2, the follow-up work [51] proposed the first algorithms for computing EGW between discrete distributions on n points to arbitrary precision in $O(n^2)$ time, subject to formal convergence guarantees. These algorithms hinge upon the dual formulation developed herein. This progress, along with the statistical theory we provide, poses EGW as a viable tool for statistical estimation and inference.

Proposition 3.1 guarantees the convergence of the EGW cost towards that of GW, as $\varepsilon \to 0$. It is natural to ask whether the optimal couplings that achieve these costs converge as well? We answer this question to the affirmative.

PROPOSITION 3.3 (Convergence of plans). *Fix $(\mu, \nu) \in \mathcal{P}_4(\mathbb{R}^{d_x}) \times \mathcal{P}_4(\mathbb{R}^{d_y})$ and let $(\varepsilon_k)_{k\in\mathbb{N}}$ be a sequence with $\varepsilon_k \searrow \varepsilon \geq 0$. For each $k \in \mathbb{N}$, let $\pi_k \in \Pi(\mu, \nu)$ be an optimal coupling for $\mathsf{S}_{\varepsilon_k}(\mu, \nu)$. Then there exists $\pi \in \Pi(\mu, \nu)$ such that $\pi_k \xrightarrow{w} \pi$ as $k \to \infty$ along a subsequence, and π is optimal for $\mathsf{S}_\varepsilon(\mu, \nu)$.*

The proof of Proposition 3.3, which is given in Section B of the Supplementary Material [70], relies on establishing Γ-convergence of the EGW functional with ε. Having that, convergence of optimal couplings follows by a tightness argument. In particular, this result implies that a sequence of optimal couplings for $\mathsf{S}_\varepsilon(\mu, \nu)$ converges, up to extracting a subsequence, to an optimal coupling for the regular $(2, 2)$-GW distance as $\varepsilon \to 0$.

3.2. *Duality.* We next derive a dual formulation for the EGW distance. This duality serves as the key component for our sample complexity analysis of empirical EGW in the next subsection. Towards the dual form, first observe that S_ε 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, that is, $\int x\, d\mu(x) = \int y\, d\nu(y) = 0$.

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

$$\mathsf{S}_\varepsilon(\mu, \nu) = \mathsf{S}^1(\mu, \nu) + \mathsf{S}^2_\varepsilon(\mu, \nu), \tag{8}$$

where

$$\begin{aligned}
&\mathsf{S}^1(\mu, \nu) \\
&\quad := \int \|x - x'\|^4\, d\mu \otimes \mu(x, x') + \int \|y - y'\|^4\, d\nu \otimes \nu(y, y') \\
&\qquad - 4\int \|x\|^2 \|y\|^2\, d\mu \otimes \nu(x, y), \\
&\mathsf{S}^2_\varepsilon(\mu, \nu) \\
&\quad := \inf_{\pi \in \Pi(\mu,\nu)} \int -4\|x\|^2\|y\|^2\, d\pi(x, y) - 8 \sum_{\substack{1\leq i\leq d_1 \\ 1\leq j\leq d_2}} \left(\int x_i y_j\, d\pi(x, y)\right)^2 + \varepsilon \mathsf{D}_{\mathsf{KL}}(\pi \| \mu \otimes \nu).
\end{aligned}$$

See (14) for the derivation. Evidently, the first term depends only on the marginals μ, ν, while the second captures the dependence on the coupling π. The following theorem establishes duality for $\mathsf{S}^2_\varepsilon(\mu, \nu)$, which, in turn, yields a dual form for $\mathsf{S}_\varepsilon(\mu, \nu)$ via the above decomposition.

THEOREM 3.4 (Entropic GW duality). *Fix $\varepsilon > 0$, 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*

$$\mathsf{S}^2_\varepsilon(\mu, \nu) = \inf_{\mathbf{A}\in\mathbb{R}^{d_x\times d_y}} 32\|\mathbf{A}\|_{\mathrm{F}}^2 + \mathsf{OT}_{\mathbf{A},\varepsilon}(\mu, \nu), \tag{9}$$

where $\mathsf{OT}_{\mathbf{A},\varepsilon}$ is the EOT 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 $\mathbf{A}^\star \in \mathcal{D}_{M_{\mu,\nu}} := [-M_{\mu,\nu}/2, M_{\mu,\nu}/2]^{d_x\times d_y}$.

The variational representation above relates the EGW to the well-understood problem of EOT. This enables leveraging knowledge on the latter to make progress in the study of EGW. In particular, this representation unlocks the sample complexity analysis in the next subsection, which relies on inserting the EOT dual from (5) into the above. Since (9) allows utilizing EOT duality for the EGW analysis, we synonymously refer to it as the EGW dual (even though it is somewhat of a misnomer, since strictly speaking, (9) is not a dual problem for $\mathsf{S}_\varepsilon(\mu,\nu)$ in the standard optimization theory sense).

The proof of Theorem 3.4 is given in Section 5.1. The key idea in deriving the above representation is to introduce the additional dual variable $\mathbf{A}$ as a means to linearize the quadratic (in fact, concave) in π term of $\mathsf{S}^2_\varepsilon(\mu,\nu)$. The resulting objective comprises two infima, over $\mathbf{A}\in\mathcal{D}_{M_{\mu,\nu}}$ and $\pi\in\Pi(\mu,\nu)$, which we may interchange. Upon doing so, we identify the inner optimization as the primal EOT problem (up to a minus sign) with the cost $c_{\mathbf{A}}$. Existence of an optimal $\mathbf{A}$ follow from continuity of the functional and compactness of the optimization domain. As the optimum is always achieved inside $\mathcal{F}_{M_{\mu,\nu}}$, we may restrict the optimization domain to $\mathbf{A}\in\mathcal{D}_M$, for any $M\ge M_{\mu,\nu}$, without changing the value. The flexibility of choosing M an optimizing over the compact set $\mathcal{D}_M$ is crucial for our sample complexity analysis.

3.3. *Sample complexity.* The dual formulation from Theorem 3.4 enables deriving, for the first time, the sample complexity of empirical EGW distances. Let $X_1,\dots,X_n$ and $Y_1,\dots,Y_n$ be independently and identically distributed (i.i.d.) samples from μ and ν, respectively, and denote their empirical measures by $\hat\mu_n=n^{-1}\sum_{i=1}^n\delta_{X_i}$ and $\hat\nu_n=n^{-1}\sum_{i=1}^n\delta_{Y_i}$. We study one- and two-sample empirical convergence, that is, the rate at which $\mathsf{S}_\varepsilon(\hat\mu_n,\nu)$ and $\mathsf{S}_\varepsilon(\hat\mu_n,\hat\nu_n)$ approach $\mathsf{S}_\varepsilon(\mu,\nu)$, under a sub-Weibull condition on the population distributions.

THEOREM 3.5 (Entropic GW sample complexity). *Fix $\varepsilon>0$ and let $(\mu,\nu)\in\mathcal{P}(\mathbb{R}^{d_x})\times\mathcal{P}(\mathbb{R}^{d_y})$ be a pair of 4-sub-Weibull distributions with parameter $\sigma^2>0$. We have*

$$\mathbb{E}\big[\big|\mathsf{S}_\varepsilon(\mu,\nu)-\mathsf{S}_\varepsilon(\hat\mu_n,\nu)\big|\big]\lesssim_{d_x,d_y}\frac{1+\sigma^4}{\sqrt n}+\varepsilon\left(1+\left(\frac{\sigma}{\sqrt\varepsilon}\right)^{9\lceil\frac{d_x}{2}\rceil+11}\right)\frac{1}{\sqrt n},$$

$$\mathbb{E}\big[\big|\mathsf{S}_\varepsilon(\mu,\nu)-\mathsf{S}_\varepsilon(\hat\mu_n,\hat\nu_n)\big|\big]\lesssim_{d_x,d_y}\frac{1+\sigma^4}{\sqrt n}+\varepsilon\left(1+\left(\frac{\sigma}{\sqrt\varepsilon}\right)^{9\lceil\frac{d_x\vee d_y}{2}\rceil+11}\right)\frac{1}{\sqrt n}.$$

Theorem 3.5 is derived in Section 5.2. Here, we provide a proof outline, and explain how the duality from Theorem 3.4 facilitates the derivation. The proof follows three main steps:

1. Decomposition: We first split the empirical estimation error of S_ε to that of its components S^1 and S^2_ε. Notice that the decomposition is not straightforward since $\mathsf{S}_\varepsilon=\mathsf{S}^1+\mathsf{S}^2_\varepsilon$ holds only for centered measures, and while we may assume this w.l.o.g. on the populations (μ,ν), centering need not hold for the empirical measures. Thus, to perform the split we first center $(\hat\mu_n,\hat\nu_n)$ by their sample means, and then further account for the bias induced by this centering step, which is shown to be at most $\sigma^2/\sqrt n$. Altogether, we obtain

$$\begin{aligned}&\mathbb{E}\big[\big|\mathsf{S}_\varepsilon(\mu,\nu)-\mathsf{S}_\varepsilon(\hat\mu_n,\hat\nu_n)\big|\big]\\&\quad\le\mathbb{E}\big[\big|\mathsf{S}^1(\mu,\nu)-\mathsf{S}^1(\hat\mu_n,\hat\nu_n)\big|\big]+\mathbb{E}\big[\big|\mathsf{S}^2_\varepsilon(\mu,\nu)-\mathsf{S}^2_\varepsilon(\hat\mu_n,\hat\nu_n)\big|\big]+\frac{\sigma^2}{\sqrt n},\end{aligned}$$

and may analyze each component separately.

2. S^1 analysis: The first term on the right-hand side (RHS) above is simple to analyze, as estimation of S^1 boils down to estimating moments of (μ,ν). Since the sub-Weibull condition implies finite moments, we establish an $O(1/\sqrt n)$ bound on the S^1 estimation error.

3. S_ε^2 analysis: The treatment of the S_ε^2 is more involved and hinges on the dual representation from Theorem 3.4. Specifically, using our dual with any $M \geq M_{\mu,\nu}$, we obtain

$$\left|\mathsf{S}_\varepsilon^2(\mu,\nu) - \mathsf{S}_\varepsilon^2(\hat{\mu}_n,\hat{\nu}_n)\right| \leq \sup_{\mathbf{A}\in\mathcal{D}_M} \left|\mathsf{OT}_{\mathbf{A},\varepsilon}(\mu,\nu) - \mathsf{OT}_{\mathbf{A},\varepsilon}(\hat{\mu}_n,\hat{\nu}_n)\right|,$$

where the RHS can be controlled by the suprema of empirical processes indexed by optimal entropic potentials. As the potentials depend on the cost $c_{\mathbf{A}}$, we analyze regularity of optimal (φ, ψ) pairs by bounding these functions and their partial derivative of arbitrary order, uniformly in $\mathbf{A} \in \mathcal{D}_M$. Given the derivative bounds, a chaining argument and entropy integral bounds yield the second term on the RHS above as a bound on the empirical convergence rate for S_ε^2. The overall rate we obtain is parametric, and hence optimal, although the dependence of the bound on σ and ε could possibly be improved.

REMARK 3.6 (Dependence on dimension). The empirical convergence rate of EGW given in Theorem 3.5 is parametric, and hence cannot be improved. The dependence of the constant in the two-sample bound on the maximal dimension, however, can be relaxed. The follow-up work [33], which was posted on arXiv several months after our original submission and arXiv upload, observed that the dependence on dimension can be improved from $d_x \vee d_y$ to $d_x \wedge d_y$, for compactly supported populations. That work studied the LCA principle from [37] in the context of EOT. Relying on our duality theory, Theorem 5.4 of [33] showed that, when μ, ν are compactly supported, an empirical convergence rate with $d_x \wedge d_y$ instead of $d_x \vee d_y$ in the constant holds true.[6] This result does not cover the full scope of Theorem 3.5, which treats unboundedly supported distributions with 4-sub-Weibull tails. The LCA principle, in its current form, is not compatible with the case where both distributions have unbounded supports, since the argument relies on uniform metric entropy bounds of the dual potential class. Still, [33], Theorem 3.13, was able to extend it to the EOT problem with quadratic cost and only one unboundedly supported distribution, assuming that it is sub-Gaussian.

REMARK 3.7 (Comparison to EOT). The EGW empirical convergence rates from Theorem 3.5 are similar to the corresponding rates for the EOT problem, which are also parametric. Specifically, the $n^{-1/2}$ rate was established in [21] for EOT between compactly supported distributions and assuming that the cost is $\mathcal{C}^\infty$ and Lipschitz, although their bound contained an undesirable exponential dependence on $1/\varepsilon$. This result was extended to sub-Gaussian distributions and quadratic cost in [46], while shaving off the said exponential factor and arriving to a bound that is similar to ours. More recently, [33] observed that the LCA principle holds for EOT, showing that the constant in front of the $n^{-1/2}$ term adapts to the smaller intrinsic dimension of the two measures.

Our approach for proving Theorem 3.5 is inspired by [46], but requires overcoming several new challenges. First, a strong duality theory, which is at the core of the proof technique, was not available until now for the EGW distance. Second, our analysis goes through the decomposition (8), which needs the distributions to be centered. While we may assume this w.l.o.g. on μ, ν, the empirical measures are generally noncentered, which necessitates a bias analysis of the EGW functional due to centering, as discussed above. Lastly, as the dual form in (9) involves optimization over $\mathbf{A} \in \mathcal{D}_M$, with $M \geq M_{\mu,\nu}$, our regularity analysis of EGW potentials must hold uniformly in $\mathbf{A}$, so as to allow the reduction to empirical processes.

[6]More precisely, [33] improves the exponent of the $\varepsilon^{-1/2}$ term in the two-sample rate bound from Theorem 3.5 to $d_x \wedge d_y$, but their overall bound still contains an implicit constant that depends on the maximal dimension $d_x \vee d_y$.

4. Gromov–Wasserstein distance.

4.1. *Duality and sample complexity.* We now consider the unregularized (2, 2)-GW distance from (7), establish duality, derive its sample complexity, and study its one-dimensional structure. Let $(\mu, \nu) \in \mathcal{P}_4(\mathbb{R}^{d_x}) \times \mathcal{P}_4(\mathbb{R}^{d_y})$ be centered w.l.o.g. and note that, similarly to the EGW case, the (2, 2)-GW distance decomposes as

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

where $\mathsf{S}^2 := \mathsf{S}^2_0$, with S^1 and S^2_ε as given after (8). To obtain a dual form for S^2, an inspection of the proof of Theorem 3.4 reveals that the same argument holds also for $\varepsilon = 0$ (i.e., any $\varepsilon \geq 0$ is allowed in that statement), up to replacing the EOT problem $\mathsf{OT}_{\mathbf{A},\varepsilon}$ in (9) with the standard (unregularized) OT problem $\mathsf{OT}_{\mathbf{A}} := \mathsf{OT}_{\mathbf{A},0}$. Recalling the definitions of $M_{\mu,\nu}$, $\mathcal{D}_{M_{\mu,\nu}}$ and $c_{\mathbf{A}}$ from Theorem 3.4, we have the following corollary.

COROLLARY 4.1 (GW duality). *For any $(\mu, \nu) \in \mathcal{P}_4(\mathbb{R}^{d_x}) \times \mathcal{P}_4(\mathbb{R}^{d_y})$, we have*

$$\mathsf{S}^2(\mu, \nu) = \inf_{\mathbf{A}\in\mathbb{R}^{d_x\times d_y}} 32\|\mathbf{A}\|_{\mathrm{F}}^2 + \mathsf{OT}_{\mathbf{A}}(\mu, \nu), \tag{10}$$

where $\mathsf{OT}_{\mathbf{A}}$ is the OT problem with cost $c_{\mathbf{A}}$ and the infimum is achieved at some $\mathbf{A}^\star \in \mathcal{D}_{M_{\mu,\nu}}$.

Given this 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 4.6 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, providing matching upper and lower rate bounds.

THEOREM 4.2 (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, that is, $\mathsf{D}(\mu, \nu) > 0$, 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

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

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

where the latter (two-sample) lower bound further assumes that the samples $X_1, \dots, X_n$ and $Y_1, \dots, Y_n$ are independent of each other.

REMARK 4.3 (Chronology of results). The originally submitted version of this work included only upper bounds on the one- and two-sample empirical convergence rates of D, where the dependence on dimension was through the maximum $d_x \vee d_y$, as opposed to the minimum as above. The follow-up work [33], which appeared online two months after our paper was uploaded to arXiv and submitted to the journal, studied the LCA principle in the context of the EOT problem. Remark 5.6 of that work, observed that the LCA principle applies to our original Theorem 4.2 and commented that the dependence on dimension can be improved to $d_x \wedge d_y$. A full proof of that claim was not provided in [33], only a high-level outline of the argument. Herein, in Section 5.3, we provide a full derivation of the upper bounds with the dependence on the smaller dimension. In addition, we establish a novel lower bound that demonstrates that these empirical convergence rates are sharp.

Theorem 4.2 is proven in Section 5.3. The upper bounds leverage the duality from Corollary 4.1 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 [37], 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 in the absence of entropic regularization and is in line with empirical convergence rates for OT; see Remark 4.6 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 $\mathbb{R}^d$ [32] (see also [31, 55]). 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 4.4 (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

$$\big(8\big(\lambda_{\min}(\mathbf{\Sigma}_\mu)^2 + \lambda_{\min}(\mathbf{\Sigma}_\nu)^2\big)\big)^{\frac{1}{4}} \inf_{\mathbf{U}\in E(d)} \mathsf{W}_2(\mu, \mathbf{U}_\sharp \nu) \le \mathsf{D}(\mu, \nu).$$

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

The lemma, which is proven in Section C of the Supplementary Material [70], 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 4.2.

REMARK 4.5 (Suboptimal (p,q)-GW rates from Lemma 4.4). Fix any (p,q) and $\mu \in \mathcal{P}_{pq}(\mathbb{R}^d)$. The upper bound from Lemma 4.4, directly yields $\mathbb{E}[\mathsf{D}_{p,q}(\hat{\mu}_n, \mu)] \lesssim \mathbb{E}[\mathsf{W}_{pq}(\hat{\mu}_n, \mu)] \lesssim n^{-1/d_x}$. Via the triangle inequality we can further obtain an $n^{-1/(d_x \vee d_y)}$ two-sample rate bound for the (p,q)-GW distance. However, as seen from the lower bounds in Theorem 4.2, this rate is suboptimal and does not adapt to the lower of the two dimensions.

REMARK 4.6 (Comparison to OT, unbounded domains and intrinsic dimension). The rates in Theorem 4.2 are inline with those for classical OT with Hölder smooth costs [42] (although our analysis is different from theirs). Over compact domains, 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. OT with unbounded domains is treated in Theorem 13 of [42], 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. Lastly, we note that empirical convergence rates of OT and the Wasserstein distance are known to present a multiscale behavior, adapting to the intrinsic dimension dimension of the supports of μ, ν; cf., for example, [37, 66]. The general form of the LCA principle from [37], Theorem 2.2, is sufficiently fine to capture that since it measures 'dimension' via uniform control over the covering number of the class of dual potentials, which follows from uniform covering of $\mathrm{spt}(\mu)$ and $\mathrm{spt}(\nu)$. Consequently, while the bounds from Theorem 4.2 only invoke the LCA principle w.r.t. the ambient spaces $\mathbb{R}^{d_x}$ and $\mathbb{R}^{d_y}$, they can be readily refined to depend only on the intrinsic dimensions of the supports. We leave a formal derivation of this refinement for future work.

4.2. *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,[7] there is no known simple solution 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 was always induced by the identity or anti-identity permutations [63], it was recently shown that this is not true in general [2]. 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, that is, $\mu = n^{-1}\sum_{i=1}^n \delta_{x_i}$ 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 $x_1 < x_2 < \cdots < x_n$ and $y_1 < y_2 < \cdots < y_n$. To compute $\mathsf{D}(\mu,\nu)$ it suffices to optimize over couplings induced by permutations [63], Theorem 9.2, (see also [43]), that is,

$$\mathsf{D}(\mu,\nu)^2 = \frac{1}{n^2}\min_{\sigma\in S_n}\sum_{i=1}^n\sum_{j=1}^n \big||x_i - x_j|^2 - |y_{\sigma(i)} - y_{\sigma(j)}|^2\big|^2, \tag{11}$$

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

$$x_i^\xi := \begin{cases} -1 & i = 1, \\ \dfrac{2i-n-1}{2}\xi & 2 \le i \le n-1, \\ 1 & i = n \end{cases} \quad \text{and} \quad y_i^\xi := \begin{cases} -1 & i = 1, \\ -1+\xi & i = 2, \\ (i-2)\xi & 3 \le i \le n. \end{cases} \tag{12}$$

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

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

permutation $\sigma_{\mathrm{cyc}}(i) = i+1 \mod n$ between $x^{\xi^\star}$ and $y^{\xi^\star}$ achieves a strictly smaller cost in (11) 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 Corollary 4.1, 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

$$\mathsf{S}^2(\mu,\nu) = \inf_{a\in[0.5W_-,0.5W_+]} 32a^2 + \inf_{\pi\in\Pi(\mu,\nu)} \int (-4x^2y^2 - 32axy)\,d\pi(x,y), \tag{13}$$

where $W_- := \inf_{\pi\in\Pi(\mu,\nu)} \int xy\,d\pi(x,y)$ and $W_+ := \sup_{\pi\in\Pi(\mu,\nu)} \int xy\,d\pi(x,y)$. Here, we have used the fact that, switching the infima order, for each $\pi \in \Pi(\mu,\nu)$, optimality is attained at $a^\star(\pi) = \frac{1}{2}\int xy\,d\pi(x,y)$. 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 (13), the GW problem is reduced to an OT problem. Hence, we investigate the optimization in a. Define $f(a) := 32a^2$ and $g(a) := \inf_{\pi\in\Pi(\mu,\nu)} \int(-4x^2y^2 - 32axy)\,d\pi(x,y)$, and note that g is concave (as the infimum of affine functions). We see that the optimization over a in (13), 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 (11); see Section D of the Supplementary Material [70] for the proof.

PROPOSITION 4.7 (Boundary values and optimal permutations). *Consider the GW problem from* (11) *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* (13). *Let* $\mathcal{S}^\star \subset S_n$ *and* $\mathcal{A}^\star \subset [0.5W_-, 0.5W_+]$ *be the argmin sets for* (11) *and* (13), *respectively. Then* $\mathcal{A}^\star \subset \{0.5W_-, 0.5W_+\}$ *if and only if* $\mathcal{S}^\star \subset \{\mathrm{id}, \overline{\mathrm{id}}\}$.

Proposition 4.7 thus implies that the identity and anti-identity can only optimize the GW distance when (13) 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, Figure 1 plots the two datasets x^ξ and

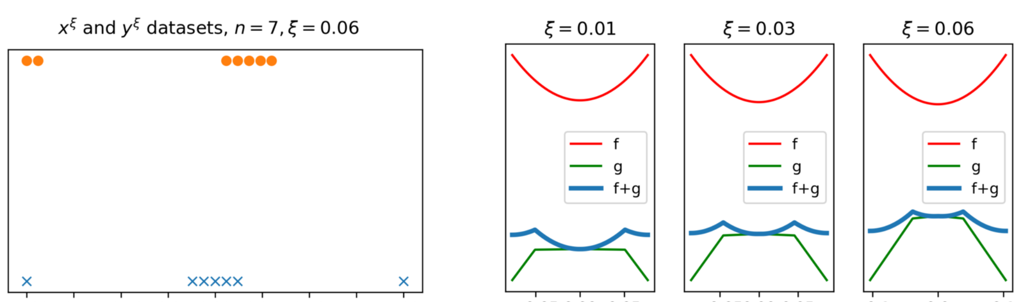

FIG. 1. *(Left) The datasets x^ξ and y^ξ from* (12), *for $n = 7$ and $\xi = 0.06$; (Right) The functions f, g and $f+g$ on the interval $a \in [0.5W_-, 0.5W_+]$, for $\xi = 0.01, 0.03, 0.06$. When $\xi = 0.01$, the minimizer of $f+g$ is attained outside the boundary and thus the corresponding optimal permutation is neither the identity nor the anti-identity.*

y^ξ from (12) 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 4.7, the optimal permutation is different from id and $\overline{\text{id}}$. 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 $\overline{\text{id}}$ in the example above.

5. Proofs of main theorems.

5.1. *Proof of Theorem* 3.4. For completeness, we first show the decomposition of $\mathsf{S}_\varepsilon(\mu,\nu)$ for centered μ, ν, given in (8). Expanding the $(2,2)$-GW cost we have

$$
\begin{aligned}
\mathsf{S}_\varepsilon(\mu,\nu) = &\int \|x-x'\|^4 \, d\mu\otimes\mu(x,x') + \int \|y-y'\|^4 \, d\nu\otimes\nu(y,y') \\
&- 4\int \|x\|^2\|y\|^2 \, d\mu\otimes\nu(x,y) \\
&+ \inf_{\pi\in\Pi(\mu,\nu)} \biggl\{ -4\int \|x\|^2\|y\|^2 \, d\pi(x,y) - 8\int \langle x,x'\rangle\langle y,y'\rangle \, d\pi\otimes\pi(x,y,x',y') \\
&+ 8\int \bigl(\langle x,x'\rangle\|y\|^2 + \|x\|^2\langle y,y'\rangle\bigr)\, d\pi\otimes\pi(x,y,x',y') + \varepsilon \mathsf{D}_{\mathsf{KL}}(\pi\|\mu\otimes\nu)\biggr\}.
\end{aligned} \tag{14}
$$

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 $\mathsf{S}_\varepsilon^2(\mu,\nu)$, respectively.

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

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

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 semi-continuity in π and $\mathbf{A}$. We conclude by identifying the EOT problem $\mathsf{OT}_{\mathbf{A},\varepsilon}$ in the last line.

5.2. *Proof of Theorem* 3.5. We only prove the two-sample case; the one-sample derivation is similar, except that in (16) ahead one would only consider the empirical process induced by μ. Proofs of technical lemmas stated throughout this proof are given in Section D in the Supplementary Material [70]. We proceed with the three steps described in the proof outline, after the theorem statement.

Decomposition: Recall from (8) that $\mathsf{S}_\varepsilon(\mu,\nu) = \mathsf{S}^1(\mu,\nu) + \mathsf{S}^2_\varepsilon(\mu,\nu)$ holds if μ, ν are centered distributions. This decomposition is convenient for analysis as it allows separately treating the marginals- and the coupling-dependents terms. Namely, we would like to have

$$\left|\mathsf{S}_\varepsilon(\mu,\nu) - \mathsf{S}_\varepsilon(\hat{\mu}_n,\hat{\nu}_n)\right| \le \left|\mathsf{S}^1(\mu,\nu) - \mathsf{S}^1(\hat{\mu}_n,\hat{\nu}_n)\right| + \left|\mathsf{S}^2_\varepsilon(\mu,\nu) - \mathsf{S}^2_\varepsilon(\hat{\mu}_n,\hat{\nu}_n)\right|.$$

However, while the EGW distance S_ε is translation invariant and we may assume $\int x\,d\mu = \int y\,d\nu = 0$ 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_ε may not hold. To amend this, we center $\hat{\mu}_n$, $\hat{\nu}_n$ and quantify the bias that this incurs on S_ε. This is stated in the following lemma, which is proven in Section E.1 of the Supplementary Material [70].

LEMMA 5.1 (Centering bias). *If μ, ν are centered, then*

$$\mathbb{E}\left[\left|\mathsf{S}_\varepsilon(\mu,\nu) - \mathsf{S}_\varepsilon(\hat{\mu}_n,\hat{\nu}_n)\right|\right] \lesssim \mathbb{E}\left[\left|\mathsf{S}^1(\mu,\nu) - \mathsf{S}^1(\hat{\mu}_n,\hat{\nu}_n)\right|\right] + \mathbb{E}\left[\left|\mathsf{S}^2_\varepsilon(\mu,\nu) - \mathsf{S}^2_\varepsilon(\hat{\mu}_n,\hat{\nu}_n)\right|\right] + \frac{\sigma^2}{\sqrt{n}}.$$

Given this decomposition, we proceed to separately treat the empirical errors of S^1 and S^2_ε.

Sample complexity of S^1: 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 Section E.2 of the Supplementary Material [70].

LEMMA 5.2 (S^1 parametric rate). *If μ, ν are* 4*-sub-Weibull with parameter $\sigma^2 > 0$, then*

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

Sample complexity of S^2_ε: It remains to analyze the sample complexity of S^2_ε. To that end, we use the dual form of S^2_1 to control its empirical error by the supremum of an empirical process indexed by optimal EGW potentials. We then derive regularity properties of the potentials, based on which standard empirical process techniques via entropy integral bounds yield the desired rate. For ease of presentation, the derivation is split into several steps.

(i) Normalization and reduction to EOT. Observe that if μ^ε, ν^ε are the pushforward measures of μ, ν through the mapping $x \mapsto \varepsilon^{-1/4}x$, then we have $\mathsf{S}^2_\varepsilon(\mu,\nu) = \varepsilon \mathsf{S}^2_1(\mu^\varepsilon,\nu^\varepsilon)$. Also note that μ^ε, ν^ε are 4-sub-Weibull distributions with parameter σ^2/ε. Thus, we henceforth set $\varepsilon = 1$ and later adapt to a general $\varepsilon > 0$ using the aforementioned observation. Invoking Theorem 3.4 for S^2_1, while optimizing over $\mathbf{A} \in \mathcal{D}_M$, for some $M \ge M_{\mu,\nu}$ to be specified later (which does not change the optimization value since $A^\star \in \mathcal{D}_{M_{\mu,\nu}}$), we obtain

$$\left|\mathsf{S}^2_1(\mu,\nu) - \mathsf{S}^2_1(\hat{\mu}_n,\hat{\nu}_n)\right| \le \sup_{\mathbf{A}\in\mathcal{D}_M} \left|\mathsf{OT}_{\mathbf{A},1}(\mu,\nu) - \mathsf{OT}_{\mathbf{A},1}(\hat{\mu}_n,\hat{\nu}_n)\right|, \tag{15}$$

which reduces the analysis to that of EOT with the cost function $c_{\mathbf{A}}$, uniformly over $\mathbf{A} \in \mathcal{D}_M$. We next analyze the regularity of optimal dual potentials for the EOT problems on the RHS above. This regularity theory is later used to decompose the RHS into suprema of empirical processes indexed by these potentials and to analyze their expected convergence rates.

(ii) Smoothness of EOT potentials. To simplify notation, we henceforth drop the subscript $\mathbf{A}$ from the EOT potentials $(\varphi_{\mathbf{A}}, \psi_{\mathbf{A}})$ for $\mathsf{OT}_{\mathbf{A},\varepsilon}(\mu, \nu)$, writing only (φ, ψ). The following lemma provides bounds on the magnitude of partial derivatives (of any order) of EOT potentials between any two sub-Weibull distribution, w.r.t. the cost $c_{\mathbf{A}}$, uniformly in $\mathbf{A} \in \mathcal{D}_M$. To state the result, for any $\sigma, M > 0$, let $\mathcal{F}_{\sigma,M}$ be the class of $\mathcal{C}^\infty(\mathbb{R}^{d_x})$ functions φ satisfying

$$\varphi(x) \le 4\sigma^2 + 8M\sqrt{2\sigma d_x d_y}\big(\sqrt{\sigma} + 2\|x\|^2\big)$$

$$-\varphi(x) \le \log 2 + 4\sigma^2 + 8M\sqrt{2\sigma d_x d_y}\left(1 + \sqrt{\sigma} + \frac{\|x\|^2}{\sqrt{2\sigma}}\right) + 8\sigma^2\big(2M\sqrt{d_x d_y}(1 + \sqrt{2\sigma}) + \|x\|^2\big)^2,$$

$$\big|D^\alpha \varphi(x)\big| \le C_\alpha\big(1 + M\sqrt{d_y} + \|x\|\big)^{|\alpha|}\big(1 + \sigma^{|\alpha|} + (1 + M\sqrt{d_x d_y})^{|\alpha|}(1 + \sigma^5 + \sigma^4\|x\|^4)^{\frac{|\alpha|}{2}}\big),$$

for all multi-indices $\alpha \in \mathbb{N}_0^{d_x}$ and some constant $C_\alpha > 0$ that depends only on α. Define the class $\mathcal{G}_{\sigma,M}$ analogously but for functions $\psi : \mathbb{R}^{d_y} \to \mathbb{R}$.

LEMMA 5.3 (Uniform regularity of EOT potentials). *Fix $M \ge M_{\mu,\nu}$, $\mathbf{A} \in \mathcal{D}_M$ and suppose that μ, ν are 4-sub-Weibull with parameter σ^2. Then there exist optimal EOT potentials (φ, ψ) for $\mathsf{OT}_{\mathbf{A},1}(\mu, \nu)$ from (4), such that $\varphi \in \mathcal{F}_{M,\sigma}$ and $\psi \in \mathcal{G}_{M,\sigma}$.*

The proof of the lemma is deferred to Section E.3 of the Supplementary Material [70]. The key idea is that given optimal EOT potentials (φ_0, ψ_0), we may define new potentials (φ, ψ) via the Schrödinger systems (6) and show that the pairs agree $\mu \otimes \nu$-a.s. Consequently, (φ, ψ) are also optimal for $\mathsf{OT}_{\mathbf{A},1}(\mu, \nu)$, but they enjoy an explicit representation via the Schrödinger systems, which evidently renders (φ, ψ) smooth functions.

Lemma 5.3 allows restricting the optimization domain in the dual form of $\mathsf{OT}_{\mathbf{A},1}$ from $L^1(\mu) \times L^1(\nu)$ to $\mathcal{F}_{M,\tilde{\sigma}} \times \mathcal{G}_{M,\tilde{\sigma}}$, for an appropriately chosen $\tilde{\sigma}$. Let $\tilde{\sigma}$ be the random variables defined as the smallest $\sigma' > 0$ such that μ, ν, $\hat{\mu}_n$, $\hat{\nu}_n$ are all 4-sub-Weibull with parameter σ'^2. Clearly, any $\varphi \in \mathcal{F}_{M,\tilde{\sigma}}$ with $M = \sqrt{2}\tilde{\sigma}$ also satisfies

$$\big|\varphi(x)\big| \le C_{d_x,d_y}\big(1 + \tilde{\sigma}^5\big)\big(1 + \|x\|^4\big),$$

$$\big|D^\alpha \varphi(x)\big| \le C_{\alpha,d_x,d_y}\big(1 + \tilde{\sigma}^{9|\alpha|/2}\big)\big(1 + \|x\|^{3|\alpha|}\big) \quad \forall \alpha \in \mathbb{N}_0^{d_x},$$

and similarly for $\psi \in \mathcal{G}_{M,\tilde{\sigma}}$. Recalling that Theorem 3.4 requires M^2 to be at least as large as the product of the 2nd moments of the involved distributions, we note that $M = \sqrt{2}\tilde{\sigma}$ is feasible for $\mathsf{OT}_{\mathbf{A},1}(\cdot,\cdot)$ between any pair from $\{\mu, \nu, \hat{\mu}_n\}$ and $(\nu, \hat{\nu}_n)$, for any $\mathbf{A} \in \mathcal{D}_M$. Lastly, define the Hölder class

$$\mathcal{F}_s = \big\{\varphi : \mathbb{R}^{d_x} \to \mathbb{R} : |\varphi| \le C_{s,d_x,d_y}\big(1 + \|\cdot\|^4\big), \big|D^\alpha \varphi\big| \le C_{s,d_x,d_y}\big(1 + \|\cdot\|^{3s}\big), \forall |\alpha| \le s\big\},$$

with $\mathcal{G}_s$ defined analogously. We conclude that for each $\mathbf{A} \in \mathcal{D}_M$, any smooth potentials (φ, ψ) for the corresponding EGW problem satisfy $(1 + \tilde{\sigma}^{5s})^{-1}\varphi \in \mathcal{F}_s$ and $(1 + \tilde{\sigma}^{5s})^{-1}\psi \in \mathcal{G}_s$. This regularity of potentials will be used to derive the parametric rate of convergence for empirical S_1^2, following the decomposition presented in the next part.

(iii) Decomposition into suprema of empirical processes. We upper bound the empirical estimation error of S_1^2 by the suprema of empirical processes indexed by optimal potentials. To simplify notation, recall the shorthand $\rho\varphi := \int \varphi \, d\rho$ for any signed Borel measure ρ. Starting from (15), we have

$$\begin{aligned}\big|\mathsf{S}_1^2(\mu, \nu) - \mathsf{S}_1^2(\hat{\mu}_n, \hat{\nu}_n)\big| &\le \sup_{\mathbf{A} \in \mathcal{D}_M} \big|\mathsf{OT}_{\mathbf{A},1}(\mu, \nu) - \mathsf{OT}_{\mathbf{A},1}(\hat{\mu}_n, \hat{\nu}_n)\big| \\ &\lesssim (1 + \tilde{\sigma}^{5s})\Big(\sup_{\varphi \in \mathcal{F}_s}\big|(\mu - \hat{\mu}_n)\varphi\big| + \sup_{\psi \in \mathcal{G}_s}\big|(\nu - \hat{\nu}_n)\psi\big|\Big),\end{aligned} \tag{16}$$

where the second inequality follows by [46], Proposition 2, which uses the fact that the optimal EOT potentials between (μ,ν), $(\hat{\mu}_n,\nu)$ and $(\hat{\mu}_n,\hat{\nu}_n)$ belong to $\mathcal{F}_s\times\mathcal{G}_s$. We have also used the fact that Lemma 5.3 holds uniformly in $\mathbf{A}\in\mathcal{D}_M$ to remove the supremum.

(iv) Sample complexity analysis. We are now in place to establish that $\mathsf{S}_1^2(\hat{\mu}_n,\hat{\nu}_n)$ converges towards $\mathsf{S}_1^2(\mu,\nu)$ at the parametric rate. More specifically, we will show

$$\mathbb{E}\big[\big|\mathsf{S}_1^2(\mu,\nu)-\mathsf{S}_1^2(\hat{\mu}_n,\hat{\nu}_n)\big|\big]\lesssim_{d_x,d_y}\big(1+\sigma^{7\lceil\frac{d_x\vee d_y}{2}\rceil+9}\big)n^{-\frac{1}{2}}. \tag{17}$$

Starting from the RHS of (16), we present the analysis of the first supremum, with the second one being treated similarly. We bound it as

$$\mathbb{E}\Big[\big(1+\tilde{\sigma}^{5s}\big)\sup_{\varphi\in\mathcal{F}_s}\big|(\mu-\hat{\mu}_n)\varphi\big|\Big]\le\sqrt{\mathbb{E}\big[\big(1+\tilde{\sigma}^{5s}\big)^2\big]\mathbb{E}\Big[\Big(\sup_{\varphi\in\mathcal{F}_s}(\mu-\hat{\mu}_n)\varphi\Big)^2\Big]}, \tag{18}$$

and proceed to bound the second term. By Theorem 3.5.1. from [24], we have

$$\mathbb{E}\Big[\Big(\sup_{\varphi\in\mathcal{F}_s}(\mu-\hat{\mu}_n)\varphi\Big)^2\Big]\lesssim_{d_x}\frac{1}{n}\mathbb{E}\bigg(\int_0^{\sqrt{\max_{\varphi\in\mathcal{F}_s}\|\varphi\|^2_{L^2(\hat{\mu}_n)}}}\sqrt{\log\big(2N\big(\xi,\mathcal{F}_s,L^2(\hat{\mu}_n)\big)\big)}\,d\xi\bigg)^2. \tag{19}$$

The integration domain is bounded by observing that

$$\max_{\varphi\in\mathcal{F}_s}\|\varphi\|^2_{L^2(\hat{\mu}_n)}\le C_{d_x,d_y}\frac{1}{n}\sum_{i=1}^n\big(1+\|x_i\|^8\big)\le C_{d_x,d_y}\big(1+\sigma^4L\big),$$

where $L:=\frac{1}{n}\sum_{i=1}^n e^{\frac{\|x_i\|^4}{2\sigma^2}}$ satisfies $\mathbb{E}[L]\le 2$. To control the integrand, we apply Corollary 2.7.4. from [64] (see also [46] Proposition 3) as follows. First, define $Q_j:=[-2^j\sqrt{\sigma},2^j\sqrt{\sigma}]^{d_x}$ for $j\in\mathbb{N}_0$, and partition $\mathbb{R}^{d_x}$ into the sets $I_j=Q_j\backslash Q_{j-1}$. Note that the Lebesgue measure of each $\{x\in\mathbb{R}^{d_x}:\|x-I_j\|\le 1\}$ is bounded by $C_{d_x}(1+2^{jd_x}\sigma^{d_x/2})$, and by Markov's inequality we further obtain $\hat{\mu}_n(I_j)\le Le^{-2^{4j-5}}$. Lastly, for any $j\in\mathbb{N}_0$ and $\varphi\in\mathcal{F}_s$, the restriction $\varphi|_{I_j}$ has a $\mathcal{C}^s(I_j)$-Hölder norm bounded by $C_{s,d_x}(1+\sigma^{3s/2})2^{3js}$. This verifies the conditions of [64], Corollary 2.7.4., which we invoke with $s=\lceil d_x/2\rceil+1$, $V=d_x/s$ and $r=2$, to get

$$\begin{aligned}
&\log N\big(\xi,\mathcal{F}_s,L^2(\hat{\mu}_n)\big)\\
&\quad\le\log N_{[\,]}\big(2\xi,\mathcal{F}_s,L^2(\hat{\mu}_n)\big)\\
&\quad\le C_{d_x,d_y}\xi^{-V}L^{\frac{V}{r}}\Bigg(\sum_{j=0}^{\infty}\big(1+2^{jd_x}\sigma^{\frac{d_x}{2}}\big)^{\frac{r}{V+r}}\big(e^{-2^{4j-5}}\big)^{\frac{V}{V+r}}\big(\big(1+\sigma^{\frac{3s}{2}}\big)2^{3js}\big)^{\frac{Vr}{V+r}}\Bigg)^{\frac{V+r}{r}}\\
&\quad\le C_{d_x,d_y}\xi^{-V}L^{\frac{V}{r}}\big(1+\sigma^{\frac{d_x+3sV}{2}}\big)\Bigg(\sum_{j=0}^{\infty}e^{-\frac{2^{4j-5}V}{V+r}}2^{\frac{(3sV+d_x)jr}{V+r}}\Bigg)^{\frac{V+r}{r}}\\
&\quad\le C_{d_x,d_y}\xi^{-\frac{d_x}{s}}L^{\frac{d_x}{2s}}\big(1+\sigma^{2d_x}\big),
\end{aligned}$$

where the last line follows because the summation is finite and only depends on d_x. Inserting this bound back into (19), we have

$$\begin{aligned}
\mathbb{E}\Big[\Big(\sup_{\varphi\in\mathcal{F}_s}(\mu-\hat{\mu}_n)\varphi\Big)^2\Big]&\lesssim_{d_x,d_y}\frac{1}{n}\mathbb{E}\bigg[\bigg(\int_0^{\sqrt{1+\sigma^4L}}\sqrt{\xi^{-\frac{d_x}{s}}L^{\frac{d_x}{2s}}\big(1+\sigma^{2d_x}\big)}\,d\xi\bigg)^2\bigg]\\
&\lesssim_{d_x,d_y}\frac{(1+\sigma^{2d_x})}{n}\mathbb{E}\big[L^{\frac{d_x}{2s}}\big(1+\sigma^4L\big)^{1-\frac{d_x}{2s}}\big]\\
&\lesssim_{d_x,d_y}\big(1+\sigma^{2(2+d_x)}\big)n^{-1}.
\end{aligned}$$

In light of (18), it remains to bound the appropriate moment of $\tilde{\sigma}$. For any $k \in \mathbb{N}$, set

$$\tau_k^2 = \sigma^2 \vee \left(\frac{k\sigma^2}{n}\sum_{i=1}^n e^{\frac{\|x_i\|^4}{2k\sigma^2}}\right) \vee \left(\frac{k\sigma^2}{n}\sum_{i=1}^n e^{\frac{\|y_i\|^4}{2k\sigma^2}}\right)$$

so that μ, ν, $\hat{\mu}_n$, $\hat{\nu}_n$ are all 4-sub-Weibull with parameter τ_k^2; cf. [46], Lemma 4. Therefore

$$\mathbb{E}[\tilde{\sigma}^{2k}] \le \mathbb{E}[\tau_k^{2k}] \le \sigma^{2k} + \frac{k^k \sigma^{2k}}{n}\mathbb{E}\left[\sum_{i=1}^n e^{\frac{\|x_i\|^4}{2\sigma^2}} + e^{\frac{\|y_i\|^4}{2\sigma^2}}\right] \le (1+4k^k)\sigma^{2k}.$$

Combining all the pieces leads to:

$$\mathbb{E}\Big[(1+\tilde{\sigma}^{5s}) \sup_{\varphi \in \mathcal{F}_s} (\mu - \hat{\mu}_n)\varphi\Big] \lesssim_{d_x,d_y} \sqrt{\mathbb{E}[(1+\tilde{\sigma}^{5s})^2]\frac{(1+\sigma^{2d_x+4})}{n}} \lesssim \frac{(1+\sigma^{9\lceil d_x/2\rceil + 11})}{\sqrt{n}}$$

with a similar bound holding for the corresponding term with $\mathcal{F}_s$ replaced by $\mathcal{G}_s$. Together with (16), these two bounds imply (17). □

5.3. *Proof of Theorem* 4.2.

5.3.1. *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). As in the proof of Theorem 3.5, we only prove the two-sample case. The one-sample result follows similarly. Derivations of technical lemmas stated throughout this proof are deferred to Section E of the Supplementary Material [70].

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 again 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. Following the same steps leading to Eqs. (E.1) and (E.2) of the Supplementary Material [70], we observe that

$$\mathbb{E}\big[|\mathsf{S}^1(\hat{\mu}_n,\hat{\nu}_n) - \mathsf{S}^1(\tilde{\mu}_n,\tilde{\nu}_n)|\big] \vee \mathbb{E}\big[|\mathsf{S}^2(\hat{\mu}_n,\hat{\nu}_n) - \mathsf{S}^2(\tilde{\mu}_n,\tilde{\nu}_n)|\big] \lesssim \frac{R^4}{\sqrt{n}},$$

which also uses the fact that any distribution whose support diameter is bounded by R is trivially 4-sub-Weibull with parameter R^4. Consequently, we may split

$$\begin{aligned} &\mathbb{E}\big[|\mathsf{D}(\mu,\nu)^2 - \mathsf{D}(\hat{\mu}_n,\hat{\nu}_n)^2|\big] \\ &\quad \le \mathbb{E}\big[|\mathsf{S}^1(\mu,\nu) - \mathsf{S}^1(\hat{\mu}_n,\hat{\nu}_n)|\big] + \mathbb{E}\big[|\mathsf{S}^2(\mu,\nu) - \mathsf{S}^2(\hat{\mu}_n,\hat{\nu}_n)|\big] + \frac{R^4}{\sqrt{n}}, \end{aligned} \tag{20}$$

and proceed with a separate analysis for S^1 and S^2.

For S^1, we follow the steps leading to Eq. (E.3) of the Supplementary Material [70] in the EGW sample complexity analysis and use the fact that μ, ν are 4-sub-Weibull with parameter R^4 to deduce

$$\mathbb{E}\big[|\mathsf{S}^1(\hat{\mu}_n,\hat{\nu}_n) - \mathsf{S}^1(\mu,\nu)|\big] \lesssim \frac{R^4}{\sqrt{n}}. \tag{21}$$

To treat S^2, we start from the variational representation from Corollary 4.1 and choose $M = R^2 \ge M_{\mu,\nu}$, which is evidently feasible. Invoking this result, we obtain

$$\big|\mathsf{S}^2(\mu,\nu) - \mathsf{S}^2(\hat{\mu}_n,\hat{\nu}_n)\big| \le \sup_{\mathbf{A} \in \mathcal{D}_{R^2}} \big|\mathsf{OT}_{\mathbf{A}}(\mu,\nu) - \mathsf{OT}_{\mathbf{A}}(\hat{\mu}_n,\hat{\nu}_n)\big|, \tag{22}$$

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

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

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 5.4 (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 Section F.1 of the Supplementary Material [70], 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 5.4, 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 optimal dual potential pairs for $\mathsf{OT}_{\mathbf{A}}(\mu, \nu)$ (see (3)). Define $\mathcal{F}_{\mathbf{A}} := \mathrm{proj}_{\mathcal{F}_R}(\Phi_{\mathbf{A}} \cap (\mathcal{F}_R \times \mathcal{G}_R))$ 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 (22), to arrive at

$$\begin{aligned} &\mathbb{E}\big[\big|\mathsf{S}^2(\mu, \nu) - \mathsf{S}^2(\hat{\mu}_n, \hat{\nu}_n)\big|\big] \\ &\quad \le \mathbb{E}\Big[\sup_{\varphi \in \bigcup_{\mathbf{A}} \mathcal{F}_{\mathbf{A}}} \big|(\mu - \hat{\mu}_n)\varphi\big|\Big] + \mathbb{E}\Big[\sup_{\psi \in \bigcup_{\mathbf{A}} \mathcal{F}_{\mathbf{A}}^c} \big|(\nu - \hat{\nu}_n)\psi\big|\Big]. \end{aligned} \tag{23}$$

As Lemma 5.4 implies that $\bigcup_{\mathbf{A}} \mathcal{F}_{\mathbf{A}} \subset \mathcal{F}_R$, 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

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

Theorem 3.2 from [35] provides a bound on the metric entropy of bounded, convex, Lipschitz functions, whereby if $\tilde{\mathcal{F}}_d := \{f : B_d(0,1) \to \mathbb{R} : f \text{ convex}, \|f\|_\infty \vee \|f\|_{\mathrm{Lip}} \le 1\}$, then $\log N(\xi, \tilde{\mathcal{F}}_d, \|\cdot\|_\infty) \le C_d \xi^{-\frac{d}{2}}$. For any $\varphi : \mathbb{R}^d \to \mathbb{R}$, define its rescaled version[8] $(S\varphi)(z) := \varphi(Rz)/(1 + C_{d_x,d_y} R^4)$, where $C_{d_x,d_y} = 10(1 + 4\sqrt{d_x d_y})$, and note that $S\varphi \in \tilde{\mathcal{F}}_{d_x}$, for any $\varphi \in \mathcal{F}_R$. We also define the map $s : x \mapsto x/R$. Combining the above, for $d_x \ge 4$, we have

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

[8] 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.

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

$$\mathbb{E}\Big[\sup_{\varphi \in \bigcup_{\mathbf{A}} \mathcal{F}_{\mathbf{A}}} \big|(\mu - \hat{\mu}_n)\varphi\big|\Big] \lesssim_{d_x, d_y} (1 + R^4) n^{-\frac{2}{d_x \vee 4}} (\log n)^{\mathbb{1}_{\{d_x = 4\}}}. \tag{24}$$

We now move to treat the second term on the RHS of (23). First, observe that one may control it by the expected supremum of an empirical process indexed by $\mathcal{G}_R$, which is bounded by $(1 + R^4) n^{-2/(d_y \vee 4)} (\log n)^{\mathbb{1}_{\{d_y = 4\}}}$ via similar steps as above. Together with (24), 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 [33], this is possible by employing the LCA principle from [37], Lemma 2.1, which states that for any cost function c and function class $\mathcal{F}$, we have $N(\xi, \mathcal{F}^c, \|\cdot\|_\infty) \le N(\xi, \mathcal{F}, \|\cdot\|_\infty)$. Starting from a rescaling step as before, we obtain

$$\mathbb{E}\Big[\sup_{\psi \in \bigcup_{\mathbf{A}} \mathcal{F}^c_{\mathbf{A}}} \big|(\nu - \hat{\nu}_n)\psi\big|\Big] \lesssim_{d_x, d_y} (1 + R^4) \mathbb{E}\Big[\sup_{\psi \in \bigcup_{\mathbf{A}} \mathcal{F}^c_{\mathbf{A}}} \big|(s_\sharp \nu - s_\sharp \hat{\nu}_n)(S\psi)\big|\Big]. \tag{25}$$

Using the LCA principle, we have the following bound on the covering number of the union of rescaled c-transformed classes, proven in Section F.2 of the Supplementary Material [70].

LEMMA 5.5. *For any $\xi > 0$, we have the covering bound*

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

Armed with the lemma, we proceed from (25) and, for $d_x \ge 4$, obtain

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

As before, when $d_x < 4$, a parametric rate bound holds instead. Inserting the above along with (23) into (24) 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 $\mathbb{E}[|\mathsf{D}(\mu, \nu) - \mathsf{D}(\hat{\mu}_n, \hat{\nu}_n)|] \le \mathsf{D}(\mu, \nu)^{-1} \mathbb{E}[|\mathsf{D}(\mu, \nu)^2 - \mathsf{D}(\hat{\mu}_n, \hat{\nu}_n)^2|]$, 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.

5.3.2. *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 $4 < d_x \le d_y$ and $R = 4$.[9] 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 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 $\mu = \mathrm{Unif}(B_d(0, 1))$ and $\nu = \mathrm{Unif}(B_d(0, 2))$.

We start from the one-sample case and establish $\mathbb{E}[|\mathsf{D}(\mu, \nu)^2 - \mathsf{D}(\hat{\mu}_n, \nu)^2|] \gtrsim n^{-2/d}$. Theorem 9.21 of [61] implies that $T : x \mapsto 2x$ is an optimal Gromov–Monge map from μ and ν,

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

and thus $\mathsf{D}(\mu,\nu)^2 = \int_{\mathcal{X}\times\mathcal{X}} |\|x-x'\|^2 - \|2x-2x'\|^2|^2\, d\mu\otimes\mu(x,x')$. 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

$$\begin{aligned}\mathsf{D}(\hat{\mu}_n,\nu)^2 &= \int |\|y-y'\|^2 - \|z-z'\|^2|^2\, d\pi_n\otimes\pi_n(y,z,y',z') \\ &= \int |\|y-y'\|^2 - \|2x-2x'\|^2|^2\, d\pi_n'\otimes\pi_n'(y,x,y',x') \qquad (26) \\ &= 4\mathsf{D}(\hat{\mu}_n,\mu)^2 - 3\int \|y-y'\|^4\, d\hat{\mu}_n\otimes\hat{\mu}_n(y,y') + 12\int \|x-x'\|^4\, d\mu\otimes\mu(x,x').\end{aligned}$$

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

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

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

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

For the first term, let $\tilde{\mu}_n$ be the centered version of $\hat{\mu}_n$ and invoke Lemma 4.4 to obtain

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

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 $\mathbb{E}[\mathsf{W}_1(\hat{\mu}_n,\mu)] \gtrsim n^{-1/d}$ [18], 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 (26), 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

$$\begin{aligned}\mathsf{D}(\hat{\mu}_n,\hat{\nu}_n)^2 - \mathsf{D}(\mu,\nu)^2 &= 4\mathsf{D}(\hat{\mu}_n,\hat{\mu}_n')^2 - 3\int \|y-y'\|^4\, d\hat{\mu}_n\otimes\hat{\mu}_n(y,y') \\ &\quad + 12\int \|y-y'\|^4\, d\hat{\mu}_n'\otimes\hat{\mu}_n'(y,y') - 9\int \|y-y'\|^4\, d\mu\otimes\mu(y,y'). \qquad (27)\end{aligned}$$

As before, we have $\mathbb{E}[\int \|y-y'\|^4\, d\hat{\mu}_n\otimes\hat{\mu}_n(y,y')] = \frac{n-1}{n}\int \|y-y'\|^4\, d\mu\otimes\mu(y,y')$ and similarly for $\mathbb{E}[\int \|y-y'\|^4\, d\hat{\mu}_n'\otimes\hat{\mu}_n'(y,y')]$, 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 Section F.3 of the Supplementary Material [70].

LEMMA 5.6 (Intermediate lower bound). *The following bound holds*:

$$\mathbb{E}\big[\mathsf{D}(\hat{\mu}_n, \hat{\mu}'_n)^2\big] \gtrsim \mathbb{E}\Big[\lambda_{\min}(\mathbf{\Sigma}_{\hat{\mu}_n})\mathbb{E}\Big[\inf_{\mathbf{U}\in O(d)} \mathsf{W}_1(\hat{\mu}_n, \mathbf{U}_\sharp \hat{\mu}'_n)^2 | X_1, \dots, X_n\Big]\Big] - 2\sqrt{\frac{M_2(\mu)}{n}}. \tag{28}$$

To treat the inner (conditional) expectation on the RHS of (28), we make use of the next lemma; see Section F.4 of the Supplementary Material [70] for the proof.

LEMMA 5.7. *For any* $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$ *with* $\mathrm{spt}(\mu), \mathrm{spt}(\nu) \subset B_d(0, 1)$, *we have*

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

where C_d *depends only on the dimension* d.

Applying the lemma, we obtain

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

Note that for any $\mathbf{U} \in O(d)$, we have $\mathbb{E}[\mathsf{W}_1(\hat{\mu}_n, \mathbf{U}_\sharp \hat{\mu}'_n)|X_1, \dots, X_n] \geq \mathsf{W}_1(\mu, \mathbf{U}_\sharp \hat{\mu}'_n) = \mathsf{W}_1(\mu, \hat{\mu}'_n)$, 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 (28), yields

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

To lower bound the expectation on the RHS, recall that by Proposition 2.1 in [18] (see also [66], 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

$$\mathbb{E}\big[\mathsf{D}(\hat{\mu}_n, \hat{\mu}'_n)^2\big] \gtrsim_d \mathbb{E}\big[\lambda_{\min}(\mathbf{\Sigma}_{\hat{\mu}_n})\big] \cdot n^{-2/d}, \tag{29}$$

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

Write $\mathbb{E}[\lambda_{\min}(\mathbf{\Sigma}_{\hat{\mu}_n})] = \mathbb{E}[\inf_{\|v\|=1} \hat{\mu}_n |v \cdot x|^2]$. 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 $\mathbb{E}[\sup_{\|v\|=1} |(\hat{\mu}_n - \mu)|v \cdot x|^2|] \leq \lambda_{\min}(\mathbf{\Sigma}_\mu)/2$. Consequently

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

Inserting this back into (29) and recalling the decomposition of the empirical estimation error from (27) concludes the proof of the two-sample lower bound. □

REMARK 5.8 (W-procrustes empirical convergence). Our two-sample analysis 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 $p \geq 1$.

6. Outlook and concluding remarks. This paper established a dual formulation for both the standard $(2, 2)$-GW distance and its entropically regularized version, between distributions supported on Euclidean spaces of different dimensions d_x and d_y. The dual forms represented GW and EGW as infima of a class of OT and EOT problems, respectively, 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 and EGW. The derived two-sample rates are $n^{-2/((d_x \wedge d_y) \vee 4)}$ (up to a log factor when $d_x \wedge d_y = 4$) for GW and $n^{-1/2}$ for EGW. The GW result accounts for compactly supported distributions, and provides matching upper and lower rate bound. For EGW, our analysis allows for unbounded domains subject to a 4-sub-Weibull condition. These results are in line with the empirical convergence rates of OT [37, 42] and EOT [33, 46].

We have also explored stability and continuity of the EGW problem in the entropic regularization parameter ε. We provided an $O(\varepsilon \log(1/\varepsilon))$ approximation bound on the GW cost and a continuity result for the optimal couplings in the weak topology. Lastly, we reexamined the open problem of the one-dimensional GW distance between discrete distributions on n points. Leveraging our duality theory, 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 represents the GW distance as a sum of concave and convex functions, illuminating that, in some regimes, the optimum is not 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 (14), in Section G in the Supplementary Material [70] we show that

$$
\begin{aligned}
&\mathsf{D}_{2,2k}(\mu, \nu)^2 \\
&\quad = 4 \sup_{a \in \mathbb{R}^\ell} \inf_{b \in \mathbb{R}^{m-\ell}} \Bigg\{ -\|a\|^2 + \|b\|^2 \\
&\qquad + \inf_{\pi \in \Pi(\mu,\nu)} \int \Bigg(-\|x\|^{2k}\|y\|^{2k} + \sum_{i=1}^{\ell} a_i g_i(x, y) - \sum_{i=\ell+1}^{m} b_{i-\ell} g_i(x, y) \Bigg) d\pi(x, y) \Bigg\},
\end{aligned} \tag{30}
$$

where $g_1, \dots, g_m$ 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 Section G in the Supplementary Material [70] 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 Corollary 4.1 once we set

$k = 1$ and assume that μ, ν are centered. A similar representation holds for the $(2, 2k)$-EGW variant, but with the entropic penalty $\varepsilon \mathsf{D}_{\mathsf{KL}}(\pi \| \mu \otimes \nu)$ added to the transportation cost in the second line above.

Notice now that the inner optimization over π specifies 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 concave in x or y. For the standard $(2, 2k)$-GW distance between compactly supported distributions, an argument similar to the proof of Theorem 4.2, would result in a two-sample convergence rate of $O(n^{-1/(d_x \wedge d_y)})$. 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. For the EGW case, under proper tail conditions (say, $4k$-sub-Weibull), smoothness of the cost would allow to reproduce the current derivation of Theorem 3.5 and arrive at the parametric convergence rate. In sum, while a duality theory for general (p, q) remains an open question, our results for the quadratic GW and EGW distances can be extended to cover any even q value.

As mentioned above, extending our duality to 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 nontrivial and may require assumptions along the lines of Theorem 13 from [42], where empirical convergence of OT on unbounded domains was treated.

Funding. Z. Goldfeld is partially supported by NSF grants CAREER CCF-2046018, DMS-2210368 and CCF-2308446, and the IBM Academic Award.

B. K. Sriperumbudur is partially supported by the NSF CAREER award DMS-1945396.

SUPPLEMENTARY MATERIAL

Supplement to: "Gromov–Wasserstein distances: Entropic regularization, duality and sample complexity". (DOI: 10.1214/24-AOS2406SUPP; .pdf). Due to space constraints, proofs of some technical lemmas and propositions from the main text are provided in the supplement.

REFERENCES

[1] Alvarez-Melis, D. and Jaakkola, T. (2018). Gromov–Wasserstein alignment of word embedding spaces. In *Proceedings of the* 2018 *Conference on Empirical Methods in Natural Language Processing* (E. Riloff, D. Chiang, J. Hockenmaier and J. Tsujii, eds.) 1881–1890. Association for Computational Linguistics, Brussels, Belgium. https://doi.org/10.18653/v1/D18-1214

[2] Beinert, R., Heiss, C. and Steidl, G. (2023). On assignment problems related to Gromov–Wasserstein distances on the real line. *SIAM J. Imaging Sci.* **16** 1028–1032. MR4605212 https://doi.org/10.1137/22M1497808

[3] Benamou, J.-D. and Brenier, Y. (2000). A computational fluid mechanics solution to the Monge–Kantorovich mass transfer problem. *Numer. Math.* **84** 375–393. MR1738163 https://doi.org/10.1007/s002110050002

[4] Bigot, J., Cazelles, E. and Papadakis, N. (2019). Central limit theorems for entropy-regularized optimal transport on finite spaces and statistical applications. *Electron. J. Stat.* **13** 5120–5150. MR4041704 https://doi.org/10.1214/19-EJS1637

[5] Blumberg, A. J., Carriere, M., Mandell, M. A., Rabadan, R. and Villar, S. (2020). MREC: a fast and versatile framework for aligning and matching point clouds with applications to single cell molecular data. arXiv preprint. Available at arXiv:2001.01666.

[6] Boissard, E. and Le Gouic, T. (2014). On the mean speed of convergence of empirical and occupation measures in Wasserstein distance. *Ann. Inst. Henri Poincaré Probab. Stat.* **50** 539–563. MR3189084 https://doi.org/10.1214/12-AIHP517

[7] BUNNE, C., ALVAREZ-MELIS, D., KRAUSE, A. and JEGELKA, S. (2019). Learning generative models across incomparable spaces. In *Proceedings of the* 36*th International Conference on Machine Learning* (K. Chaudhuri and R. Salakhutdinov, eds.). *Proceedings of Machine Learning Research* **97** 851–861. PMLR, Long Beach, CA, USA.

[8] CARLIER, G., DUVAL, V., PEYRÉ, G. and SCHMITZER, B. (2017). Convergence of entropic schemes for optimal transport and gradient flows. *SIAM J. Math. Anal.* **49** 1385–1418. MR3635459 https://doi.org/10.1137/15M1050264

[9] CHEN, Y., GEORGIOU, T. T. and PAVON, M. (2016). On the relation between optimal transport and Schrödinger bridges: A stochastic control viewpoint. *J. Optim. Theory Appl.* **169** 671–691. MR3489825 https://doi.org/10.1007/s10957-015-0803-z

[10] CHIZAT, L., ROUSSILLON, P., LÉGER, F., VIALARD, F.-X. and PEYRÉ, G. (2020). Faster Wasserstein distance estimation with the Sinkhorn divergence. In *Advances in Neural Information Processing Systems* (H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan and H. Lin, eds.) **33** 2257–2269. Curran Associates, Red Hook.

[11] COMMANDER, C. W. (2005). A survey of the quadratic assignment problem, with applications. *Morehead Electron. J. of Appl. Math.* **4**. MATH-2005-01.

[12] CONFORTI, G. and TAMANINI, L. (2021). A formula for the time derivative of the entropic cost and applications. *J. Funct. Anal.* **280** Paper No. 108964, 48. MR4232667 https://doi.org/10.1016/j.jfa.2021.108964

[13] DEL BARRIO, E., GONZÁLEZ SANZ, A., LOUBES, J.-M. and NILES-WEED, J. (2023). An improved central limit theorem and fast convergence rates for entropic transportation costs. *SIAM J. Math. Data Sci.* **5** 639–669. MR4616887 https://doi.org/10.1137/22M149260X

[14] DEL BARRIO, E. and LOUBES, J.-M. (2019). Central limit theorems for empirical transportation cost in general dimension. *Ann. Probab.* **47** 926–951. MR3916938 https://doi.org/10.1214/18-AOP1275

[15] DELON, J., DESOLNEUX, A. and SALMONA, A. (2022). Gromov–Wasserstein distances between Gaussian distributions. *J. Appl. Probab.* **59** 1178–1198. MR4507687 https://doi.org/10.1017/jpr.2022.16

[16] DEMETCI, P., SANTORELLA, R., SANDSTEDE, B., NOBLE, W. S. and SINGH, R. (2022). SCOT: Single-cell multi-omics alignment with optimal transport. *J. Comput. Biol.* **29** 3–18. MR4370741 https://doi.org/10.1089/cmb.2021.0446

[17] DEREICH, S., SCHEUTZOW, M. and SCHOTTSTEDT, R. (2013). Constructive quantization: Approximation by empirical measures. *Ann. Inst. Henri Poincaré Probab. Stat.* **49** 1183–1203. MR3127919 https://doi.org/10.1214/12-AIHP489

[18] DUDLEY, R. M. (1968). The speed of mean Glivenko–Cantelli convergence. *Ann. Math. Stat.* **40** 40–50. MR0236977 https://doi.org/10.1214/aoms/1177697802

[19] DUMONT, T., LACOMBE, T. and VIALARD, F.-X. (2024). On the existence of Monge maps for the Gromov–Wasserstein problem. *Found. Comput. Math.* 1–48.

[20] FOURNIER, N. and GUILLIN, A. (2015). On the rate of convergence in Wasserstein distance of the empirical measure. *Probab. Theory Related Fields* **162** 707–738. MR3383341 https://doi.org/10.1007/s00440-014-0583-7

[21] GENEVAY, A., CHIZAT, L., BACH, F., CUTURI, M. and PEYRÉ, G. (2019). Sample complexity of Sinkhorn divergences. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics* (K. Chaudhuri and M. Sugiyama, eds.). *Proceedings of Machine Learning Research* **89** 1574–1583. PMLR, Naha, Okinawa, Japan.

[22] GENTIL, I., LÉONARD, C. and RIPANI, L. (2017). About the analogy between optimal transport and minimal entropy. *Ann. Fac. Sci. Toulouse Math.* (6) **26** 569–601. MR3669966 https://doi.org/10.5802/afst.1546

[23] GIGLI, N. and TAMANINI, L. (2020). Benamou–Brenier and duality formulas for the entropic cost on $\mathsf{RCD}^*(K,N)$ spaces. *Probab. Theory Related Fields* **176** 1–34. MR4055185 https://doi.org/10.1007/s00440-019-00909-1

[24] GINÉ, E. and NICKL, R. (2016). *Mathematical Foundations of Infinite-Dimensional Statistical Models. Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge Univ. Press, New York. MR3588285 https://doi.org/10.1017/CBO9781107337862

[25] GOLDFELD, Z. and GREENEWALD, K. (2020). Gaussian-smoothed optimal transport: Metric structure and statistical efficiency. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics* (S. Chiappa and R. Calandra, eds.). *Proceedings of Machine Learning Research* **108** 3327–3337. PMLR, Online.

[26] GOLDFELD, Z., KATO, K., NIETERT, S. and RIOUX, G. (2024). Limit distribution theory for smooth p-Wasserstein distances. *Ann. Appl. Probab.* **34** 2447–2487. MR4728174 https://doi.org/10.1214/23-aap2028

[27] GOLDFELD, Z., KATO, K., RIOUX, G. and SADHU, R. (2024). Statistical inference with regularized optimal transport. *Inf. Inference* **13** Paper No. 13, 68. MR4701828 https://doi.org/10.1093/imaiai/iaad056

[28] GOLDFELD, Z., KATO, K., RIOUX, G. and SADHU, R. (2024). Limit theorems for entropic optimal transport maps and Sinkhorn divergence. *Electron. J. Stat.* **18** 980–1041. MR4718466 https://doi.org/10.1214/24-ejs2217

[29] GONZÁLEZ-SANZ, A. and HUNDRIESER, S. (2023). Weak limits for empirical entropic optimal transport: Beyond smooth costs. arXiv preprint. Available at arXiv:2305.09745.

[30] GONZÁLEZ-SANZ, A., LOUBES, J.-M. and NILES-WEED, J. (2022). Weak limits of entropy regularized optimal transport; potentials, plans and divergences. arXiv preprint. Available at arXiv:2207.07427.

[31] GOODALL, C. (1991). Procrustes methods in the statistical analysis of shape. *J. Roy. Statist. Soc. Ser. B* **53** 285–339. MR1108330

[32] GRAVE, E., JOULIN, A. and BERTHET, Q. (2019). Unsupervised alignment of embeddings with Wasserstein procrustes. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics* (K. Chaudhuri and M. Sugiyama, eds.). *Proceedings of Machine Learning Research* **89** 1880–1890. PMLR, Naha, Okinawa, Japan.

[33] GROPPE, M. and HUNDRIESER, S. (2023). Lower complexity adaptation for empirical entropic optimal transport. arXiv preprint. Available at arXiv:2306.13580.

[34] GUNSILIUS, F. and XU, Y. (2021). Matching for causal effects via multimarginal unbalanced optimal transport. arXiv preprint, updated on July 9, 2022. Available at arXiv:2112.04398.

[35] GUNTUBOYINA, A. and SEN, B. (2013). Covering numbers for convex functions. *IEEE Trans. Inf. Theory* **59** 1957–1965. MR3043776 https://doi.org/10.1109/TIT.2012.2235172

[36] HUNDRIESER, S., KLATT, M., STAUDT, T. and MUNK, A. (2022). A unifying approach to distributional limits for empirical optimal transport. arXiv preprint. Available at arXiv:2202.12790.

[37] HUNDRIESER, S., STAUDT, T. and MUNK, A. (2024). Empirical optimal transport between different measures adapts to lower complexity. *Ann. Inst. Henri Poincaré Probab. Stat.* **60** 824–846. MR4757509 https://doi.org/10.1214/23-aihp1369

[38] KLATT, M., TAMELING, C. and MUNK, A. (2020). Empirical regularized optimal transport: Statistical theory and applications. *SIAM J. Math. Data Sci.* **2** 419–443. MR4105566 https://doi.org/10.1137/19M1278788

[39] KOEHL, P., DELARUE, M. and ORLAND, H. (2023). Computing the Gromov–Wasserstein distance between two surface meshes using optimal transport. *Algorithms* **16** 131.

[40] LE, K., LE, D. Q., NGUYEN, H., DO, D., PHAM, T. and HO, N. (2022). Entropic Gromov–Wasserstein between Gaussian distributions. In *Proceedings of the* 39*th International Conference on Machine Learning* (K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu and S. Sabato, eds.). *Proceedings of Machine Learning Research* **162** 12164–12203. PMLR, Baltimore, MD, USA.

[41] MANOLE, T., BALAKRISHNAN, S., NILES-WEED, J. and WASSERMAN, L. (2021). Plugin estimation of smooth optimal transport maps. arXiv preprint. Available at arXiv:2107.12364.

[42] MANOLE, T. and NILES-WEED, J. (2024). Sharp convergence rates for empirical optimal transport with smooth costs. *Ann. Appl. Probab.* **34** 1108–1135. MR4700254 https://doi.org/10.1214/23-aap1986

[43] MARON, H. and LIPMAN, Y. (2018). (Probably) concave graph matching. In *Advances in Neural Information Processing Systems* (S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi and R. Garnett, eds.). **31**. Curran Associates, Red Hook.

[44] MÉMOLI, F. (2009). Spectral Gromov–Wasserstein distances for shape matching. In 2009 *IEEE* 12*th International Conference on Computer Vision Workshops*. *ICCV Workshops* 256–263. IEEE Press, New York.

[45] MÉMOLI, F. (2011). Gromov–Wasserstein distances and the metric approach to object matching. *Found. Comput. Math.* **11** 417–487. MR2811584 https://doi.org/10.1007/s10208-011-9093-5

[46] MENA, G. and NILES-WEED, J. (2019). Statistical bounds for entropic optimal transport: Sample complexity and the central limit theorem. In *Advances in Neural Information Processing Systems* (H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox and R. Garnett, eds.). **32**. Curran Associates, Red Hook.

[47] NUTZ, M. and WIESEL, J. (2022). Entropic optimal transport: Convergence of potentials. *Probab. Theory Related Fields* **184** 401–424. MR4498514 https://doi.org/10.1007/s00440-021-01096-8

[48] PEYRÉ, G., CUTURI, M. et al. (2019). Computational optimal transport: With applications to data science. Foundations and trends® in machine learning, Vol. 11. 355–607.

[49] PEYRÉ, G., CUTURI, M. and SOLOMON, J. (2016). Gromov–Wasserstein averaging of kernel and distance matrices. In *Proceedings of the* 33*rd International Conference on Machine Learning* (M. F. Balcan and K. Q. Weinberger, eds.). *Proceedings of Machine Learning Research* **48** 2664–2672. PMLR, New York.

[50] RIGOLLET, P. and STROMME, A. J. (2022). On the sample complexity of entropic optimal transport. arXiv preprint. Available at arXiv:2206.13472.

[51] RIOUX, G., GOLDFELD, Z. and KATO, K. (2023). Entropic Gromov–Wasserstein distances: Stability, algorithms, and distributional limits. arXiv preprint. Available at arXiv:2306.00182.

[52] RÖMISCH, W. (2004). *Delta Method, Infinite Dimensional. Encyclopedia of Statistical Sciences*. Wiley, New York.

[53] SANTAMBROGIO, F. (2015). *Optimal Transport for Applied Mathematicians: Calculus of Variations, PDEs, and Modeling. Progress in Nonlinear Differential Equations and Their Applications* **87**. Birkhäuser, Cham. MR3409718 https://doi.org/10.1007/978-3-319-20828-2

[54] SCETBON, M., PEYRÉ, G. and CUTURI, M. (2022). Linear-time Gromov Wasserstein distances using low rank couplings and costs. In *Proceedings of the* 39*th International Conference on Machine Learning* (K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu and S. Sabato, eds.). *Proceedings of Machine Learning Research* **162** 19347–19365. PMLR, Baltimore, MD, USA.

[55] SCHÖNEMANN, P. H. (1966). A generalized solution of the orthogonal Procrustes problem. *Psychometrika* **31** 1–10. MR0215870 https://doi.org/10.1007/BF02289451

[56] SEJOURNE, T., VIALARD, F.-X. and PEYRÉ, G. (2021). The unbalanced Gromov–Wasserstein distance: Conic formulation and relaxation. In *Advances in Neural Information Processing Systems* (M. Ranzato, A. Beygelzimer, Y. Dauphin, P. S. Liang and J. W. Vaughan, eds.) **34** 8766–8779. Curran Associates, Red Hook.

[57] SHAPIRO, A. (1990). On concepts of directional differentiability. *J. Optim. Theory Appl.* **66** 477–487. MR1080259 https://doi.org/10.1007/BF00940933

[58] SOLOMON, J., PEYRÉ, G., KIM, V. G. and SRA, S. (2016). Entropic metric alignment for correspondence problems. *ACM Trans. Graph.* **35** 1–13.

[59] SOMMERFELD, M. and MUNK, A. (2018). Inference for empirical Wasserstein distances on finite spaces. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **80** 219–238. MR3744719 https://doi.org/10.1111/rssb.12236

[60] STROMME, A. J. (2023). Minimum intrinsic dimension scaling for entropic optimal transport. arXiv preprint. Available at arXiv:2306.03398.

[61] STURM, K.-T. (2023). The space of spaces: Curvature bounds and gradient flows on the space of metric measure spaces. *Mem. Amer. Math. Soc.* **290** v+111. MR4659681 https://doi.org/10.1090/memo/1443

[62] TAMELING, C., SOMMERFELD, M. and MUNK, A. (2019). Empirical optimal transport on countable metric spaces: Distributional limits and statistical applications. *Ann. Appl. Probab.* **29** 2744–2781. MR4019874 https://doi.org/10.1214/19-AAP1463

[63] TITOUAN, V., FLAMARY, R., COURTY, N., TAVENARD, R. and CHAPEL, L. (2019). Sliced Gromov–Wasserstein. In *Advances in Neural Information Processing Systems* (H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox and R. Garnett, eds.). **32**. Curran Associates, Red Hook.

[64] VAN DER VAART, A. W. and WELLNER, J. A. (2023). *Weak Convergence and Empirical Processes—with Applications to Statistics. Springer Series in Statistics*. Springer, Cham. Second edition [of 1385671]. MR4628026 https://doi.org/10.1007/978-3-031-29040-4

[65] VILLANI, C. (2009). *Optimal Transport: Old and New. Grundlehren der Mathematischen Wissenschaften* [*Fundamental Principles of Mathematical Sciences*] **338**. Springer, Berlin. MR2459454 https://doi.org/10.1007/978-3-540-71050-9

[66] WEED, J. and BACH, F. (2019). Sharp asymptotic and finite-sample rates of convergence of empirical measures in Wasserstein distance. *Bernoulli* **25** 2620–2648. MR4003560 https://doi.org/10.3150/18-BEJ1065

[67] XU, H., LUO, D. and CARIN, L. (2019). Scalable Gromov–Wasserstein learning for graph partitioning and matching. In *Advances in Neural Information Processing Systems* (H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox and R. Garnett, eds.). **32**. Curran Associates, Red Hook.

[68] XU, H., LUO, D., ZHA, H. and DUKE, L. C. (2019). Gromov–Wasserstein learning for graph matching and node embedding. In *Proceedings of the* 36*th International Conference on Machine Learning* (K. Chaudhuri and R. Salakhutdinov, eds.). *Proceedings of Machine Learning Research* **97** 6932–6941. PMLR, Long Beach, CA, USA.

[69] YAN, Y., LI, W., WU, H., MIN, H., TAN, M. and WU, Q. (2018). Semi-supervised optimal transport for heterogeneous domain adaptation. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI*-18 2969–2975. International Joint Conferences on Artificial Intelligence Organization. https://doi.org/10.24963/ijcai.2018/412

[70] ZHANG, Z., GOLDFELD, Z., MROUEH, Y. and SRIPERUMBUDUR, B. K. (2024). Supplement to "Gromov–Wasserstein Distances: Entropic Regularization, Duality and Sample Complexity." https://doi.org/10.1214/24-AOS2406SUPP

[71] ZHANG, Z., MROUEH, Y., GOLDFELD, Z. and SRIPERUMBUDUR, B. (2022). Cycle consistent probability divergences across different spaces. In *Proceedings of the* 25*th International Conference on Artificial Intelligence and Statistics* (G. Camps-Valls, F. J. R. Ruiz and I. Valera, eds.). *Proceedings of Machine Learning Research* **151** 7257–7285. PMLR, Online.