Sliced Inner Product Gromov–Wasserstein Distances

Xiaoyun Gong[*]
Cornell University
Ithaca, NY 14853
xg332@cornell.edu

Gabriel Rioux[*]
Imperial College London
London, SW7 2AZ
g.rioux@ic.ac.uk

Ziv Goldfeld
Cornell University
Ithaca, NY 14853
goldfeld@cornell.edu

Abstract

The Gromov–Wasserstein (GW) problem provides a framework for aligning heterogeneous datasets by matching their intrinsic geometry, but its statistical and computational scaling remains an issue for high-dimensional problems. Slicing techniques offer an appealing route to scalability, but, unlike Wasserstein distances, GW problems do not generally admit closed-form solutions in one-dimension. We resolve this problem for the GW problem with inner product cost (IGW), propose a sliced IGW distance that enjoys a natural rotational invariance property, and comprehensively study its structural and computational properties. Numerical experiments validating our theory are presented, followed by applications to heterogeneous clustering of text data and language model representation comparison.

1 Introduction

The Gromov–Wasserstein (GW) problem enables aligning heterogeneous probability spaces by matching their intrinsic geometry, without the need for shared coordinates or pointwise labels [1]. This versatility has made GW central to structure-aware matching and inference, including network/graph comparison [2], latent correspondence learning [3, 4], and graph-isomorphism testing [5, 6], among others. However, scaling GW alignment to modern regimes remains challenging as empirical rates exhibit the curse of dimensionality [7], while computation generally involves solving nonconvex quadratic problems [1, 2] which are NP hard to solve in general.

An effective route to obtaining scalable statistical distances is via *slicing*, i.e., aggregating distance values across one-dimensional projections. For Wasserstein distances, this strategy succeeds as each 1D subproblem admits a simple closed-form solution [8]. This tractability underlies broad applications in computing barycenters [9], generative modeling [10–13], and representation learning [14–16], along with a growing statistical and algorithmic theory [17–22]. More recent work has explored constructing original transport from sliced transport [23, 24], identifying informative slicing directions [16, 25–27], as well as leveraging sliced Wasserstein distances as part of the objective in other algorithms/tasks [28–31]. By contrast, GW generally lacks a closed-form solution in one dimension, limiting the development of a parallel slicing framework, see, e.g., [32].

[*]Denotes equal contribution to this work.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

Contributions We consider the GW problem with inner product cost (abbreviated IGW), between probability distributions $\mu \in \mathcal{P}(\mathbb{R}^{d_x})$ and $\nu \in \mathcal{P}(\mathbb{R}^{d_y})$, supported on arbitrary Euclidean spaces:

$$\mathsf{IGW}(\mu,\nu) \coloneqq \left(\inf_{\pi \in \Pi(\mu,\nu)} \iint |\langle x, x' \rangle - \langle y, y' \rangle|^2 \, d\pi \otimes \pi(x, y, x', y') \right)^{1/2}, \tag{1}$$

where $\Pi(\mu,\nu)$ is the set of all couplings of μ and ν. IGW quantifies the least amount of inner product distortion achievable when optimizing over all possible correspondences of the two spaces. As such, it captures a meaningful notion of discrepancy between heterogeneous spaces with a natural inner product structure, which motivated its adoption in practice. When $d_x = d_y = 1$, we first establish a closed-form solution for this problem which reduces to sorting if μ and ν are uniformly distributed on the same (finite) number of points, akin to the Wasserstein case. This unlocks a slicing paradigm for IGW which is the main focus of this work.

The natural construction of the average-sliced IGW (namely, averaging IGW values across one-dimensional projections), is not necessarily invariant to orthogonal transformations, which is a key property of IGW. To address this, inspired by [33], we define a rotationally invariant sliced IGW, denoted $\underline{\mathsf{IGW}}$ (see Definition 4.1), by introducing an additional outer minimization over orthogonal transformations of one of the marginals. Crucially, the slicing enables explicit formulas for the inner IGW and the minimization restores invariance to orthogonal transformations. With this, $\underline{\mathsf{IGW}}$ defines a pseudometric on the space of probability measures on $\mathbb{R}^d$ and is topologically equivalent to IGW under certain structural conditions. As $\underline{\mathsf{IGW}}$ involves an optimization over the Stiefel manifold of an objective defined in terms of an integral over the unit sphere, we study Monte Carlo averaging as a proxy for integrating over the sphere and propose a subgradient method for optimizing over the Stiefel manifold. Non-asymptotic guarantees are provided for both approximations, resulting in a principled computation framework for average-sliced IGW. The paper concludes with numerical experiments validating the derived theory and providing applications of this framework to comparing embeddings of LLMs and clustering of heterogeneous datasets.

Related Literature The IGW distance, introduced in [34], has been applied for various machine learning tasks, including latent correspondence learning [3], LLM representation comparison and heterogeneous clustering of text data [35], cross-lingual alignment [36], among others. Its popularity stems, in part, from its nice analytic properties. Indeed, in contrast to the standard GW distance with quadratic cost, closed-form solutions for IGW with Gaussian marginals are known [3, 35, 37] and the existence of solutions induced by maps holds under mild conditions [38].

The computational burden associated with solving GW problems (in particular, IGW) motivated the study of computationally efficient approximate solvers. This includes approaches based on entropic regularization [4, 39–43], convex/conic relaxations [44, 45], neural approximation [46], semi-definite relaxations [47], and direct optimization over bi-directional (parametrized) Gromov–Monge maps [48]. [33] proposed a sliced quadratic GW distance and studied its metric and topological properties. However, this approach is contingent on a closed-form solution for quadratic GW (i.e., with squared norms in place of inner products); the formula provided in [33] was later shown to be invalid via a counterexample [32]. Another line of work applies slicing to a lower bound of a variant of GW [49].

The motivation to unlock slicing for GW stems from the practical success of sliced Wasserstein distances [9], which overcome the statistical and computational issues of standard Wasserstein distance [20]. Average-slicing is typically implemented via Monte Carlo integration [20, 50], for which quantitative error bounds were provided in [22]. From the statistical standpoint, sliced Wasserstein distances enjoy fast dimension-independent convergence rates [19, 20, 22, 51, 52], as well as a comprehensive limit distribution theory [19, 53–55]. Given their attractive algorithmic and statistical profile, sliced Wasserstein distances have been broadly applied for machine learning tasks such as generative modeling [10–13], barycenter computation [9], autoencoders [14, 15], locality-sensitive hashing [29], and representation learning [16].

2 Background and Preliminaries

We briefly review definitions and preliminary results concerning the IGW problem. Let $\|\cdot\|$ and $\langle \cdot, \cdot \rangle$ denote the Euclidean norm and inner product. The operator and Frobenius norms of a matrix

$\mathbf{A} \in \mathbb{R}^{d\times k}$ are denoted by $\|\mathbf{A}\|_{\mathrm{op}}$ and $\|\mathbf{A}\|_{\mathrm{F}}$, respectively. We write $\mathbf{I}_d \in \mathbb{R}^{d\times d}$ for the identity matrix. For $d_x \le d_y$, $\mathrm{St}(d_x, d_y) \coloneqq \left\{\boldsymbol{\Delta} \in \mathbb{R}^{d_y\times d_x} : \boldsymbol{\Delta}^\intercal\boldsymbol{\Delta} = \mathbf{I}_{d_x}\right\}$ is the Stiefel manifold of orthonormal d_x-frames in $\mathbb{R}^{d_y}$. Write $\mathcal{P}(\mathbb{R}^d)$ for the space of Borel probability measures over $\mathbb{R}^d$, and let $\mathcal{P}_p(\mathbb{R}^d)$ be its restriction to distributions with finite p-th absolute moments. For $\mu \in \mathcal{P}(\mathbb{R}^d)$, we denote its support by $\mathrm{spt}(\mu)$, while $\boldsymbol{\Sigma}_\mu$, $\mathbf{R}_\mu$ and $M_p(\mu)$ denote the covariance matrix, second moment matrix, and p-th absolute moment of μ, respectively. We write $T_\sharp\mu$ for the pushforward measure of μ through a measurable map T. For $\theta \in \mathbb{R}^d$, we identify $\theta^\intercal$ with function $x \mapsto \theta^\intercal x$, and write $(\theta^\intercal)_\sharp\mu$ for the corresponding pushforward. Weak convergence of probability measures is denoted as $\nu_n \xrightarrow{w} \nu$. We use $\lesssim_x$ to denote inequalities up to constants that only depend on x; the subscript is dropped when the constant is universal.

Inner Product Gromov–Wasserstein Distance Throughout, we analyze the GW distance with inner product cost (IGW) between $\mu \in \mathcal{P}(\mathbb{R}^{d_x})$ and $\nu \in \mathcal{P}(\mathbb{R}^{d_y})$ as defined in (1). Notably, IGW is invariant under orthogonal transformations (but is not translation-invariant [3]) and defines a pseudometric in the sense that it is symmetric, satisfies the triangle inequality, and is nonnegative with $\mathsf{IGW}(\mu,\nu) = 0$ if and only if there exists a $\mu\otimes\mu$-a.s. unitary isomorphism $T : \mathrm{spt}(\mu) \to \mathrm{spt}(\nu)$ with $T_\sharp\mu = \nu$ [56, Proposition 3.1]. It is known that the IGW distance can be connected to a class of parametrized optimal transport (OT) problems as follows; see [4, Appendix E] and [56, Lemma 2.1].

Lemma 2.1 (IGW duality)**.** *Fix $(\mu,\nu) \in \mathcal{P}_2(\mathbb{R}^{d_x}) \times \mathcal{P}_2(\mathbb{R}^{d_y})$, and $M_{\mu,\nu} \coloneqq \sqrt{M_2(\mu)M_2(\nu)}$. Then,*

$$\mathsf{IGW}^2(\mu,\nu) = \|\mathbf{R}_\mu\|_{\mathrm{F}}^2 + \|\mathbf{R}_\nu\|_{\mathrm{F}}^2 + \inf_{\mathbf{A}\in\mathbb{R}^{d_x\times d_y}} 8\|\mathbf{A}\|_{\mathrm{F}}^2 + \mathsf{IOT}_{\mathbf{A}}(\mu,\nu), \tag{2}$$

where $\mathsf{IOT}_{\mathbf{A}}(\mu,\nu) \coloneqq \inf_{\pi\in\Pi(\mu,\nu)}\int c_{\mathbf{A}}\,d\pi$ is the OT problem with cost function $c_{\mathbf{A}} : (x,y) \in \mathbb{R}^{d_x} \times \mathbb{R}^{d_y} \mapsto -8x^\intercal\mathbf{A}y$ and the infimum is achieved at some $\mathbf{A}^\star \in \mathcal{D}_{M_{\mu,\nu}} \coloneqq [-M_{\mu,\nu}/2, M_{\mu,\nu}/2]^{d_x\times d_y}$.

Comparison and Topological Properties The IGW problem is comparable to the 2-Wasserstein distance, W_2, as stated in [56, Lemma 3.2];

$$\mathsf{IGW}(\mu,\nu) \le \left(2M_2(\mu) + 2M_2(\nu)\right)^{\frac{1}{2}}\mathsf{W}_2(\mu,\nu), \quad \forall\mu,\nu \in \mathcal{P}_2(\mathbb{R}^d), \tag{3}$$

where $\mathsf{W}_2(\mu,\nu)^2 \coloneqq \inf_{\pi\in\Pi(\mu,\nu)}\int \|x-y\|^2 d\pi(x,y)$. The cited lemma also provides a lower bound in terms of W_2 (up to an orthogonal transformation of one of the marginals) when $\boldsymbol{\Sigma}_\mu$ and $\boldsymbol{\Sigma}_\nu$ are nonsingular. Our subsequent derivations require a stronger lower bound that removes this nonsingularity condition; see Appendix A.1 for the proof.

Proposition 2.2 (IGW and Wasserstein comparison)**.** *Suppose that $\mu \in \mathcal{P}_2(\mathbb{R}^{d_x})$ and $\nu \in \mathcal{P}_2(\mathbb{R}^{d_y})$ are not point masses. Letting $\mathbf{P}\boldsymbol{\Lambda}_{\pi^\star}\mathbf{Q}^\intercal$ be the singular value decomposition of $\int xy^\intercal d\pi^\star(x,y)$ for some solution $\pi^\star$ of $\mathsf{IGW}(\mu,\nu)$ with $\mathbf{P} \in \mathbb{R}^{d_x\times d_x}$, and $\mathbf{Q} \in \mathbb{R}^{d_y\times d_x}$, we have that $\sqrt{\kappa}\mathsf{W}_2((\boldsymbol{\Delta}\mathbf{P}^\intercal)_\sharp\mu, (\mathbf{Q}^\intercal)_\sharp\nu) \le \mathsf{IGW}(\mu,\nu)$, where $\boldsymbol{\Delta} = [\mathbf{I}_{d_x} \;\; \mathbf{0}]^\intercal \in \mathbb{R}^{d_y\times d_x}$, $\kappa \coloneqq \frac{1}{2}\min_{i\in\mathcal{N}}\left\{\int x_i^2 d(\boldsymbol{\Delta}\mathbf{P}^\intercal)_\sharp\mu(x) + \int y_i^2 d(\mathbf{Q}^\intercal)_\sharp\nu(y)\right\} > 0$, and $\mathcal{N} \coloneqq \left\{i \in \{1,\dots,d_y\} : \int x_i^2 d(\boldsymbol{\Delta}\mathbf{P}^\intercal)_\sharp\mu(x) + \int y_i^2 d(\mathbf{Q}^\intercal)_\sharp\nu(y) \ne 0\right\}$.*

Proposition 2.2 and its proof enable us to characterize convergence under IGW. To our knowledge, this result is new and stands in analogy to the equivalence between p-Wasserstein convergence and weak convergence of measures plus convergence of p-th moments, see [57, Theorem 6.9].

Proposition 2.3 (IGW and weak convergence)**.** *Fix $(\mu_n)_{n\in\mathbb{N}} \subset \mathcal{P}_2(\mathbb{R}^d)$ and $\mu \in \mathcal{P}_2(\mathbb{R}^d)$. Then, $\mathsf{IGW}(\mu_n,\mu) \to 0$ if and only if there exists $\boldsymbol{\Delta}_n \in \mathrm{St}(d,d)$ such that $(\boldsymbol{\Delta}_n)_\sharp\mu_n \xrightarrow{w} \mu$ and $M_2(\mu_n) \to M_2(\mu)$. Furthermore, there exists $\boldsymbol{\Delta} \in \mathrm{St}(d,d)$, such that $\boldsymbol{\Delta}_\sharp\mu_n \xrightarrow{w} \mu$ along a subsequence.*

Proposition 2.3 is proved in Appendix A.2. Observe that the sequence $\mu_n = \left((-1)^n\mathrm{Id}\right)_\sharp\mu$ satisfies $\mathsf{IGW}(\mu_n,\mu) = 0$ for every $n \in \mathbb{N}$, but μ_{2n} and μ_{2n+1} converge weakly to μ and $(-\mathrm{id})_\sharp\mu$ respectively. As such, it is necessary to pass to a subsequence to guarantee that $\boldsymbol{\Delta}_\sharp\mu_n \xrightarrow{w} \mu$ for some $\boldsymbol{\Delta} \in \mathrm{St}(d,d)$.

3 Closed-Form Solutions for Univariate IGW Alignment

Computation of a sliced IGW distance requires a means to efficiently solve univariate IGW problems. By comparison, the p-Wasserstein distance on the line has $\mathsf{W}_p(\mu,\nu) = \|F_\mu^{-1} - F_\nu^{-1}\|_{L^p([0,1])}$, where

Algorithm 1 North-West corner rule

Given univariate $\mu = \sum_{i=1}^{N} a_i \delta_{x^{(i)}}$ and $\nu = \sum_{j=1}^{M} b_j \delta_{y^{(j)}}$ with sorted support points

$(i, j) \leftarrow (1, 1)$
while $i \leq N$, $j \leq M$ **do**
  $\gamma_{ij} \leftarrow \min\{a_i, b_j\}$
  $(a_i, b_j) \leftarrow (a_i - \gamma_{ij}, b_j - \gamma_{ij})$
  if $a_i = 0$ **then** $i \leftarrow i + 1$
  else if $b_j = 0$ **then** $j \leftarrow j + 1$
return $\pi_{\mathrm{NW}}(\mu, \nu) \in \Pi(\mu, \nu)$

F_μ^{-1} is the quantile function associated with $\mu \in \mathcal{P}(\mathbb{R})$. When μ and ν are uniform distributions on n points, this reduces to a sorting problem. While Theorem 4.2.4 in [34] derives a solution for univariate IGW, their result assumes absolute continuity of one of the marginals, which is violated in the discrete setting. Our next result, proven in Appendix A.3, provides a fully general treatment.

Theorem 3.1 (Univariate IGW)**.** *Fix $\mu, \nu \in \mathcal{P}_2(\mathbb{R})$ and F_μ^{-1}, F_ν^{-1} as their quantile functions. Then, $\mathsf{IGW}(\mu,\nu)^2 = M_2(\mu)^2 + M_2(\nu)^2 - 2\max\left\{\left(\int xy\, d\pi^\star(x,y)\right)^2, \left(\int xy\, d\pi_\star(x,y)\right)^2\right\}$, where $\pi^\star = \left(F_\mu^{-1}, F_\nu^{-1}\right)_\sharp \upsilon$ and $\pi_\star = \left(-F^{-1}_{(-\mathrm{id})_\sharp \mu}, F_\nu^{-1}\right)_\sharp \upsilon$, with $\upsilon = \mathrm{Unif}([0,1])$ and the optimal coupling for $\mathsf{IGW}(\mu,\nu)$ is that which achieves the maximum. Moreover,*

1. *if μ, ν are uniformly distributed on $\mathrm{spt}(\mu) = \{x^{(i)}\}_{i=1}^N$ and $\mathrm{spt}(\nu) = \{y^{(j)}\}_{j=1}^N$ with $x^{(1)} < x^{(2)} < \cdots < x^{(N)}$ and $y^{(1)} < y^{(2)} < \cdots < y^{(N)}$, then either $\pi_{\mathrm{id}} = (\mathrm{id}, T_{\mathrm{id}})_\sharp \mu$ or $\pi_{\check{\mathrm{id}}} = (\mathrm{id}, T_{\check{\mathrm{id}}})_\sharp \mu$ is an optimal coupling where $T_{\mathrm{id}} : x^{(i)} \mapsto y^{(i)}$ and $T_{\check{\mathrm{id}}} : x^{(i)} \mapsto y^{(N-i+1)}$ are the identity and anti-identity permutations.*
2. *if μ, ν have ordered support as above, but positive weights $(a_i)_{i=1}^N$, $(b_j)_{j=1}^M$ summing to 1, then either $\pi_{\mathrm{NW}}(\mu,\nu)$, the output of Algorithm 1, or $(-\mathrm{id}, \mathrm{id})_\sharp \pi_{\mathrm{NW}}((-\mathrm{id})_\sharp \mu, \nu)$ is optimal.*

To our knowledge, the above general result has not appeared previously in the literature and depends strongly on the structure of IGW. The special case in item 2 is a corrected version of Proposition 5 in [58] which includes an erroneous sorting of weight vectors and does not reverse the order of the first coordinate of the coupling with the reversed marginal. Item 1 was claimed to hold in [33] for the quadratic GW problem, but was shown to fail in [32].

4 Sliced Inner Product Gromov–Wasserstein Distance

The closed-form solution to the univariate IGW problem enables slicing methods for high-dimensional IGW alignment. To define the average-sliced IGW distance, write $\mathbb{S}^{d-1}$ for the unit sphere in $\mathbb{R}^d$ and let σ_d be the Haar measure on $\mathbb{S}^{d-1}$. We assume, without loss of generality, that $d_x \leq d_y$.

Definition 4.1 (Sliced IGW distance)**.** The average–sliced IGW distance between $\mu \in \mathcal{P}_2(\mathbb{R}^{d_x})$ and $\nu \in \mathcal{P}_2(\mathbb{R}^{d_y})$ is $\underline{\mathsf{IGW}}(\mu,\nu) := \inf_{\boldsymbol{\Delta} \in \mathrm{St}(d_x, d_y)} \left(\int_{\mathbb{S}^{d_y-1}} \mathsf{IGW}\left((\theta^\intercal \boldsymbol{\Delta})_\sharp \mu, (\theta^\intercal)_\sharp \nu\right)^2 d\sigma_{d_y}(\theta)\right)^{1/2}$.

The above definition differs from standard sliced Wasserstein distances by the outside minimization over the Stiefel manifold. This step is used to retain the orthogonal invariance of the original IGW distance [33] (see Proposition 4.4). For instance, consider a uniform distribution over $[-1,1]\times[-2,2]$ and a version which is rotated by $90°$. Projecting with angle $(1,0)$, we get uniform distributions on $[-1,1]$ and $[-2,2]$ respectively, which have nonzero IGW distance. By continuity, the integrated IGW would be nonzero without optimizing over rotations, see Appendix H.1 for a numerical study.

The slicing procedure enables us to leverage the formula for univariate IGW described above to accelerate computation. Notably, solving univariate IGW requires the same sorting-based procedure as sliced Wasserstein distances and sliced MMD with certain choices of kernel [59]. Thus, while sliced Wasserstein/MMD could be made invariant to orthogonal transformations by adding a similar minimization over elements of the Stiefel manifold (à la Procrustes Wasserstein [60]), the computational, structural, and statistical properties of such schemes are not expected to differ

substantially from those developed herein. In effect, the computational overhead of minimizing a nonsmooth function (due to sorting) over the Stiefel manifold appears unavoidable and the empirical approximation rates match those for sliced Wasserstein distances up to log factors, see Section 4. We compare the proposed sliced IGW to other methods for aligning distributions in Appendix H.2.

We underscore that the infimum over the Stiefel manifold is always attained.

Proposition 4.2 (Attainment)**.** *Fix $\mu \in \mathcal{P}_2(\mathbb{R}^{d_x})$ and $\nu \in \mathcal{P}_2(\mathbb{R}^{d_y})$. Then, there exists $\bar{\mathbf{\Delta}} \in \mathrm{St}(d_x, d_y)$ for which $\underline{\mathsf{IGW}}(\mu,\nu)^2 = \int_{\mathbb{S}^{d_y-1}} \mathsf{IGW}((\theta^\intercal \bar{\mathbf{\Delta}})_\sharp \mu, (\theta^\intercal)_\sharp \nu)^2 d\sigma_{d_y}(\theta)$.*

Proposition 4.2 follows by leveraging compactness of $\mathrm{St}(d_x, d_y)$ and continuity of the objective in $\mathbf{\Delta}$ as described in Appendix A.4. As an example, we present a closed-form solution for $\underline{\mathsf{IGW}}$ between centered Gaussians, which we denote by $\mathcal{N}(0, \mathbf{\Sigma})$ where $\mathbf{\Sigma}$ is the covariance matrix, see Appendix C for details. This characterization provides a baseline for subsequent numerical experiments.

Example 4.3 ($\underline{\mathsf{IGW}}$ between Gaussians)**.** For $\mathbf{\Sigma}_\mu \in \mathbb{R}^{d_x \times d_x}$ and $\mathbf{\Sigma}_\nu \in \mathbb{R}^{d_y \times d_y}$, with eigenvalues $\lambda_i(\mathbf{\Sigma}_\mu)_{i=1}^{d_x}$ ordered from largest to smallest and similarly for $\mathbf{\Sigma}_\nu$,

$$\underline{\mathsf{IGW}}(\mathcal{N}(0,\mathbf{\Sigma}_\mu), \mathcal{N}(0,\mathbf{\Sigma}_\nu))^2 = \frac{\big(\mathrm{Tr}(\mathbf{\Sigma}_\mu) - \mathrm{Tr}(\mathbf{\Sigma}_\nu)\big)^2 + 2\sum_{i=1}^{d_y}\big(\lambda_i(\mathbf{\Sigma}_\mu) - \lambda_i(\mathbf{\Sigma}_\nu)\big)^2}{d_y(d_y+2)}, \tag{4}$$

where $\lambda_i(\mathbf{\Sigma}_\mu) = 0$ for $i > d_x$. Furthermore, the infimum over $\mathrm{St}(d_x, d_y)$ is achieved at $\mathbf{\Delta} = \mathbf{V}\mathbf{J}\mathbf{U}^\intercal$ where $\mathbf{J} = [\mathbf{I}_{d_x}\ \mathbf{0}]^\intercal$ and $\mathbf{U}$, $\mathbf{V}$ are obtained by diagonalizing $\mathbf{\Sigma}_\mu = \mathbf{U}\mathbf{D}_\mu\mathbf{U}^\intercal$ and $\mathbf{\Sigma}_\nu = \mathbf{V}\mathbf{D}_\nu\mathbf{V}^\intercal$.

We now move to a comprehensive study of metric, topological, and computational properties of $\underline{\mathsf{IGW}}$.

Metric and Topological Structure We first show that, like IGW itself, the sliced variant defines a pseudometric which, under primitive conditions, nullifies if and only if the high-dimensional IGW distance is zero. To state this result, we recall that the characteristic function of $\mu \in \mathcal{P}(\mathbb{R}^d)$ is given by $\Phi_\mu(t) \coloneqq \mathbb{E}_\mu[e^{it^\intercal X}],\ t \in \mathbb{R}^d$. We say that Φ_μ is analytic if, for each $t \in \mathbb{R}^d$ there exists a neighborhood U of t on which Φ_μ admits an absolutely and uniformly convergent power series representation [61, Definition 2.3.2].

Proposition 4.4 (Pseudometric)**.** *$\big(\mathcal{P}_2(\mathbb{R}^d), \underline{\mathsf{IGW}}\big)$ is a pseudometric space. Furthermore (i) if $d > 2$ there exist $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$ for which $\underline{\mathsf{IGW}}(\mu,\nu) = 0$, but $\mathsf{IGW}(\mu,\nu) \neq 0$ and (ii) if $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$ have analytic characteristic functions, $\underline{\mathsf{IGW}}(\mu,\nu) = 0 \iff \mathsf{IGW}(\mu,\nu) = 0$.*

It is easy to see that if $\mathsf{IGW}(\mu,\nu) = 0$, then $\underline{\mathsf{IGW}}(\mu,\nu) = 0$. The converse requires more care and is shown to fail without the analyticity assumption. Intuitively, if $\underline{\mathsf{IGW}}(\mu,\nu) = 0$, this only gives information about (almost every) one-dimensional projection of the measures. Lifting this information to the original measures requires certain structural assumptions such as analyticity of the characteristic function as illustrated in the proof (Appendix A.5). This condition holds, for instance, if μ, ν are compactly supported, see [62, Theorem 7.1.14]. We further show in Appendix D that when μ, ν are supported on a lattice, we again have $\underline{\mathsf{IGW}}(\mu,\nu) = 0 \iff \mathsf{IGW}(\mu,\nu) = 0$.

Next we consider the (quotient) topology induced by the sliced IGW distance and establish its equivalence to that of IGW under suitable conditions.

Proposition 4.5 (Topology)**.** *Fix $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ and $(\mu_n)_{n\in\mathbb{N}} \subset \mathcal{P}_2(\mathbb{R}^d)$. If $\mathsf{IGW}(\mu_n,\mu) \to 0$, then $\underline{\mathsf{IGW}}(\mu_n,\mu) \to 0$. Further, if $\underline{\mathsf{IGW}}(\mu_n,\mu) \to 0$ and every weak subsequential limit of μ_n has an analytic characteristic function, then $\mathsf{IGW}(\mu_n,\mu) \to 0$ if μ has an analytic characteristic function.*

As in the proof of Proposition 4.4, the difficulty stems from showing that $\underline{\mathsf{IGW}}(\mu_n,\mu) \to 0$ implies that $\mathsf{IGW}(\mu_n,\mu) \to 0$, see Appendix A.6 for complete details.

We conclude with a comparison inequality between sliced IGW and the sliced 2-Wasserstein distance, defined as $\underline{\mathsf{W}}_2(\alpha,\beta)^2 \coloneqq \int_{\mathbb{S}^{d-1}} \mathsf{W}_2\big((\theta^\intercal)_\sharp\alpha, (\theta^\intercal)_\sharp\beta\big)^2 d\sigma_d(\theta)$ for $\alpha, \beta \in \mathcal{P}_2(\mathbb{R}^d)$.

Proposition 4.6 (Comparison)**.** *Fix $\mu \in \mathcal{P}_2(\mathbb{R}^{d_x})$, $\nu \in \mathcal{P}_2(\mathbb{R}^{d_y})$ and $\mathbf{\Delta} = [\mathbf{I}_{d_x}\ \mathbf{0}]^\intercal \in \mathbb{R}^{d_y \times d_x}$,*

$$\underline{\mathsf{IGW}}(\mu,\nu)^2 \leq \int_{\mathbb{S}^{d-1}} \mathsf{IGW}\big((\theta^\intercal\mathbf{\Delta})_\sharp\mu, (\theta^\intercal)_\sharp\nu\big)^2 d\sigma_{d_y}(\theta) \leq 2\big(M_2(\mu) + M_2(\nu)\big)\,\underline{\mathsf{W}}_2((\mathbf{\Delta})_\sharp\mu, \nu)^2. \tag{5}$$

This result follows essentially by applying (3) slice-wise and enables transferring known empirical convergence rates for $\underline{\mathsf{W}}_2$ to $\underline{\mathsf{IGW}}$, as we may control $\mathbb{E}[\underline{\mathsf{IGW}}(\hat{\mu}_n,\mu)]$ by $\sqrt{\mathbb{E}[\underline{\mathsf{W}}_2(\hat{\mu}_n,\mu)^2]}$, where

$\hat{\mu}_n \coloneqq \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ is the empirical measure from n independent and identically distributed (i.i.d.) samples $X_1, \dots, X_n \overset{i.i.d.}{\sim} \mu$. For instance, the following result follows by applying [20, Corollary 2], see Appendix A.7 for the proof of Proposition 4.6 and this corollary.

Corollary 4.7 (Empirical convergence rates)**.** *Let $\mu \in \mathcal{P}_q(\mathbb{R}^{d_x})$, $\nu \in \mathcal{P}_q(\mathbb{R}^{d_y})$ for $q > 2$ and $n \geq 2$. There exists $C_q > 0$ depending only on q such that (i) $\mathbb{E}\big[\underline{\mathsf{IGW}}(\hat{\mu}_n, \mu)\big] \leq C_q \sqrt{M_2(\mu)}\, M_q(\mu)^{1/q} r_q(n)$, and (ii) $\mathbb{E}\Big[\big|\underline{\mathsf{IGW}}(\hat{\mu}_n, \hat{\nu}_n) - \underline{\mathsf{IGW}}(\mu, \nu)\big|\Big] \leq C_q\Big(\sqrt{M_2(\mu)}\, M_q(\mu)^{1/q} + \sqrt{M_2(\nu)}\, M_q(\nu)^{1/q}\Big) r_q(n)$, where $r_q(n) = n^{-(q-2)/(2q)}$ if $q \in (2,4)$ and $r_q(n) = n^{-1/4}(\log n)^{\mathbb{1}\{q=4\}1/2}$ if $q \geq 4$.*

Faster rates are attainable under additional assumptions. To wit, if μ is log-concave, $\mathbb{E}[\underline{\mathsf{IGW}}(\hat{\mu}_n, \mu)] = O(\sqrt{\log(n)/n})$, which is optimal up to $\log$ factors [22, Theorem 1, Remark 11].

Averaging via Monte Carlo Integration Since $\underline{\mathsf{IGW}}$ requires integrating the one-dimensional IGW costs over $\mathbb{S}^{d_y-1}$, a natural approximation is Monte Carlo (MC) integration over random directions as in the sliced Wasserstein case [13, 22]. Fixing $\mu \in \mathcal{P}_2(\mathbb{R}^{d_x})$ and $\nu \in \mathcal{P}_2(\mathbb{R}^{d_y})$ with $d_x \leq d_y$, we write $\underline{\mathsf{IGW}}(\mu, \nu) = \inf_{\boldsymbol{\Delta} \in \mathrm{St}(d_x, d_y)} \underline{\mathsf{IGW}}_{\boldsymbol{\Delta}}(\mu, \nu)$, where $\underline{\mathsf{IGW}}_{\boldsymbol{\Delta}}(\mu, \nu)^2 \coloneqq \int_{\mathbb{S}^{d_y-1}} g_{\boldsymbol{\Delta}}(\theta)\, d\sigma_{d_y}(\theta)$, for $g_{\boldsymbol{\Delta}}(\theta) \coloneqq \mathsf{IGW}\big((\theta^\intercal \boldsymbol{\Delta})_\sharp \mu, (\theta^\intercal)_\sharp \nu\big)^2$. To approximate this integral for a fixed $\boldsymbol{\Delta} \in \mathrm{St}(d_x, d_y)$ given $\Theta_1, \dots, \Theta_m \overset{i.i.d.}{\sim} \sigma_{d_y}$, we characterize the error of the MC estimator of $\underline{\mathsf{IGW}}_{\boldsymbol{\Delta}}(\mu, \nu)^2$, defined by $\widehat{\underline{\mathsf{IGW}}}_{\boldsymbol{\Delta}, m}(\mu, \nu)^2 \coloneqq \frac{1}{m}\sum_{j=1}^m g_{\boldsymbol{\Delta}}(\Theta_j)$. This will allow us to characterize the MC error of $\widehat{\underline{\mathsf{IGW}}}_m(\mu, \nu)^2 \coloneqq \inf_{\boldsymbol{\Delta} \in \mathrm{St}(d_x, d_y)} \widehat{\underline{\mathsf{IGW}}}_{\boldsymbol{\Delta}, m}(\mu, \nu)^2$.

Proposition 4.8 (MC error)**.** *Fix $\mu \in \mathcal{P}_2(\mathbb{R}^{d_x})$, $\nu \in \mathcal{P}_2(\mathbb{R}^{d_y})$, $m \geq 2$, and $\boldsymbol{\Delta} \in \mathrm{St}(d_x, d_y)$. Then, $\mathbb{E}\Big[\Big|\widehat{\underline{\mathsf{IGW}}}_{\boldsymbol{\Delta}, m}(\mu, \nu)^2 - \underline{\mathsf{IGW}}_{\boldsymbol{\Delta}}(\mu, \nu)^2\Big|\Big] \leq \frac{4\sqrt{2}(2\pi)^{1/4}(M_2(\mu)+M_2(\nu))^2}{\sqrt{m d_y}}$. Furthermore, $\mathbb{E}\Big[\Big|\widehat{\underline{\mathsf{IGW}}}_m(\mu, \nu)^2 - \underline{\mathsf{IGW}}(\mu, \nu)^2\Big|\Big] \lesssim \big(M_2(\mu)^2 + M_2(\nu)^2\big)\sqrt{\Big(d_x d_y - \frac{d_x(d_x-1)}{2}\Big)\frac{\log(m)}{m}}$.*

The proof of this result relies on Lipschitz continuity in θ and $\boldsymbol{\Delta}$ of $\mathsf{IGW}\big((\theta^\intercal \boldsymbol{\Delta})_\sharp \mu, (\theta^\intercal)_\sharp \nu\big)^2$ and employs a Rademacher symmetrization argument. Complete details are included in Appendix A.8.

To account for both the empirical error and the MC error, note that the fully data-driven estimator $\widehat{\underline{\mathsf{IGW}}}_{m,n} \coloneqq \Big(\inf_{\boldsymbol{\Delta} \in \mathrm{St}(d_x, d_y)} \frac{1}{m}\sum_{j=1}^m \mathsf{IGW}\big((\Theta_j^\intercal \boldsymbol{\Delta})_\sharp \hat{\mu}_n, (\Theta_j^\intercal)_\sharp \hat{\nu}_n\big)^2\Big)^{1/2}$, based on m MC slices and n samples from each distribution, approximates $\underline{\mathsf{IGW}}(\mu, \nu)$ with error at most $\big|\widehat{\underline{\mathsf{IGW}}}_{m,n} - \widehat{\underline{\mathsf{IGW}}}(\hat{\mu}_n, \hat{\nu}_n)\big| + \big|\widehat{\underline{\mathsf{IGW}}}(\hat{\mu}_n, \hat{\nu}_n) - \underline{\mathsf{IGW}}(\mu, \nu)\big|$ by the triangle inequality. These terms can be controlled (in expectation) by appealing to Corollary 4.7 and Proposition 4.8, noting that $|a - b| \leq \sqrt{|a^2 - b^2|}$ for $a, b \geq 0$ and applying Jensen's inequality.

Remark 4.9 (Comparison to sliced Wasserstein)**.** *Compared with the Wasserstein case [22], whose rate matches that where $\boldsymbol{\Delta}$ is fixed in Proposition 4.8, the bound for IGW incurs an additional $\sqrt{\log(m)}$ factor in the MC error. This extra logarithmic term arises from the covering argument used to control the discrepancy between infima over the Stiefel manifold and this scaling is tight (see [63]). Nevertheless, this does not preclude that a different argument could shed this factor.*

Optimization over the Stiefel Manifold To this point, we have quantified the MC error in Proposition 4.8 and derived a closed-form solution for univariate IGW in Theorem 3.1. To account for the minimization over $\mathrm{St}(d_x, d_y)$, suppose we are given fixed $(\theta^{(i)})_{i=1}^R$ (realizations of samples from σ_{d_y}) and that μ, ν are supported on at most N atoms. We aim to solve

$$\inf_{\boldsymbol{\Delta} \in \mathbb{R}^{d_y \times d_x}} F(\boldsymbol{\Delta}) \coloneqq \frac{1}{R}\sum_{i=1}^R \mathsf{IGW}\left(((\theta^{(i)})^\intercal \boldsymbol{\Delta})_\sharp \mu, ((\theta^{(i)})^\intercal)_\sharp \nu\right)^2 \text{ s.t. } \|\boldsymbol{\Delta}^\intercal \boldsymbol{\Delta} - \mathbf{I}_{d_x}\|_{\mathrm{F}}^2 = 0, \quad (6)$$

where the sum is a proxy for the integral, following Proposition 4.8. Letting $\mathbf{C}_\pi = \int xy^\intercal d\pi(x, y)$ for $\pi \in \Pi(\mu, \nu)$, Lemma B.1 yields that $\mathsf{IGW}\left((\theta^\intercal \boldsymbol{\Delta})_\sharp \mu, (\theta^\intercal)_\sharp \nu\right)^2 = \left(\theta^\intercal \boldsymbol{\Delta} \mathbf{R}_\mu \boldsymbol{\Delta}^\intercal \theta\right)^2 + \left(\theta^\intercal \mathbf{R}_\nu \theta\right)^2 +$

Algorithm 2 Subgradient method for sliced IGW

Fix $\mathbf{\Delta}_1$ with $\|\mathbf{\Delta}_1^\intercal \mathbf{\Delta}_1 - \mathbf{I}_{d_x}\|_{\mathrm{F}} \leq 1/6$, $0 \leq K \in \mathbb{N}$, $\beta > 0$, and a step sequence $(\eta_k)_{k\in\mathbb{N}}$,

1: **for** $k = 1, \dots, K$ **do**

2: $\quad \mathbf{G}_k \leftarrow DA_{[\mathbf{\Delta}_k]}(\mathbf{S}_k) + \beta \mathbf{\Delta}_k(\mathbf{\Delta}_k^\intercal \mathbf{\Delta}_k - \mathbf{I}_{d_x})$ for some $\mathbf{S}_k \in \partial F(A(\mathbf{\Delta}_k))$

3: $\quad \mathbf{\Delta}_{k+1} \leftarrow \mathbf{\Delta}_k - \eta_k \mathbf{G}_k$

$\inf_{\pi\in\Pi(\mu,\nu)} -2\left(\theta^\intercal \mathbf{\Delta}\mathbf{C}_\pi\theta\right)^2$, so that the objective in (6) is a sum of nonsmooth and nonconvex functions (in $\mathbf{\Delta}$). Indeed, $\mathbf{\Delta} \mapsto \inf_{\pi\in\Pi(\mu,\nu)} -2\left(\theta^\intercal \mathbf{\Delta}\mathbf{C}_\pi\theta\right)^2$ is not necessarily differentiable (see Proposition 4.10) and is concave (see Proposition 2.9 and Exercise 2.20 (a) in [64]). Notably, this function may fail to be differentiable at points where multiple optimal couplings exist; while one of the identity or anti-identity permutations is always optimal for fixed Δ at each θ, this depends on the choice of θ so that the infimum over $\Pi(\mu,\nu)$ cannot be replaced by a fixed evaluation at one of those two couplings.

Given the particular structure of F (nonconvex, nonsmooth, and not clearly weakly-convex), standard approaches to optimization over the Stiefel manifold [65–67] are not subject to formal convergence guarantees. To obtain provably convergent algorithms in the current setting, we thus follow the alternative approach of [68, 69] which propose to bypass the intricacies of Riemannian optimization methods via a "constraint dissolving" approach. Simply put, these works establish a correspondence between solving a problem over a Riemannian manifold and solving an unconstrained optimization problem. In the particular case of the Stiefel manifold, [69] proposes to minimize

$$H(\mathbf{\Delta}) = F(A(\mathbf{\Delta})) + \frac{\beta}{4}\left\|\mathbf{\Delta}^\intercal\mathbf{\Delta} - \mathbf{I}_{d_x}\right\|_{\mathrm{F}}^2, \text{ for } A(\mathbf{\Delta}) = \frac{1}{8}\mathbf{\Delta}\left(15\mathbf{I}_{d_x} - 10\mathbf{\Delta}^\intercal\mathbf{\Delta} + 3(\mathbf{\Delta}^\intercal\mathbf{\Delta})^2\right) \quad (7)$$

over the entire space $\mathbb{R}^{d_y\times d_x}$ for a suitable choice of $\beta > 0$. The composition with $A(\cdot)$ aims to correct for the curvature of the manifold and the added term penalizes deviations from $\mathrm{St}(d_x, d_y)$. As the new objective, H, inherits the nonconvexity and nonsmoothness of F, we minimize it via a standard subgradient method, Algorithm 2, whose convergence properties were studied in [70]. Throughout, the subdifferential of a locally Lipschitz function is understood in the sense of Clarke [71].

Proposition 4.10 (Subdifferential)**.** *For any $\mathbf{\Delta} \in \mathbb{R}^{d_y\times d_x}$, the Clarke subdifferential of H is*

$$\partial H(\mathbf{\Delta}) = DA_{[\mathbf{\Delta}]}\left(\partial F(A(\mathbf{\Delta}))\right) + \beta\mathbf{\Delta}\left(\mathbf{\Delta}^\intercal\mathbf{\Delta} - \mathbf{I}_{d_x}\right) \text{ for}$$

$$DA_{[\mathbf{\Delta}]}(\mathbf{\Xi}) = \frac{1}{8}\mathbf{\Xi}\left(15\mathbf{I}_{d_x} - 10\mathbf{\Delta}^\intercal\mathbf{\Delta} + 3(\mathbf{\Delta}^\intercal\mathbf{\Delta})^2\right) - \mathbf{\Delta}\Phi\left(\mathbf{\Delta}^\intercal\mathbf{\Xi}\right) + \frac{3}{2}\mathbf{\Delta}\Phi\left(\Phi(\mathbf{\Delta}^\intercal\mathbf{\Xi})\left(\mathbf{\Delta}^\intercal\mathbf{\Delta} - \mathbf{I}_{d_x}\right)\right)$$

where, for a square matrix $\mathbf{M}$, we set $\Phi(\mathbf{M}) = \frac{1}{2}(\mathbf{M} + \mathbf{M}^\intercal)$, and

$$\partial F(\mathbf{\Delta}) = \frac{1}{R}\sum_{i=1}^{R}\left(4\left((\theta^{(i)})^\intercal\mathbf{\Delta}\mathbf{R}_\mu\mathbf{\Delta}^\intercal\theta^{(i)}\right)\left(\mathbf{R}_\mu\mathbf{\Delta}^\intercal\theta^{(i)}(\theta^{(i)})^\intercal\right)^\intercal + \partial G^{\theta^{(i)}}(\mathbf{\Delta})\right),$$

where $\partial G^\theta(\mathbf{\Delta}) = \mathrm{conv}\left\{-4\theta^\intercal\mathbf{\Delta}\mathbf{C}_{\pi_{\theta,\mathbf{\Delta}}}\theta\left(\mathbf{C}_{\pi_{\theta,\mathbf{\Delta}}}\theta\theta^\intercal\right)^\intercal : \pi_{\theta,\mathbf{\Delta}} \in \arg\min_{\pi\in\Pi(\mu,\nu)} -2\left(\theta^\intercal\mathbf{\Delta}\mathbf{C}_\pi\theta\right)^2\right\}$ for each $\theta \in \mathbb{S}^{d_y-1}$ and $\mathbf{\Delta} \in \mathbb{R}^{d_y\times d_x}$, with $\mathrm{conv}(\cdot)$ denoting the closed convex hull of a set.

The proof of Proposition 4.10 follows by applying a known chain rule for H, Proposition 3.6 in [69], and by computing the subdifferential of F, see Appendix A.9 for details.

As computing a subgradient of H at a point requires computing an optimal coupling of the IGW problem between univariate distributions, it is worth recalling that Theorem 3.1 provides a closed form solution to this problem. If the sliced marginals are uniformly distributed on the same number of atoms an optimal coupling can be found by sorting and Algorithm 1 can be applied if they are not.

Having characterized the subdifferential, we employ Algorithm 2 to obtain critical points of (7). Namely, the algorithm aims to recover a point $\bar{\mathbf{\Delta}} \in \mathbb{R}^{d_y\times d_x}$ for which $0 \in \partial H(\bar{\mathbf{\Delta}})$ which, by Proposition 2.3.2 in [71], is a necessary first order condition for optimality. By contrast, a Riemannian critical point for the original problem (6) is a point $\mathbf{\Delta} \in \mathrm{St}(d_x, d_y)$ for which $0 \in \partial_R F(\mathbf{\Delta})$, where $\partial_R F(\mathbf{\Delta})$ is the Riemannian Clarke subdifferential [72, Theorem 5.1], $\partial_R F(\mathbf{\Delta}) = \left\{\mathbf{Z} - \frac{1}{2}\mathbf{\Delta}(\mathbf{\Delta}^\intercal\mathbf{Z} + \mathbf{Z}^\intercal\mathbf{\Delta}) : \mathbf{Z} \in \partial F(\mathbf{\Delta})\right\}$. Here, the term $\mathbf{Z} - \frac{1}{2}\mathbf{\Delta}(\mathbf{\Delta}^\intercal\mathbf{Z} + \mathbf{Z}^\intercal\mathbf{\Delta})$ corresponds to the projection of $\mathbf{Z} \in \partial F(\mathbf{\Delta})$ onto the tangent space to $\mathrm{St}(d_x, d_y)$ at $\mathbf{\Delta}$ (cf. e.g., [73] Example 3.6.2). The derivation of this formula is given in Appendix E. Remarkably, despite this difference, the cluster points of Algorithm 2 correspond to Riemannian critical points for F.

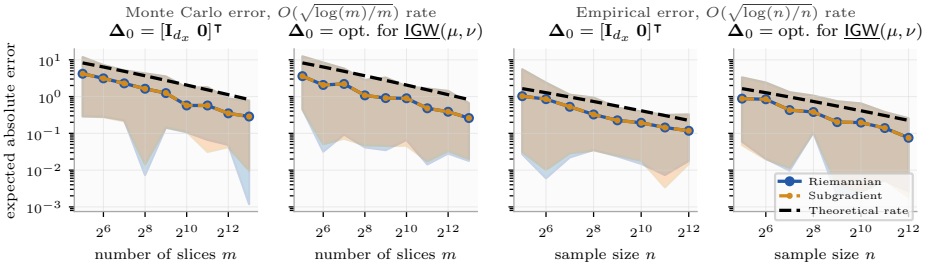


Figure 1: **Left:** Validation of the MC error from Proposition 4.8. The estimated value of $\mathbb{E}[|\widehat{\underline{\mathsf{IGW}}}_m(\mu,\nu)^2 - \underline{\mathsf{IGW}}(\mu,\nu)^2|]$ follows the theoretical rate $O(\sqrt{\log(m)/m})$ across optimization methods and initializations. **Right:** Validation of the empirical convergence rate from Corollary 4.7. For fixed m, the estimated value of $\mathbb{E}[|\widehat{\underline{\mathsf{IGW}}}_{m,n} - \underline{\mathsf{IGW}}(\mu,\nu)|]$ scales as $C_1 + C_2\sqrt{\log(n)/n}$. In both plots the shaded area corresponds to the maximum and minimum values from the 25 realizations.

Theorem 4.11 (Convergence of Algorithm 2)**.** *Suppose that the step sequence η_k satisfies $\sum_{k=1}^{\infty}\eta_k = \infty$, $\sum_{k=1}^{\infty}\eta_k^2 < \infty$, and $0 \le \eta_k \le \frac{1}{2\beta}$ where $\beta \ge \max\{162\alpha, 2L_1\}$ for α and L_1 as in Proposition F.4 and Corollary F.1. Then, the iterates $(\mathbf{\Delta}_k)_{k\in\mathbb{N}}$ generated by Algorithm 2 are bounded and each cluster point of $(\mathbf{\Delta}_k)_{k\in\mathbb{N}}$ is a Riemannian critical point of F over the Stiefel manifold.*

The proof of this result relies on the convergence of subgradient methods for certain generic classes of functions [70] and the properties of the modified objective H; see Appendix F for details. In particular, we leverage the fact that if $0 \in \partial H(\mathbf{\Delta})$ for some $\mathbf{\Delta} \in \mathrm{St}(d_x, d_y)$, then $0 \in \partial_R F(\mathbf{\Delta})$. This framework is compatible with the constraint-dissolving approach we adopt because of the correspondence it provides between stationary points of H and Riemannian stationary points of F. Extending our analysis to the nonasymptotic regime is of interest, but convergence rate results for subgradient methods [74–76] only guarantee a weaker version of stationarity which forfeits the correspondence to Riemannian stationary points. Given that the algorithm is only shown to converge asymptotically to a critical point a large number of iterations may be required in practice leading to slow convergence as seen in the experiments. The per-iteration complexity of Algorithm 2 is characterized in Appendix G.1.

5 Experiments

Validation of Monte-Carlo error and convergence rate. These experiments empirically validate Corollary 4.7 and Proposition 4.8 up to the optimization error. Since this requires access to the population value of $\underline{\mathsf{IGW}}(\mu,\nu)$, we use the Gaussian setting from Example 4.3. Specifically, we take $\mu = \mathcal{N}(0, \mathbf{\Sigma}_\mu)$ and $\nu = \mathcal{N}(0, \mathbf{\Sigma}_\nu)$ with $d_x = 5$ and $d_y = 10$, where the covariance matrices are randomly generated. In this setting, both $\underline{\mathsf{IGW}}(\mu,\nu)$ and the projected one-dimensional costs can be computed directly from covariance matrices, without discretizing μ and ν, see Appendix G.1 for details and the expressions for $\mathbf{\Sigma}_\mu, \mathbf{\Sigma}_\nu$.

We first illustrate the $\sqrt{\log(m)/m}$ MC rate for $\mathbb{E}\big[\big|\widehat{\underline{\mathsf{IGW}}}_m(\mu,\nu)^2 - \underline{\mathsf{IGW}}(\mu,\nu)^2\big|\big]$, where the expectation is over the i.i.d. projection directions $\Theta_1, \dots, \Theta_m \sim \sigma_{d_y}$, recall Proposition 4.8. For each $m \in \{2^r\}_{r=5}^{13}$, we draw a fresh collection of m directions, approximately compute $\widehat{\underline{\mathsf{IGW}}}_m(\mu,\nu)^2 = \inf_{\mathbf{\Delta}\in\mathrm{St}(d_x,d_y)} \widehat{\underline{\mathsf{IGW}}}_{\mathbf{\Delta},m}(\mu,\nu)^2$, and repeat this procedure 25 times. We compare the subgradient method from Algorithm 2 with a practical Riemannian subgradient method described in Appendix G.2; only the former is covered by Theorem 4.11, but the two methods produce similar values in this experiment. We also compare two initializations: $\mathbf{\Delta}_0 = [\mathbf{I}_{d_x}\ \mathbf{0}]^\intercal$ and the population optimizer from Example 4.3. The results in Fig. 1 are consistent with the theoretical rate.

We next validate the empirical convergence rate from Corollary 4.7. Fixing $m = 3000$ and $n \in \{2^r\}_{r=5}^{12}$, we construct $\hat{\mu}_n, \hat{\nu}_n$ from n samples, sample m projection directions, compute $\widehat{\underline{\mathsf{IGW}}}_{m,n}$, and average $\big|\widehat{\underline{\mathsf{IGW}}}_{m,n} - \underline{\mathsf{IGW}}(\mu,\nu)\big|$ over 25 repetitions with new samples. Since μ and ν are Gaussian, the discussion after Corollary 4.7 yields a rate of $C_1 + C_2\sqrt{\log(n)/n}$ for $C_1, C_2 > 0$. The results in Fig. 1 again agree with the theory even though $\widehat{\underline{\mathsf{IGW}}}_{m,n}$ is only estimated using the optimization

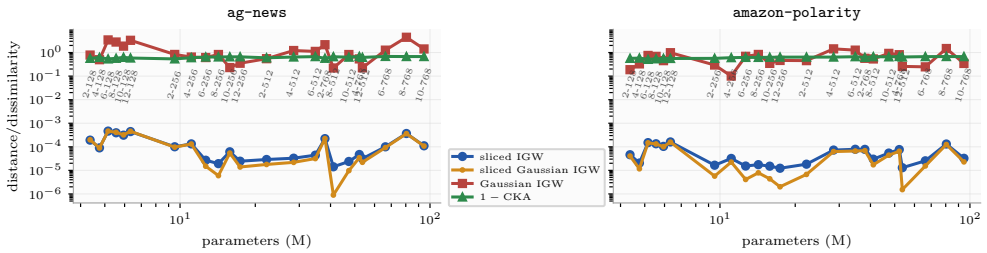


Figure 2: Comparison of distances between embeddings of sentences from `ag-news` and `amazon-polarity`. Each point corresponds to the distance between the embeddings generated by the student and teacher models with the various metrics. The labels indicate the number of layers L and embedding dimension H for the student architecture, abbreviated as L-H.

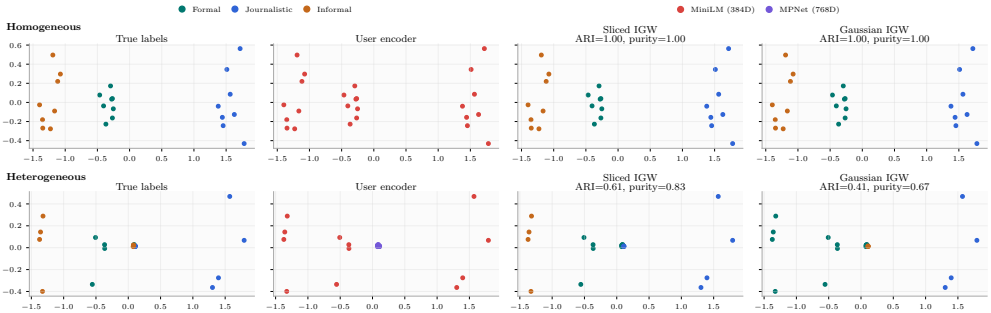


Figure 3: Results of user clustering experiment. The 2d representations of the users are obtained via 2d MDS applied to the sliced IGW distance matrix. The top row corresponds to the homogeneous experiment whereas the bottom row is the heterogeneous one. The columns are organized as follows: **Left:** ground truth clusters, **Center left:** model used to encode the user texts, **Center right:** result of clustering based on sliced IGW pairwise distances, **Right:** clustering from Gaussian IGW pairwise distances. Cluster colors in the two right columns are for illustrative purposes.

methods which are not guaranteed to achieve global optimality. See Appendix G.1 for further details including runtime. Appendix H.3 includes a similar experiment with heavy-tailed distributions admitting finite moments only of order $q \in (2, 4)$ to illustrate the slow rates in Corollary 4.7.

Comparison of LLM representations. Following the representation-comparison framework of [35], we compare embeddings generated by a teacher model, `bert-base-uncased`, and the 23 distilled BERT student models from [77]. The teacher has $L = 12$ transformer layers and embedding dimension $H = 768$, while the students vary over depths $L \in \{2, 4, 6, 8, 10, 12\}$ and embedding dimensions $H \in \{128, 256, 512, 768\}$, excluding the teacher-sized model. Intuitively, the students are trained to recover the teacher's predictions, and hence are expected to learn representations with similar geometry, although these representations need not be expressed in a common basis. This renders invariant distances like IGW and sliced IGW particularly suitable for this comparison.

We first embed 800 text passages from the `amazon-polarity` [78] and `ag-news` [79] datasets using the teacher and student models. For a student-teacher pair, μ^S denotes the centered empirical distribution over the student embeddings and ν^T is defined similarly for the teacher. We next compute $\underline{\mathrm{IGW}}(\mu^S, \nu^T)$ using the Riemannian subgradient method with $m = 500$ slices initialized at the Gaussian alignment from Example 4.3. We compare this with $\underline{\mathrm{IGW}}(\mathcal{N}(0, \mathbf{\Sigma}_S), \mathcal{N}(0, \mathbf{\Sigma}_T))$ and $\mathrm{IGW}(\mathcal{N}(0, \mathbf{\Sigma}_S), \mathcal{N}(0, \mathbf{\Sigma}_T))$ (which is available in closed form, see Appendix G.3), where $\mathbf{\Sigma}_S, \mathbf{\Sigma}_T$ are empirical covariance matrices, and $1 - \mathrm{CKA}(X^S, Y^T)$ [80], the centered kernel alignment of the embeddings X^S and Y^T from the student and teacher, see (20) for a definition. The results are shown in Fig. 2, demonstrating that the empirical sliced IGW distance is consistent with the trends of the other metrics while forgoing the assumption that the data is paired or marginally Gaussian. Importantly, the raw values are not comparable and there is no ground truth to benchmark against so that only the trends of the different metrics can be compared. Dataset sources, licenses, and runtime details are reported in Appendix G.3.

Clustering heterogeneous user text data. Our final application is to clustering users which are represented by a distribution of text embeddings. Precisely, we form 8 synthetic users for each of the primary styles: Formal (`arXiv-summarization` [81]), Journalistic (`ag-news` [79]), and Informal (`IMDB` [82]) based on 120 sentence embeddings per user for a total of 24 users. We consider both a homogeneous setting, where all user text is encoded with `all-MiniLM-L6-v2` in dimension 384, and a heterogeneous setting, where half of the users of each style are encoded with `all-MiniLM-L6-v2` and the rest with `all-mpnet-base-v2` in dimension 768. Each user is built from 110 texts in their primary style and 10 documents from the 2 non-primary styles (the number of each of the 2 styles is chosen at random) to complicate the task.

To cluster the users according to their primary style, we compute the pairwise IGW between Gaussian approximations of the user data embeddings as in the previous experiment and the pairwise sliced IGW between empirical measures (again with the Riemannian subgradient method). Importantly, sliced IGW forgoes the restrictive modeling assumptions of Gaussian IGW. We underscore that the centered kernel alignment cannot be applied here, as unlike the previous experiment, there is no *a priori* pairing of the data points. For each distance matrix $\mathbf{D}$, we form a self-tuning affinity matrix with entries $\mathbf{K}_{ij} = \exp(-\mathbf{D}_{ij}^2/(\sigma_i\sigma_j))$, where σ_i is the distance from the i-th user to its third nearest neighbor and then apply spectral clustering [83] with 3 clusters to estimate the true clusters.

To visualize the results of the clustering in Fig. 3, we apply multidimensional scaling (MDS) to obtain 2d embeddings of the users using the sliced-IGW pairwise distance matrix. We repeat the process described above three times with new users generated by randomly permuting the embedded source documents. In all three repetitions, both methods recover the true clusters exactly in the homogeneous setting, achieving an Adjusted Rand Index (ARI) [84] of 1 and purity score of 1. In the heterogeneous setting, the clustering from sliced IGW achieves ARI values of $(0.61, 0.44, 0.52)$ and purity scores of $(0.83, 0.75, 0.79)$, compared with ARI values of $(0.41, 0.17, 0.17)$ and purity scores of $(0.67, 0.67, 0.67)$ for Gaussian IGW where the i-th coordinate corresponds to the i-th experiment. We attribute this improved performance to the fact that sliced IGW captures the underlying structure of the data without enforcing a Gaussianity modeling assumption. Dataset sources, licenses, and runtime details are reported in Appendix G.4. We also include an ablation study in Appendix H.4 to analyze the effects of the experimental choices on these results.

6 Conclusion and Future Directions

This work has demonstrated that the IGW problem between one-dimensional distributions admits a closed form solution, enabling the development of a slicing paradigm for IGW. The metric and topological properties of the average-sliced IGW were derived and shown to coincide with those of the standard IGW under structural conditions and, surprisingly, may differ if those conditions are not met. In addition, the rate of convergence for empirical sliced IGW was provided along with an analysis of the MC error induced when approximating the integral over the sphere by a finite number of slices. A subgradient method was provided to account for the minimization over the Stiefel manifold which was shown to provably converge to a Riemannian critical point. Synthetic numerical experiments were provided to empirically validate the theoretical results in addition to experiments on comparing learned embeddings from LLMs and clustering heterogeneous data.

A number of important open questions persist in the study of sliced GW problems. Chief among them is to establish other costs for which the univariate GW problem can be solved exactly. Another important direction is to develop more efficient algorithms for handling the optimization over the Stiefel manifold. In effect, the proposed approach based on the constraint dissolving technique was considered as it can be shown to converge to a Riemannian critical point whereas methods based on direct Riemannian optimization techniques do not appear to be provably convergent for the objective at hand, but are expected to be more computationally efficient. Furthermore, as noted previously, if $\underline{\mathsf{IGW}}(\mu, \nu) = 0$, $\mathsf{IGW}(\mu, \nu)$ is not necessarily zero unless additional conditions are met (e.g., μ, ν have analytic characteristic functions or are supported on lattices). It is thus of interest to characterize more classes of distributions under which this implication does hold.

AI Usage Statement: ChatGPT 6 was used to assist with copy editing and assisted with writing code for Section H: Additional Experiments.

Acknowledgments and Disclosure of Funding

Z. Goldfeld is partially supported by NSF grants CCF-1947801, CCF-2046018, and DMS-2210368, and the 2020 IBM Academic Award.

References

[1] Facundo Mémoli. Gromov–Wasserstein distances and the metric approach to object matching. *Foundations of Computational Mathematics*, 11:417–487, 2011.

[2] Samir Chowdhury and Facundo Mémoli. The Gromov–Wasserstein distance between networks and stable network invariants. *Information and Inference: A Journal of the IMA*, 8(4):757–787, 2019.

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

[4] Gabriel Rioux, Ziv Goldfeld, and Kengo Kato. Entropic Gromov–Wasserstein distances: Stability and algorithms. *Journal of Machine Learning Research*, 25(363):1–52, 2024.

[5] Gabriel Rioux, Ziv Goldfeld, and Kengo Kato. Limit laws for Gromov–Wasserstein alignment with applications to testing graph isomorphisms. *arXiv preprint arXiv:2410.18006*, 2024.

[6] Gabriel Rioux, Joanna Marks, Riccardo Passeggeri, and Ziv Goldfeld. Discrete Gromov-Wasserstein duality: Algorithms and isomorphism testing. *arXiv preprint arXiv:2609.03094*, 2026.

[7] Zhengxin Zhang, Ziv Goldfeld, Youssef Mroueh, and Bharath K Sriperumbudur. Gromov–Wasserstein distances: Entropic regularization, duality and sample complexity. *The Annals of Statistics*, 52(4):1616–1645, 2024.

[8] Filippo Santambrogio. *Optimal transport for applied mathematicians*, volume 87. Springer, 2015.

[9] Julien Rabin, Gabriel Peyré, Julie Delon, and Marc Bernot. Wasserstein barycenter and its application to texture mixing. In *Scale Space and Variational Methods in Computer Vision*, volume 6667 of *Lecture Notes in Computer Science*, pages 435–446. Springer, 2011.

[10] Ishan Deshpande, Yuan-Ting Hu, Ruoyu Sun, Ayis Pyrros, Nasir Siddiqui, Sanmi Koyejo, Zhizhen Zhao, David A. Forsyth, and Alexander G. Schwing. Max-sliced Wasserstein distance and its use for GANs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10648–10656, 2019.

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

[12] Jie Li, Dan Xu, and Shaowen Yao. Sliced Wasserstein distance for neural style transfer. *Computers & Graphics*, 102:89–98, 2022.

[13] Kimia Nadjahi, Alain Durmus, Umut Simsekli, and Roland Badeau. Asymptotic guarantees for learning generative models with the sliced-Wasserstein distance. *Advances in Neural Information Processing Systems*, 32, 2019.

[14] Suman Adhya and Debarshi Kumar Sanyal. S2wtm: Spherical sliced-Wasserstein autoencoder for topic modeling. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 23211–23225, 2025.

[15] Soheil Kolouri, Phillip E. Pope, Charles E. Martin, and Gustavo K. Rohde. Sliced Wasserstein auto-encoders. In *International Conference on Learning Representations (ICLR)*, 2019.

[16] Nika Naderializadeh, Danial Salehi, Xinran Liu, and Soheil Kolouri. Constrained sliced Wasserstein embedding. *arXiv preprint arXiv:2506.02203*, 2025.

[17] March T Boedihardjo. Sharp bounds for max-sliced Wasserstein distances. *Foundations of Computational Mathematics*, pages 1–32, 2025.

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

[19] Tudor Manole, Sivaraman Balakrishnan, and Larry Wasserman. Minimax confidence intervals for the sliced Wasserstein distance. *Electronic Journal of Statistics*, 16(1):2252 – 2345, 2022.

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

[21] Kimia Nadjahi, Alain Durmus, Pierre E Jacob, Roland Badeau, and Umut Simsekli. Fast approximation of the sliced-Wasserstein distance using concentration of random projections. *Advances in Neural Information Processing Systems*, 34:12411–12424, 2021.

[22] Sloan Nietert, Ziv Goldfeld, Ritwik Sadhu, and Kengo Kato. Statistical, robustness, and computational guarantees for sliced Wasserstein distances. *Advances in Neural Information Processing Systems*, 35:28179–28193, 2022.

[23] Xinran Liu, Rocio Diaz Martin, Yikun Bai, Ashkan Shahbazi, Matthew Thorpe, Akram Aldroubi, and Soheil Kolouri. Expected sliced transport plans. *arXiv preprint arXiv:2410.12176*, 2024.

[24] Khai Nguyen, Hai Nguyen, and Nhat Ho. Fast estimation of Wasserstein distances via regression on sliced Wasserstein distances. *arXiv preprint arXiv:2509.20508*, 2025.

[25] Xinran Liu, Elaheh Akbari, Rocio Diaz Martin, Navid NaderiAlizadeh, and Soheil Kolouri. Efficient transferable optimal transport via min-sliced transport plans. *arXiv preprint arXiv:2511.19741*, 2025.

[26] Khai Nguyen and Nhat Ho. Amortized projection optimization for sliced Wasserstein generative models. *Advances in Neural Information Processing Systems*, 35:36985–36998, 2022.

[27] Khai Nguyen, Tongzheng Ren, and Nhat Ho. Markovian sliced Wasserstein distances: Beyond independent projections. *Advances in Neural Information Processing Systems*, 36:39812–39841, 2023.

[28] Pengfei Chen, Rongzhen Zhao, Tianjing He, Kongyuan Wei, and Qidong Yang. Unsupervised domain adaptation of bearing fault diagnosis based on join sliced Wasserstein distance. *ISA transactions*, 129:504–519, 2022.

[29] Yuzhe Lu, Xinran Liu, Andrea Soltoggio, and Soheil Kolouri. Slosh: Set locality sensitive hashing via sliced-Wasserstein embeddings. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2566–2576, 2024.

[30] Kimia Nadjahi, Valentin De Bortoli, Alain Durmus, Roland Badeau, and Umut Şimşekli. Approximate Bayesian computation with the sliced-Wasserstein distance. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5470–5474. IEEE, 2020.

[31] Ashkan Shahbazi, Elaheh Akbari, Darian Salehi, Xinran Liu, Navid NaderiAlizadeh, and Soheil Kolouri. Espformer: Doubly-stochastic attention with expected sliced transport plans. *arXiv preprint arXiv:2502.07962*, 2025.

[32] Robert Beinert, Cosmas Heiss, and Gabriele Steidl. On assignment problems related to Gromov–Wasserstein distances on the real line. *SIAM Journal on Imaging Sciences*, 16(2):1028–1032, 2023.

[33] Titouan Vayer, Rémi Flamary, Romain Tavenard, Laetitia Chapel, and Nicolas Courty. Sliced Gromov-Wasserstein. *Advances in Neural Information Processing Systems*, 32, 2019.

[34] Titouan Vayer. A contribution to optimal transport on incomparable spaces, 2020. PhD thesis, Institut Polytechnique de Paris.

[35] Sanjit Dandapanthula, Aleksandr Podkopaev, Shiva Prasad Kasiviswanathan, Aaditya Ramdas, and Ziv Goldfeld. Optimal transportation and alignment between Gaussian measures. *arXiv preprint arXiv:2512.03579*, 2025.

[36] Xavier Aramayo Carrasco, Maksim Nekrashevich, Petr Mokrov, Evgeny Burnaev, and Alexander Korotin. Uncovering challenges of solving the continuous Gromov-Wasserstein problem. *arXiv preprint arXiv:2303.05978*, 2023.

[37] Julie Delon, Agnes Desolneux, and Antoine Salmona. Gromov–Wasserstein distances between Gaussian distributions. *Journal of Applied Probability*, 59(4):1178–1198, 2022.

[38] Théo Dumont, Théo Lacombe, and François-Xavier Vialard. On the existence of Monge maps for the Gromov–Wasserstein problem. *Foundations of Computational Mathematics*, pages 1–48, 2024.

[39] Venkatkrishna Karumanchi, Gabriel Rioux, and Ziv Goldfeld. Approximation analysis of the entropic penalty in quadratic programming. *arXiv preprint arXiv:2509.20031*, 2025.

[40] Joanna Marks, Gabriel Rioux, and Riccardo Passeggeri. Averaged mirror descent and dual gradient methods: Convergent algorithms for entropic Gromov-Wasserstein problems. *arXiv preprint arXiv:2609.31848*, 2026.

[41] Gabriel Peyré, Marco Cuturi, and Justin Solomon. Gromov-Wasserstein averaging of kernel and distance matrices. In *International conference on machine learning*, pages 2664–2672. PMLR, 2016.

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

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

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

[45] Cédric Vincent-Cuaz, Rémi Flamary, Marco Corneli, Titouan Vayer, and Nicolas Courty. Semi-relaxed Gromov–Wasserstein divergence with applications on graphs. *arXiv preprint arXiv:2110.02753*, 2021.

[46] Tao Wang and Ziv Goldfeld. Neural entropic optimal transport and gromov–wasserstein alignment. *IEEE Transactions on Information Theory*, 72(4):2424–2443, 2026.

[47] Junyu Chen, Binh T. Nguyen, Shang Hui Koh, and Yong Sheng Soh. Semidefinite relaxations of the Gromov–Wasserstein distance. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 69814–69839. Curran Associates, Inc., 2024.

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

[49] Moritz Piening and Robert Beinert. A novel sliced fused Gromov-Wasserstein distance. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 24828–24836, 2026.

[50] Soheil Kolouri, Kimia Nadjahi, Umut Simsekli, Roland Badeau, and Gustavo Rohde. Generalized sliced Wasserstein distances. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[51] Tianyi Lin, Zeyu Zheng, Elynn Chen, Marco Cuturi, and Michael I Jordan. On projection robust optimal transport: Sample complexity and model misspecification. In *International Conference on Artificial Intelligence and Statistics*, pages 262–270. PMLR, 2021.

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

[53] Ziv Goldfeld, Kengo Kato, Gabriel Rioux, and Ritwik Sadhu. Statistical inference with regularized optimal transport. *Information and Inference: A Journal of the IMA*, 13(1):iaad056, 2024.

[54] Jiaqi Xi and Jonathan Niles-Weed. Distributional convergence of the sliced Wasserstein process. *Advances in Neural Information Processing Systems*, 35:13961–13973, 2022.

[55] Xianliang Xu and Zhongyi Huang. Central limit theorem for the sliced 1-Wasserstein distance and the max-sliced 1-Wasserstein distance. *arXiv preprint arXiv:2205.14624*, 2022.

[56] Zhengxin Zhang, Ziv Goldfeld, Kristjan Greenewald, Youssef Mroueh, and Bharath K Sriperumbudur. Gradient flows and Riemannian structure in the Gromov–Wasserstein geometry. *Foundations of Computational Mathematics*, pages 1–93, 2025.

[57] Cédric Villani. *Optimal transport: old and new*, volume 338. Springer, 2008.

[58] Clément Bonet, Titouan Vayer, Nicolas Courty, François Septier, and Lucas Drumetz. Subspace detours meet gromov–wasserstein. *Algorithms*, 14(12):366, 2021.

[59] Johannes Hertrich, Christian Wald, Fabian Altekrüger, and Paul Hagemann. Generative sliced MMD flows with Riesz kernels. In *International Conference on Learning Representations*, volume 2024, pages 20923–20949, 2024.

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

[61] Steven George Krantz. *Function theory of several complex variables*, volume 340. American Mathematical Soc., 2001.

[62] Lars Hörmander. *The analysis of Linear Partial Differential Operators I: Distribution Theory and Fourier Analysis*. Springer-Verlag, 1983.

[63] Boxin Zhao, Mladen Kolar, and Jinchi Lv. Smart: A spectral transfer approach to multi-task learning. *arXiv preprint arXiv:2604.20161*, 2026.

[64] R. Tyrrell Rockafellar and Roger JB Wets. *Variational analysis*. Springer, 1998.

[65] Zhihui Zhu, Tianyu Ding, Daniel Robinson, Manolis Tsakiris, and René Vidal. A linearly convergent method for non-smooth non-convex optimization on the Grassmannian with applications to robust subspace and dictionary learning. *Advances in Neural Information Processing Systems*, 32, 2019.

[66] Xiao Li, Shixiang Chen, Zengde Deng, Qing Qu, Zhihui Zhu, and Anthony Man-Cho So. Weakly convex optimization over Stiefel manifold using Riemannian subgradient-type methods. *SIAM Journal on Optimization*, 31(3):1605–1634, 2021.

[67] Wen Huang and Ke Wei. Riemannian proximal gradient methods. *Mathematical Programming*, 194(1):371–413, 2022.

[68] Nachuan Xiao, Xin Liu, and Kim-Chuan Toh. Dissolving constraints for Riemannian optimization. *Mathematics of Operations Research*, 49(1):366–397, 2024.

[69] Xiaoyin Hu, Nachuan Xiao, Xin Liu, and Kim-Chuan Toh. A constraint dissolving approach for nonsmooth optimization over the Stiefel manifold. *IMA Journal of Numerical Analysis*, 44(6):3717–3748, 2024.

[70] Damek Davis, Dmitriy Drusvyatskiy, Sham Kakade, and Jason D Lee. Stochastic subgradient method converges on tame functions. *Foundations of Computational Mathematics*, 20(1):119–154, 2020.

[71] Frank H. Clarke. *Optimization and Nonsmooth Analysis*. Society for Industrial and Applied Mathematics, 1990.

[72] Wei Hong Yang, Lei-Hong Zhang, and Ruyi Song. Optimality conditions for the nonlinear programming problems on Riemannian manifolds. *Pacific Journal of Optimization*, 10(2):415–434, 2014.

[73] P.-A. Absil, R. Mahony, and R. Sepulchre. *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, 2008.

[74] Jingzhao Zhang, Hongzhou Lin, Stefanie Jegelka, Suvrit Sra, and Ali Jadbabaie. Complexity of finding stationary points of nonconvex nonsmooth functions. In *International Conference on Machine Learning*, pages 11173–11182. PMLR, 2020.

[75] Damek Davis, Dmitriy Drusvyatskiy, Yin Tat Lee, Swati Padmanabhan, and Guanghao Ye. A gradient sampling method with complexity guarantees for lipschitz functions in high and low dimensions. *Advances in Neural Information Processing Systems*, 35:6692–6703, 2022.

[76] Siyu Kong and Adrian S Lewis. The cost of nonconvexity in deterministic nonsmooth optimization. *Mathematics of Operations Research*, 49(4):2385–2401, 2024.

[77] Iulia Turc, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Well-read students learn better: On the importance of pre-training compact models. In *International Conference on Learning Representations*, 2020.

[78] Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. MTEB: Massive text embedding benchmark. *arXiv preprint arXiv:2210.07316*, 2022.

[79] Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems*, 2015.

[80] Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of Neural Network representations revisited. In *International Conference on Machine Learning*, pages 3519–3529. PMLR, 2019.

[81] Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. A discourse-aware attention model for abstractive summarization of long documents. In Marilyn Walker, Heng Ji, and Amanda Stent, editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 615–621, New Orleans, Louisiana, 2018. Association for Computational Linguistics.

[82] Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In Dekang Lin, Yuji Matsumoto, and Rada Mihalcea, editors, *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA, 2011. Association for Computational Linguistics.

[83] Lihi Zelnik-Manor and Pietro Perona. Self-Tuning spectral clustering. In *Advances in Neural Information Processing Systems*, volume 17, 2004.

[84] Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of classification*, 2(1):193–218, 1985.

[85] Erhan Bayraktar and Gaoyue Guo. Strong equivalence between metrics of Wasserstein type. *Electronic Communications in Probability*, 26:1 – 13, 2021.

[86] Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019.

[87] Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, 2nd edition, 2026. Available at `https://www.math.uci.edu/~rvershyn/papers/HDP-book/HDP-2.pdf`.

[88] Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

[89] Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient flows: in metric spaces and in the space of probability measures*. Springer Science & Business Media, 2008.

[90] Gerald B. Folland. How to integrate a polynomial over a sphere. *The American Mathematical Monthly*, 108(5):446–448, 2001.

[91] John von Neumann. Some matrix inequalities and metrization of matrix space. Mitt. Forschungsinst. Math. Mech. Kujbyschew-Univ. Tomsk 1, 286-300, 1937.

[92] Vladimir I. Bogachev. *Measure Theory*, volume 1. Springer, 2007.

[93] John M Lee. *Introduction to smooth manifolds*. Springer, 2013.

[94] Michel Coste. An introduction to semialgebraic geometry. Technical report, Institut de Recherche Mathématiques de Rennes, 2002.

[95] Quentin Lhoest, Albert Villanova del Moral, Yacine Jernite, Abhishek Thakur, Patrick von Platen, Suraj Patil, Julien Chaumond, Mariama Drame, Julien Plu, Lewis Tunstall, Joe Davison, Hassan Sajjad, Gunjan Chhablani, Bhavitvya Malik, Simon Brandeis, Teven Le Scao, Victor Sanh, Canwen Xu, Nicolas Patry, Angelina McMillan-Major, Philipp Schmid, Sylvain Gugger, Clement Delangue, Théo Matussiere, Lysandre Debut, Stas Bekman, Pierric Cistac, Thibault Goehringer, Victor Mustar, François Lagunas, Alexander Rush, and Thomas Wolf. Datasets: A community library for natural language processing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 175–184. Association for Computational Linguistics, 2021.

[96] Antonio Gulli. Ag's corpus of news articles, 2004. `http://groups.di.unipi.it/~gulli/AG_corpus_of_news_articles.html`.

[97] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October 2020. Association for Computational Linguistics.

[98] Rémi Flamary, Nicolas Courty, Alexandre Gramfort, Mokhtar Z. Alaya, Aurélie Boisbunon, Stanislas Chambon, Laetitia Chapel, Adrien Corenflos, Kilian Fatras, Nemo Fournier, Léo Gautheron, Nathalie T.H. Gayraud, Hicham Janati, Alain Rakotomamonjy, Ievgen Redko, Antoine Rolet, Antony Schutz, Vivien Seguy, Danica J. Sutherland, Romain Tavenard, Alexander Tong, and Titouan Vayer. Pot: Python optimal transport. *Journal of Machine Learning Research*, 22(78):1–8, 2021.

[99] Davide Adamo, Marco Corneli, Manon Vuillien, and Emmanuelle Vila. An in depth look at the Procrustes-Wasserstein distance: properties and barycenters. In *Forty-second International Conference on Machine Learning*, 2025.

A Proofs of Main Results

A.1 Proof of Proposition 2.2

We divide the proof of Proposition 2.2 into a number of auxiliary lemmas whose proofs can be found in Appendices B.1 and B.2

Lemma A.1. *Fix $\mu \in \mathcal{P}_2(\mathbb{R}^{d_x}), \nu \in \mathcal{P}_2(\mathbb{R}^{d_y})$ and let $\pi^\star$ solve $\mathsf{IGW}(\mu,\nu)$. Then, if $\mathbf{P}\mathbf{\Lambda}_{\pi^\star}\mathbf{Q}^\intercal$ is the singular value decomposition of $\int xy^\intercal d\pi^\star(x,y)$ for $\mathbf{P} \in \mathbb{R}^{d_x \times d_x}$ and $\mathbf{Q} \in \mathbb{R}^{d_y \times d_y}$,*

$$\mathsf{W}_2(\tilde{\mu},\tilde{\nu})^2 \leq \frac{2}{\iota}\left(\mathsf{IGW}(\mu,\nu)^2 - \sum_{k\in\mathcal{K}}\left(\int x_k^2 d\tilde{\mu}(x)\right)^2 - \sum_{j\in\mathcal{J}}\left(\int y_j^2 d\tilde{\nu}(y)\right)^2\right) + \sum_{j\in\mathcal{J}}\int y_j^2 d\tilde{\nu}(y) + \sum_{k\in\mathcal{K}}\int x_k^2 d\tilde{\mu}(x),$$

where $\tilde{\mu} = (\mathbf{\Delta}\mathbf{P}^\intercal)_\sharp\mu$, and $\tilde{\nu} = (\mathbf{Q}^\intercal)_\sharp\nu$ for $\mathbf{\Delta} = [\mathbf{I}_{d_x} \quad \mathbf{0}]^\intercal \in \mathrm{St}(d_x,d_y) \subset \mathbb{R}^{d_y\times d_x}$,

$$\mathcal{I} \coloneqq \left\{i \in \{1,\dots,d_x\} : \int x_i^2 d\tilde{\mu}(x) \neq 0, \int y_i^2 d\tilde{\nu}(y) \neq 0\right\},$$
$$\mathcal{J} \coloneqq \left\{i \in \{1,\dots,d_y\} \cap \mathcal{I}^c : \int y_i^2 d\tilde{\nu}(y) \neq 0\right\}, \mathcal{K} \coloneqq \left\{k \in \{1,\dots,d_y\} \cap \mathcal{I}^c : \int x_k^2 d\tilde{\mu}(x) \neq 0\right\},$$

and $\iota \coloneqq \min_{i\in\mathcal{I}}\left\{\int x_i^2 d\tilde{\mu}(x) + \int y_i^2 d\tilde{\nu}(y)\right\}$ provided that $\mathcal{I} \neq \emptyset$.

Lemma A.2. *In the setting of Lemma A.1, if $\mathcal{I} = \emptyset$, then μ or ν is a point mass.*

Proof of Proposition 2.2. Proposition 2.2 is a direct consequence of Lemmas A.1 and A.2. Indeed, as μ and ν are not point masses, $\mathcal{I}$ is nonempty by Lemma A.2, so that Lemma A.1 is applicable. Under the convention that the minimum over an empty set is $+\infty$, the sum over an empty set is 0, and $+\infty \cdot 0 = 0$, we set $\varsigma \coloneqq \min_{j\in\mathcal{J}}\left\{\int y_j^2 d\tilde{\nu}(y)\right\}$, $\zeta \coloneqq \min_{k\in\mathcal{K}}\left\{\int x_i^2 d\tilde{\mu}(x)\right\}$, and obtain that

$$\sum_{j\in\mathcal{J}}\left(\int y_j^2 d\tilde{\nu}(y)\right)^2 \geq \varsigma\sum_{j\in\mathcal{J}}\int y_j^2 d\tilde{\nu}(y),\ \sum_{k\in\mathcal{K}}\left(\int x_i^2 d\tilde{\mu}(x)\right)^2 \geq \zeta\sum_{k\in\mathcal{K}}\int x_i^2 d\tilde{\mu}(x),$$

where we have again used the notation $\tilde{\mu} = (\mathbf{\Delta}\mathbf{P}^\intercal)_\sharp\mu$, and $\tilde{\nu} = (\mathbf{Q}^\intercal)_\sharp\nu$. Applying these bounds in Lemma A.1 yields

$$\mathsf{W}_2(\tilde{\mu},\tilde{\nu})^2 \leq \frac{2}{\iota}\left(\mathsf{IGW}(\mu,\nu)^2 - \sum_{k\in\mathcal{K}}\left(\int x_k^2 d\tilde{\mu}(x)\right)^2 - \sum_{j\in\mathcal{J}}\left(\int y_j^2 d\tilde{\nu}(y)\right)^2\right) + \varsigma^{-1}\sum_{j\in\mathcal{J}}\left(\int y_j^2 d\tilde{\nu}(y)\right)^2 + \zeta^{-1}\sum_{k\in\mathcal{K}}\left(\int x_k^2 d\tilde{\mu}(x)\right)^2,$$

where $1/\infty$ is understood as 0. Using (16) and (19), we have that $\mathsf{IGW}(\mu,\nu)^2 - \sum_{j\in\mathcal{J}} M_2(\tilde{\nu})_j^2 - \sum_{k\in\mathcal{K}} M_2(\tilde{\mu})_k^2 \geq 0$ so that, setting $\kappa' = \frac{1}{2}\min\{\iota, 2\varsigma, 2\zeta\}$, which is finite since $\mathcal{I}$ is nonempty, $\mathsf{W}_2(\tilde{\mu},\tilde{\nu})^2 \leq (\kappa')^{-1}\mathsf{IGW}(\mu,\nu)^2$. The claimed result then follows by noting that $\kappa' \geq \frac{1}{2}\min\{\iota,\varsigma,\zeta\} = \kappa$ where κ is the constant from Proposition 2.2. □

A.2 Proof of Proposition 2.3

If $\mathsf{IGW}(\mu_n, \mu) \to 0$, we have from (19) that[2]

$$\begin{aligned}
\mathsf{IGW}(\mu_n,\mu)^2 &\geq \sum_{i\in\mathcal{I}} \left(\left(\int x_i^2 d(\mathbf{\Delta}\mathbf{P}_n^\intercal)_\sharp \mu_n(x) \right)^2 + \left(\int y_i^2 d(\mathbf{Q}_n^\intercal)_\sharp \mu(y) \right)^2 - 2 \left(\int x_i y_i d\tilde{\pi}_n(x,y) \right)^2 \right) \\
&+ \sum_{k\in\mathcal{K}} \left(\int x_k^2 d(\mathbf{\Delta}\mathbf{P}_n^\intercal)_\sharp \mu_n(x) \right)^2 + \sum_{j\in\mathcal{J}} \left(\int y_j^2 d(\mathbf{Q}_n^\intercal)_\sharp \mu(y) \right)^2 \\
&= \sum_{i\in\mathcal{I}} \left(\left(\int x_i^2 d(\mathbf{\Delta}\mathbf{P}_n^\intercal)_\sharp \mu_n(x) \right)^2 + \left(\int y_i^2 d(\mathbf{Q}_n^\intercal)_\sharp \mu(y) \right)^2 - 2 \left(\int x_i y_i d\tilde{\pi}_n(x,y) \right)^2 \right) \\
&+ \sum_{k\in\mathcal{K}} \left(\left(\int x_k^2 d(\mathbf{\Delta}\mathbf{P}_n^\intercal)_\sharp \mu_n(x) \right)^2 + \left(\int y_k^2 d(\mathbf{Q}_n^\intercal)_\sharp \mu(y) \right)^2 - 2 \left(\int x_k y_k d\tilde{\pi}_n(x,y) \right)^2 \right) \\
&+ \sum_{j\in\mathcal{J}} \left(\left(\int x_j^2 d(\mathbf{\Delta}\mathbf{P}_n^\intercal)_\sharp \mu_n(x) \right)^2 + \left(\int y_j^2 d(\mathbf{Q}_n^\intercal)_\sharp \mu(y) \right)^2 - 2 \left(\int x_j y_j d\tilde{\pi}_n(x,y) \right)^2 \right) \\
&= \sum_{i=1}^{d} \left(\left(\int x_i^2 d(\mathbf{\Delta}\mathbf{P}_n^\intercal)_\sharp \mu_n(x) \right)^2 + \left(\int y_i^2 d(\mathbf{Q}_n^\intercal)_\sharp \mu(y) \right)^2 - 2 \left(\int x_i y_i d\tilde{\pi}_n(x,y) \right)^2 \right) \\
&\geq \sum_{i=1}^{d} \left(\int x_i^2 d(\mathbf{\Delta}\mathbf{P}_n^\intercal)_\sharp \mu_n(x) - \int y_i^2 d(\mathbf{Q}_n^\intercal)_\sharp \mu(y) \right)^2 \geq 0,
\end{aligned}$$

where the first equality is due to the fact that $\int y_k^2 d(\mathbf{Q}_n^\intercal)_\sharp \mu(y) = 0$ and $\int x_k y_k d\tilde{\pi}_n(x,y) = 0$ for every $k \in \mathcal{K}$ and a similar conclusion applies to $\mathcal{J}$ (recall the proof of Lemma A.1). The before last inequality follows by the Cauchy-Schwarz inequality;

$$\left(\int x_i y_i d\tilde{\pi}_n(x,y) \right)^2 \leq \int x_i^2 d\tilde{\pi}_n(x,y) \int y_i^2 d\tilde{\pi}_n(x,y) = \int x_i^2 d(\mathbf{\Delta}\mathbf{P}_n^\intercal)_\sharp \mu_n(x) \int y_i^2 d(\mathbf{Q}_n^\intercal)_\sharp \mu(y).$$

It follows that, for every $i \in \{1,\dots,d\}$, $\int x_i^2 d(\mathbf{\Delta}\mathbf{P}_n^\intercal)_\sharp \mu_n(x) - \int y_i^2 d(\mathbf{Q}_n^\intercal)_\sharp \mu(y) \to 0$ so that

$$\sum_{i=1}^{d} \int x_i^2 d(\mathbf{\Delta}\mathbf{P}_n^\intercal)_\sharp \mu_n(x) - \sum_{i=1}^{d} \int y_i^2 d(\mathbf{Q}_n^\intercal)_\sharp \mu(y) = M_2(\mu_n) - M_2(\mu) \to 0.$$

Furthermore, our previous application of the Cauchy-Schwarz inequality is asymptotically tight, that is, for each $i \in \{1,\dots,d\}$,

$$0 \leq \int x_i^2 d(\mathbf{\Delta}\mathbf{P}_n^\intercal)_\sharp \mu_n(x) \int y_i^2 d(\mathbf{Q}_n^\intercal)_\sharp \mu(y) - \left(\int x_i y_i d\tilde{\pi}_n(x,y) \right)^2 \to 0,$$

whereby

$$\begin{aligned}
&\left(\int x_i y_i d\tilde{\pi}_n(x,y) \right)^2 - \left(\int y_i^2 d(\mathbf{Q}_n^\intercal)_\sharp \mu(y) \right)^2 \\
&= \left(\int x_i y_i d\tilde{\pi}_n(x,y) \right)^2 - \int x_i^2 d(\mathbf{\Delta}\mathbf{P}_n^\intercal)_\sharp \mu_n(x) \int y_i^2 d(\mathbf{Q}_n^\intercal)_\sharp \mu(y) \\
&+ \left(\int x_i^2 d(\mathbf{\Delta}\mathbf{P}_n^\intercal)_\sharp \mu_n(x) - \int y_i^2 d(\mathbf{Q}_n^\intercal)_\sharp \mu(y) \right) \int y_i^2 d(\mathbf{Q}_n^\intercal)_\sharp \mu(y) \to 0,
\end{aligned}$$

i.e, there exists a sequence $(\alpha_n)_{n\in\mathbb{N}}$ such that $\alpha_n \in \{-1,1\}^d$ and $R_{n,i} \coloneqq (\alpha_n)_i \int x_i y_i d\tilde{\pi}_n(x,y) - \int y_i^2 d(\mathbf{Q}_n^\intercal)_\sharp \mu(y) \to 0$ for each $i \in \{1,\dots,d\}$. Conclude that

$$\begin{aligned}
&0 \leq \mathsf{W}_2\left((\mathrm{diag}(\alpha_n))_\sharp (\mathbf{\Delta}\mathbf{P}_n^\intercal)_\sharp \mu_n, (\mathbf{Q}_n^\intercal)_\sharp \mu \right)^2 \\
&\leq M_2\left((\mathbf{\Delta}\mathbf{P}_n^\intercal)_\sharp \mu_n \right) + M_2((\mathbf{Q}_n^\intercal)_\sharp \mu) - 2 \sum_{i=1}^{d} (\alpha_n)_i \int x_i y_i d\tilde{\pi}_n(x,y) \to 2M_2(\mu) - 2M_2(\mu) = 0.
\end{aligned}$$

[2]We highlight that this step in the proof does not require the assumption that μ_n and μ not be point masses.

It follows from Theorem 6.9 in [57] that $(\mathbf{Q}_n \operatorname{diag}(\alpha_n)\mathbf{\Delta}\mathbf{P}_n^\intercal)_\sharp \mu_n \xrightarrow{w} \mu$ so that the claimed result follows by noting that $\mathbf{\Delta}_n \coloneqq \mathbf{Q}_n \operatorname{diag}(\alpha_n)\mathbf{\Delta}\mathbf{P}_n^\intercal \in \mathrm{St}(d,d)$.

On the other hand, if there exists $\mathbf{\Delta}_n \in \mathrm{St}(d,d)$ such that $(\mathbf{\Delta}_n)_\sharp \mu_n \xrightarrow{w} \mu$ and $M_2\big(\mu_n\big) \to M_2(\mu)$, it holds that

$$\mathsf{IGW}(\mu_n, \mu) = \mathsf{IGW}((\mathbf{\Delta}_n)_\sharp \mu_n, \mu) \leq (2M_2(\mu_n) + 2M_2(\mu))^{\frac{1}{2}}\, \mathsf{W}_2((\mathbf{\Delta}_n)_\sharp \mu_n, \mu) \to 0,$$

as follows from invariance of IGW to unitary isomorphisms and (3).

Now, assume that $(\mathbf{\Delta}_n)_\sharp\, \mu_n \xrightarrow{w} \mu$ for some $\mathbf{\Delta}_n \in \mathrm{St}(d,d)$ and that the second moments converge. Note that the Stiefel manifold $\mathrm{St}(d,d)$ is compact in the sense that it forms a closed and bounded subset of $\mathbb{R}^{d\times d}$. Consequently, there exists a subsequence n' along which $\mathbf{\Delta}_{n'} \to \mathbf{\Delta} \in \mathrm{St}(d,d)$ as matrices. Now, for any bounded Lipschitz function $f : \mathbb{R}^d \to \mathbb{R}$ with Lipschitz constant L,

$$\begin{aligned}\left|\int f d(\mathbf{\Delta}_\sharp \mu_{n'} - \mu)\right| &\leq \left|\int f(\mathbf{\Delta}(\cdot)) - f(\mathbf{\Delta}_{n'}(\cdot)) d\mu_{n'}\right| + \left|\int f d\left((\mathbf{\Delta}_{n'})_\sharp \mu_{n'} - \mu\right)\right| \\ &\leq L\|\mathbf{\Delta} - \mathbf{\Delta}_{n'}\|_{\mathrm{F}} M_1(\mu_{n'}) + \left|\int f d\left((\mathbf{\Delta}_{n'})_\sharp \mu_{n'} - \mu\right)\right|.\end{aligned}$$

As $(\mathbf{\Delta}_n)_\sharp\, \mu_n \xrightarrow{w} \mu$, the leftmost term converges to 0 by the Portmanteau theorem. As for the first term, we have that $\|x\| \leq 1 + \|x\|^2$ for all $x \in \mathbb{R}^d$ so that Definition 6.8 in [57] asserts that $M_1(\mu_n) \to M_1(\mu)$ as $n \to \infty$ (recall that $(\mathbf{\Delta}_n)_\sharp\, \mu_n \xrightarrow{w} \mu$ and convergence of second moments implies convergence in 2-Wasserstein distance). As $\|\mathbf{\Delta} - \mathbf{\Delta}_{n'}\|_{\mathrm{F}} \to 0$, it follows that $\int f d(\mathbf{\Delta}_\sharp \mu_{n'}) \to \int f d\mu$ as $n \to \infty$. Since f was an arbitrary bounded Lipschitz function, we conclude that $\mathbf{\Delta}_\sharp \mu_{n'} \xrightarrow{w} \mu$ by the Portmanteau theorem. □

We remark here that Proposition 4.5 admits a simpler form in the case that μ and ν are distributions on the real line. We record the result here, as it will be used in the proof of Proposition 4.5 ahead.

Corollary A.3. *If $\mu, \nu \in \mathcal{P}_2(\mathbb{R})$,*

$$\frac{1}{2}(M_2(\mu) + M_2(\nu)) \min\left\{\mathsf{W}_2(\mu,\nu)^2, \mathsf{W}_2((-\operatorname{id})_\sharp \mu, \nu)^2\right\} \leq \mathsf{IGW}(\mu,\nu)^2. \tag{8}$$

If $(\mu_n)_{n\in\mathbb{N}} \subset \mathcal{P}_2(\mathbb{R})$ is such that $\mathsf{IGW}(\mu_n, \mu) \to 0$, then $(s_n \operatorname{id})_\sharp\, \mu_n \xrightarrow{w} \mu$ and $M_2(\mu_n) \to M_2(\mu)$ where

$$s_n = \begin{cases} 1, & \text{if } \mathsf{W}_2(\mu_n,\mu) \leq \mathsf{W}_2((-\operatorname{id})_\sharp \mu_n, \mu), \\ -1, & \text{if } \mathsf{W}_2(\mu_n,\mu) > \mathsf{W}_2((-\operatorname{id})_\sharp \mu_n, \mu). \end{cases}$$

If $(a_n \operatorname{id})_\sharp \mu_n \to \mu$ for some $a_n \in \{-1, 1\}$ and $M_2(\mu_n) \to M_2(\mu)$ then $\mathsf{IGW}(\mu_n, \mu) \to 0$.

The proof of Corollary A.3 is included in Appendix B.3.

A.3 Proof of Theorem 3.1

We have from Appendix E in [4] that, for any $\rho \in \mathcal{P}_2(\mathbb{R}^{d_x})$ and $\eta \in \mathcal{P}_2(\mathbb{R}^{d_y})$,

$$\mathsf{IGW}(\rho,\eta)^2 = \|\mathbf{R}_\rho\|_{\mathrm{F}}^2 + \|\mathbf{R}_\eta\|_{\mathrm{F}}^2 + \inf_{\pi\in\Pi(\rho,\eta)} -2\left\|\int xy^\intercal d\pi(x,y)\right\|_{\mathrm{F}}^2.$$

Instantiating this result with the univariate distributions μ, ν, the infimum can be written as

$$\inf_{\pi\in\Pi(\mu,\nu)} -2\left(\int xy\, d\pi(x,y)\right)^2 = -2\max\left\{\left(\int xy\, d\pi^{\min}(x,y)\right)^2, \left(\int xy\, d\pi^{\max}(x,y)\right)^2\right\}, \tag{9}$$

where $\pi^{\min} \in \arg\min_{\pi\in\Pi(\mu,\nu)} \int xy\, d\pi(x,y)$ and $\pi^{\max} \in \arg\max_{\pi\in\Pi(\mu,\nu)} \int xy\, d\pi(x,y)$. It is easy to see that

$$\begin{aligned}\arg\max_{\pi\in\Pi(\mu,\nu)} \int xy\, d\pi(x,y) &= \arg\min_{\pi\in\Pi(\mu,\nu)} \int -xy\, d\pi(x,y), \\ \arg\min_{\pi\in\Pi(\mu,\nu)} \int xy\, d\pi(x,y) &= \left\{(-\operatorname{id}, \operatorname{id})_\sharp \check{\pi} : \check{\pi} \in \arg\min_{\pi\in\Pi((-\operatorname{id})_\sharp\mu,\nu)} \int -xy\, d\pi(x,y)\right\}.\end{aligned}$$

With the objectives in this form, we note that, for any $\pi \in \Pi(\mu, \nu)$,

$$\int -2xy \, d\pi(x,y) = \int (x-y)^2 \, d\pi(x,y) - \int x^2 \, d\mu(x) - \int y^2 \, d\nu(y),$$

so that $\pi^{\max}$ corresponds to an optimal coupling for the quadratic OT problem between μ and ν and, likewise, $\pi^{\min}$ is related to a solution of the quadratic OT problem between $(-\operatorname{id})_\sharp \mu$ and ν.

It is well known (cf. e.g., [8, Theorem 2.9]) that the unique optimal coupling for the quadratic OT problem between univariate measures α, β is $\pi^\star = \left(F_\alpha^{-1}, F_\beta^{-1}\right)_\sharp(\mathrm{Unif}([0,1]))$. Applying this fact with $(\alpha, \beta) = (\mu, \nu)$ and $(\alpha, \beta) = ((-\operatorname{id})_\sharp \mu, \nu)$ proves the general claim. The statement for uniform distributions on the same number of points follows immediately by inserting the expressions for the relevant CDFs into the derived formula. In this case $\pi^\star$ and $\pi_\star$ are induced by the identity and anti-identity permutations, respectively as claimed.

In the case that $\mu = \sum_{i=1}^N a_i \delta_{x^{(i)}} \in \mathcal{P}(\mathbb{R})$ and $\nu = \sum_{j=1}^M b_j \delta_{y^{(j)}} \in \mathcal{P}(\mathbb{R})$ where $(a_i)_{i=1}^N, (b_j)_{j=1}^M$ are positive and sum to 1 it suffices to show that $\pi_{\mathrm{NW}}(\mu, \nu) = (F_\mu^{-1}, F_\nu^{-1})_\sharp(\mathrm{Unif}([0,1]))$. To this end, let $A_0 = 0$ and $A_i = \sum_{k=1}^i a_k$ for $k \leq N$, then $F_\mu^{-1}(u) = x^{(i)}$ if $u \in (A_{i-1}, A_i]$; the endpoints can be disregarded as they have null measure. Letting $(B_j)_{j=0}^M$ be defined similarly for ν,

$$\begin{aligned}(F_\mu^{-1}, F_\nu^{-1})_\sharp(\mathrm{Unif}([0,1]))(\{(x^{(i)}, y^{(j)})\}) &= \mathrm{Unif}([0,1])\left(((A_{i-1}, A_i] \cap (B_{j-1}, B_j])\right) \\ &= \max\left\{\min\{A_i, B_j\} - \max\{A_{i-1}, B_{j-1}\}, 0\right\},\end{aligned}$$

for each $i \in \{1, \ldots, N\}$ and $j \in \{1, \ldots, M\}$.

We now show that $\pi_{\mathrm{NW}}(\mu, \nu)(\{(x^{(i)}, y^{(j)})\})$ admits the same formula. At the start of a given iteration of the loop where (i, j) is active, we have either just incremented j (in which case all weight from $y^{(1)}, \ldots, y^{(j-1)}$ has been assigned, no mass from $y^{(j)}$ has been assigned and, in particular, $A_i \geq B_{j-1} \geq A_{i-1}$) or i has just been incremented (in which case the previous facts hold with x in place of y, i in place of j and A in place of B). In the former case, if we were to increment j again, b_j is less than or equal to a_i minus all of the weight from $x^{(i)}$ which has already been assigned. Thus, $A_i \geq B_j$, $B_{j-1} \geq A_{i-1}$, and so

$$\pi_{\mathrm{NW}}(\mu, \nu)(\{(x^{(i)}, y^{(j)})\}) = b_j = B_j - B_{j-1} = \max\{\min\{A_i, B_j\} - \max\{A_{i-1}, B_{j-1}\}, 0\}.$$

Still under the former case, if we were to increment i, we assign all previously unassigned mass from $x^{(i)}$ to $y^{(j)}$ (possibly 0 mass) which currently has no mass assigned to it and this unassigned mass is less than or equal to b_j. Since, after this operation, we have assigned a total mass of A_i from μ, this means that $A_i - B_{j-1}$ mass was assigned and, as a consequence,

$$\pi_{\mathrm{NW}}(\mu, \nu)(\{(x^{(i)}, y^{(j)})\}) = A_i - B_{j-1} = \max\{\min\{A_i, B_j\} - \max\{A_{i-1}, B_{j-1}\}, 0\}.$$

The case where we increment i in the previous iteration admits the same analysis by symmetry. If the pair (i, j) is never active throughout the iterations this means that all of the mass from $x^{(i)}$ has been assigned before the index j occurs or vice-versa and so the north-west corner rule coupling assigns no mass to this pair. In the former case $A_{i-1} < A_i \leq B_{j-1} < B_j$, whereby

$$\max\{\min\{A_i, B_j\} - \max\{A_{i-1}, B_{j-1}\}, 0\} = \max\{A_i - B_{j-1}, 0\} = 0.$$

The other case follows by symmetry. Finally, we account for the initial iteration for which this formula clearly holds since either $x^{(i)}$ assigns all of its mass to $y^{(j)}$ or the opposite holds. In sum, $\pi_{\mathrm{NW}}(\mu, \nu) = (F_\mu^{-1}, F_\nu^{-1})_\sharp(\mathrm{Unif}([0,1]))$.

Conclude from the first part of the proof that $\pi_{\mathrm{NW}}(\mu, \nu)$ or $(-\operatorname{id}, \operatorname{id})_\sharp \pi_{\mathrm{NW}}((-\operatorname{id})_\sharp \mu, \nu)$ is optimal. □

A.4 Proof of Proposition 4.2

As shown in Lemma A.5 ahead, for each fixed $\theta \in \mathbb{S}^{d_y - 1}$, the function $\boldsymbol{\Delta} \in \mathrm{St}(d_x, d_y) \mapsto \mathsf{IGW}((\theta^\intercal \boldsymbol{\Delta})_\sharp \mu, (\theta^\intercal)_\sharp \nu)^2$ is Lipschitz continuous relative to $\|\cdot\|_{\mathrm{F}}$ with constant at most $L'_{\mu,\nu} = 4M_2(\mu)^2 + 4M_2(\mu)M_2(\nu)$. It follows that, for any $\boldsymbol{\Delta}, \boldsymbol{\Delta}' \in \mathrm{St}(d_x, d_y)$,

$$\begin{aligned}&\left| \int_{\mathbb{S}^{d_y - 1}} \mathsf{IGW}((\theta^\intercal \boldsymbol{\Delta})_\sharp \mu, (\theta^\intercal)_\sharp \nu)^2 - \mathsf{IGW}((\theta^\intercal \boldsymbol{\Delta}')_\sharp \mu, (\theta^\intercal)_\sharp \nu)^2 d\sigma_{d_y}(\theta) \right| \\ &\qquad \leq L'_{\mu,\nu} \int_{\mathbb{S}^{d_y - 1}} \|\boldsymbol{\Delta} - \boldsymbol{\Delta}'\|_{\mathrm{F}} d\sigma_{d_y}(\theta) = L'_{\mu,\nu} \|\boldsymbol{\Delta} - \boldsymbol{\Delta}'\|_{\mathrm{F}},\end{aligned}$$

so that $\boldsymbol{\Delta} \in \mathrm{St}(d_x, d_y) \mapsto \int_{\mathbb{S}^{d_y-1}} \mathsf{IGW}((\theta^\intercal \boldsymbol{\Delta})_\sharp \mu, (\theta^\intercal)_\sharp \nu)^2 d\sigma_{d_y}(\theta)$ is Lipschitz continuous with the same constant. As $\mathrm{St}(d_x, d_y)$ is compact relative to $\|\cdot\|_{\mathrm{F}}$, each continuous function over $\mathrm{St}(d_x, d_y)$ achieves its minimum, proving the claim. □

A.5 Proof of Proposition 4.4

The proof of this result relies upon the following technical lemma, whose proof is provided in Appendix B.4 ahead.

Lemma A.4. *For any $\mu \in \mathcal{P}_2(\mathbb{R}^{d_x}), \nu \in \mathcal{P}_2(\mathbb{R}^{d_y})$, and $\boldsymbol{\Delta} \in \mathrm{St}(d_x, d_y)$, the functions $a \in \mathbb{R}^{d_y} \mapsto \mathsf{W}_2((a^\intercal \boldsymbol{\Delta})_\sharp \mu, (a^\intercal)_\sharp \nu)$ and $a \in \mathbb{R}^{d_y} \mapsto \mathsf{W}_2((-\mathrm{id})_\sharp (a^\intercal \boldsymbol{\Delta})_\sharp \mu, (a^\intercal)_\sharp \nu)$ are continuous.*

We first prove the pseudometric property. For any $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$, Proposition 4.2 implies that $\underline{\mathsf{IGW}}(\mu, \nu)^2 = \int \mathsf{IGW}((\theta^\intercal \boldsymbol{\Delta})_\sharp \mu, (\theta^\intercal)_\sharp \nu)^2 d\sigma_d(\theta)$ for some choice of $\boldsymbol{\Delta} \in \mathrm{St}(d, d)$. Since $\mathsf{IGW}((\theta^\intercal \boldsymbol{\Delta})_\sharp \mu, (\theta^\intercal)_\sharp \nu)^2 \geq 0$ for every $\theta \in \mathbb{S}^{d-1}$, $\underline{\mathsf{IGW}}(\mu, \nu) \geq 0$. If $\mu = \nu$, $0 \leq \underline{\mathsf{IGW}}(\mu, \nu)^2 \leq \int \mathsf{IGW}((\theta^\intercal)_\sharp \mu, (\theta^\intercal)_\sharp \mu)^2 d\sigma_d(\theta) = 0$ so that $\underline{\mathsf{IGW}}(\mu, \nu) = 0$. To show symmetry, notice that, for any $\boldsymbol{\Delta} \in \mathrm{St}(d, d)$, after the change of variables $\vartheta = \boldsymbol{\Delta}^\intercal \theta$,

$$\int \mathsf{IGW}((\theta^\intercal \boldsymbol{\Delta})_\sharp \mu, (\theta^\intercal)_\sharp \nu)^2 d\sigma_d(\theta) = \int \mathsf{IGW}((\vartheta^\intercal)_\sharp \mu, ((\boldsymbol{\Delta}\vartheta)^\intercal)_\sharp \nu)^2 d\sigma_d(\vartheta).$$

Take the infimum over $\boldsymbol{\Delta}$ on both sides, we get

$$\begin{aligned}
\underline{\mathsf{IGW}}(\mu, \nu)^2 &= \inf_{\Delta \in \mathrm{St}(d,d)} \int \mathsf{IGW}((\vartheta^\intercal)_\sharp \mu, ((\boldsymbol{\Delta}\vartheta)^\intercal)_\sharp \nu)^2 d\sigma_d(\vartheta) \\
&= \inf_{\Delta \in \mathrm{St}(d,d)} \int \mathsf{IGW}(((\boldsymbol{\Delta}\vartheta)^\intercal)_\sharp \nu, (\vartheta^\intercal)_\sharp \mu)^2 d\sigma_d(\vartheta) = \underline{\mathsf{IGW}}(\nu, \mu)^2.
\end{aligned}$$

The last equality follows by noting that $\mathrm{St}(d, d)$ is the set of all orthonormal matrices whereby $\{\boldsymbol{\Delta} : \boldsymbol{\Delta} \in \mathrm{St}(d, d)\} = \{\boldsymbol{\Delta}^\intercal : \boldsymbol{\Delta} \in \mathrm{St}(d, d)\}$.

As for the triangle inequality, if $\mu, \nu, \rho \in \mathcal{P}_2(\mathbb{R}^d)$, Minkowski's inequality in $L^2(\mathbb{S}^{d-1}, \sigma_d)$ and the triangle inequality for IGW yield that

$$\begin{aligned}
\underline{\mathsf{IGW}}(\mu, \nu) &= \min_{\boldsymbol{\Delta} \in \mathrm{St}(d,d)} \left(\int \mathsf{IGW}((\theta^\intercal \boldsymbol{\Delta})_\sharp \mu, (\theta^\intercal)_\sharp \nu)^2 d\sigma_d(\theta) \right)^{1/2} \\
&\leq \min_{\boldsymbol{\Delta} \in \mathrm{St}(d,d)} \left(\int \left(\mathsf{IGW}((\theta^\intercal \boldsymbol{\Delta})_\sharp \mu, (\theta^\intercal \boldsymbol{\Gamma})_\sharp \rho) + \mathsf{IGW}((\theta^\intercal \boldsymbol{\Gamma})_\sharp \rho, (\theta^\intercal)_\sharp \nu) \right)^2 d\sigma_d(\theta) \right)^{1/2} \\
&\leq \min_{\boldsymbol{\Delta} \in \mathrm{St}(d,d)} \left(\int \mathsf{IGW}((\theta^\intercal \boldsymbol{\Delta})_\sharp \mu, (\theta^\intercal \boldsymbol{\Gamma})_\sharp \rho)^2 d\sigma_d(\theta) \right)^{1/2} \\
&\qquad + \left(\int \mathsf{IGW}((\theta^\intercal \boldsymbol{\Gamma})_\sharp \rho, (\theta^\intercal)_\sharp \nu)^2 d\sigma_d(\theta) \right)^{1/2},
\end{aligned}$$

where $\boldsymbol{\Gamma} \in \mathrm{St}(d, d)$ is arbitrary. The first term satisfies

$$\min_{\boldsymbol{\Delta} \in \mathrm{St}(d,d)} \int \mathsf{IGW}((\theta^\intercal \boldsymbol{\Delta})_\sharp \mu, (\theta^\intercal \boldsymbol{\Gamma})_\sharp \rho)^2 d\sigma_d(\theta) = \underline{\mathsf{IGW}}(\mu, \boldsymbol{\Gamma}_\sharp \rho)^2 = \underline{\mathsf{IGW}}(\mu, \rho)^2,$$

where we recall that $\underline{\mathsf{IGW}}$ is invariant to orthogonal transformations. We then tighten the upper bound in the previous display by minimizing over the choice of $\boldsymbol{\Gamma}$;

$$\begin{aligned}
\underline{\mathsf{IGW}}(\mu, \nu) &\leq \underline{\mathsf{IGW}}(\mu, \rho) + \min_{\boldsymbol{\Gamma} \in \mathrm{St}(d,d)} \left(\int \mathsf{IGW}((\theta^\intercal \boldsymbol{\Gamma})_\sharp \rho, (\theta^\intercal)_\sharp \nu)^2 d\sigma_d(\theta) \right)^{1/2} \\
&= \underline{\mathsf{IGW}}(\mu, \rho) + \underline{\mathsf{IGW}}(\rho, \nu),
\end{aligned}$$

as desired. With that, we move to establish Items (i) and (ii).

Item (ii): Equivalence under analyticity condition. We show that when μ, ν have analytic characteristic functions, $\underline{\mathsf{IGW}}(\mu, \nu) = 0$ if and only if $\mathsf{IGW}(\mu, \nu) = 0$. Suppose first that $\underline{\mathsf{IGW}}(\mu, \nu) = 0$. By Proposition 4.2, there exists $\boldsymbol{\Delta} \in \mathrm{St}(d, d)$ for which $\mathsf{IGW}((\theta^\intercal \boldsymbol{\Delta})_\sharp \mu, (\theta^\intercal)_\sharp \nu) = 0$ for every $\theta \in \mathcal{O}$

where $\mathcal{O} \subset \mathbb{S}^{d-1}$ has full σ_d measure. Thus, for all such $\theta \in \mathcal{O}$ there exists $\mathbf{D}_\theta \in \mathrm{St}(1,1) = \{-1, 1\}$ for which $\mu_\theta := (\mathbf{D}_\theta)_\sharp(\theta^\intercal \boldsymbol{\Delta})_\sharp \mu = (\theta^\intercal)_\sharp \nu =: \nu_\theta$.

Suppose first that there exists some $\theta \in \mathcal{O}$ at which $(-\mathrm{id})_\sharp(\theta^\intercal \boldsymbol{\Delta})_\sharp \mu \neq (\theta^\intercal)_\sharp \nu$. By Lemma A.4, there exists a neighborhood $N_\theta \subset \mathbb{R}^d$ of θ on which $\mathsf{W}_2((-\mathrm{id})_\sharp((\cdot)^\intercal \boldsymbol{\Delta})_\sharp \mu, ((\cdot)^\intercal)_\sharp \nu) > 0$ so that this inequality holds, in particular, at every $\vartheta \in N_\theta \cap \mathcal{O}$ which has positive measure relative to σ_d. Conclude that $\mathsf{W}_2((\vartheta^\intercal \boldsymbol{\Delta})_\sharp \mu, (\vartheta^\intercal)_\sharp \nu) = 0$ every $\vartheta \in N_\theta \cap \mathcal{O}$ and, by continuity, on the entire set $\mathbb{S}^{d-1} \cap N_\theta$ so that, for every $\vartheta \in \mathbb{S}^{d-1} \cap N_\theta$, $\Phi_{(\vartheta^\intercal \boldsymbol{\Delta})_\sharp \mu}(t) = \Phi_{(\vartheta^\intercal)_\sharp \nu}(t)$ for each $t \in \mathbb{R}$, i.e.,

$$\int e^{its} d(\vartheta^\intercal \boldsymbol{\Delta})_\sharp \mu(s) = \int e^{i(t\vartheta)^\intercal x} d\boldsymbol{\Delta}_\sharp \mu(x) = \int e^{i(t\vartheta)^\intercal y} d\nu(y).$$

In sum, $\Phi_{\boldsymbol{\Delta}_\sharp \mu}(t) = \Phi_\nu(t)$ for every $t \in \mathcal{T} := \{s\vartheta : s \in \mathbb{R} \text{ and } \vartheta \in \mathbb{S}^{d-1} \cap N_\theta\}$. It is easy to see that $\mathcal{T}$ contains an open ball centered at θ for instance, see Fig. 4 for an illustration. However, since $\Phi_{\boldsymbol{\Delta}_\sharp \mu}$ and Φ_ν are analytic and coincide on some open set, Corollary 2.3.8 in [61] guarantees that $\Phi_{\boldsymbol{\Delta}_\sharp \mu} = \Phi_\nu$ on $\mathbb{R}^d$ and so $\boldsymbol{\Delta}_\sharp \mu = \nu$.

By the same token, if there exists $\theta \in \mathcal{O}$ at which $(\theta^\intercal \boldsymbol{\Delta})_\sharp \mu \neq (\theta^\intercal)_\sharp \nu$, then $(-\boldsymbol{\Delta})_\sharp \mu = \nu$. Finally, if $(-\mathrm{id})_\sharp(\theta^\intercal \boldsymbol{\Delta})_\sharp \mu = (\theta^\intercal)_\sharp \nu$ and $(\theta^\intercal \boldsymbol{\Delta})_\sharp \mu = (\theta^\intercal)_\sharp \nu$ for every $\theta \in \mathcal{O}$, the squared sliced 2-Wasserstein distances,

$$\int \mathsf{W}_2((-\mathrm{id})_\sharp(\theta^\intercal \boldsymbol{\Delta})_\sharp \mu, (\theta^\intercal)_\sharp \nu)^2 d\sigma_d(\theta) \text{ and } \int \mathsf{W}_2((\theta^\intercal \boldsymbol{\Delta})_\sharp \mu, (\theta^\intercal)_\sharp \nu)^2 d\sigma_d(\theta)$$

are both 0 so that Proposition 2.2 in [85] implies that $(-\boldsymbol{\Delta})_\sharp \mu = \nu$ and $(\boldsymbol{\Delta})_\sharp \mu = \nu$.

For the other direction, if $\mathsf{IGW}(\mu, \nu) = 0$, there exists $\boldsymbol{\Delta} \in \mathrm{St}(d, d)$ for which $\boldsymbol{\Delta}_\sharp \mu = \nu$, so

$$\begin{aligned}\underline{\mathsf{IGW}}(\mu, \nu)^2 &\leq \int \mathsf{IGW}((\theta^\intercal \boldsymbol{\Delta})_\sharp \mu, (\theta^\intercal)_\sharp \nu)^2 d\sigma_d(\theta) \\ &\leq \int 2\left(M_2((\theta^\intercal \boldsymbol{\Delta})_\sharp \mu) + M_2((\theta^\intercal)_\sharp \nu)\right) \mathsf{W}_2((\theta^\intercal \boldsymbol{\Delta})_\sharp \mu, (\theta^\intercal)_\sharp \nu)^2 d\sigma_d(\theta),\end{aligned}$$

where the upper bound is due to (3). Since $\boldsymbol{\Delta}_\sharp \mu = \nu$, the right hand side is 0 by Proposition 2.2 in [85] whereby $\underline{\mathsf{IGW}}(\mu, \nu) = 0$ proving the claim.

Item (i): Necessity of analyticity. We show by a direct construction that there exist measures ρ_1, ρ_2 with $\underline{\mathsf{IGW}}(\rho_1, \rho_2) = 0$ but $\mathsf{IGW}(\rho_1, \rho_2) > 0$.

Let $\kappa \in \mathcal{P}_2(\mathbb{R}^d)$ $(d > 2)$ be the base measure with density $g(x) = \frac{K}{(1+\|x\|^2)^m}$, where K is the normalizing constant. We take $m > \frac{d+2}{2}$ so that κ has bounded second moment. Now, let

$$\chi(x) = \begin{cases} \exp(-\frac{1}{1-\|x\|^2}), & \text{if } \|x\| < 1, \\ 0, & \text{otherwise.} \end{cases}$$

Notice that $\chi(\frac{x-z}{r})$ is a smooth compactly supported function on $B(z, r) := \{x \in \mathbb{R}^d : \|x - z\| \leq r\}$. Let $v_j = 3je_1 + e_j$ for $j = 1, \ldots, d$ where the e_j are the standard unit vectors and set $\eta_1(x) = \sum_{j=1}^d \left(\chi\left(\frac{x-v_j}{1/10}\right) - \chi\left(\frac{x+v_j}{1/10}\right)\right)$ and $\eta_2(x) = \chi\left(\frac{x-e_2}{1/10}\right) - \chi\left(\frac{x+e_2}{1/10}\right)$. Note that η_1 is supported in $\bigcup_{j=1}^d (B(v_j, 1/10) \cup B(-v_j, 1/10))$, which is a subset of $C_1 := x \in \mathbb{R}^d : |x_1| \geq |x_2|$. While η_2 is supported in $B(e_2, 1/10) \cup B(-e_2, 1/10)$, which is a subset of $C_2 := x \in \mathbb{R}^d : |x_2| \geq |x_1|$ so that η_1 and η_2 have disjoint support and are both odd functions. Furthermore, η_1 and η_2 are Schwartz functions, that is,

$$\sup_{x \in \mathbb{R}^d} \left|x^\beta \partial^\alpha \eta_i(x)\right| < \infty \quad \text{for } i = 1, 2 \text{ and every multi-index } \alpha, \beta,$$

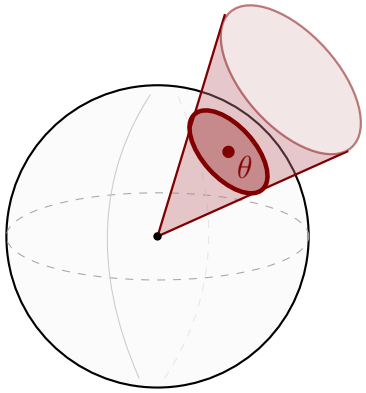


Figure 4: Illustration of the proof of Item (ii) in Proposition 4.4. For any $\theta \in \mathbb{S}^{d-1}$ where the identity mapping is strictly optimal, there exists a spherical cap around θ on which the same mapping remains optimal. This optimality extends to the entire cone extending from the origin through that cap, via a scaling argument. Since the characteristic functions of the projected measures coincide along all rays in that cone, they also coincide on the cone itself. Analyticity then forces the characteristic functions to coincide everywhere, hence $\mathsf{IGW}(\mu, \nu) = 0$.

As the Fourier transform defines an isomorphism on the set of all Schwartz functions, η_1 and η_2 have inverse Fourier transforms, $\check{\eta}_1, \check{\eta}_2$, which are also Schwartz functions and, for $j = 1, 2$,

$$\check{\eta_j}(x) = (2\pi)^{-d} i \int_{\mathbb{R}^d} \sin(t^\intercal x)\eta_j(t)dt,$$

where the final equality follows from the fact that η_j is odd. Conclude that $h_j \coloneqq i\check{\eta}_j$ is a real-valued odd Schwartz function. Conclude that h_1 and h_2 integrate to 0 and satisfy $|h_j(x)| \leq M(1+|x|^2)^{-m}$ for some choice of $M > 0$. Defining f_1, f_2 via

$$f_1(x) \coloneqq g(x) + \frac{K}{2M}h_1(x) + \frac{K}{2M}h_2(x) \geq 0, \quad f_2(x) \coloneqq g(x) + \frac{K}{2M}h_1(x) - \frac{K}{2M}h_2(x) \geq 0,$$

we see that f_1, f_2 integrate to 1 so that they are probability density functions for some distributions which we denote by ρ_1 and ρ_2. Their characteristic functions are

$$\Phi_{\rho_1}(t) = \hat{g}(t) + \frac{iK}{2M}\left(\eta_1(t) + \eta_2(t)\right), \quad \Phi_{\rho_2}(t) = \hat{g}(t) + \frac{iK}{2M}\left(\eta_1(t) - \eta_2(t)\right).$$

For any $t \in \mathrm{Int}(C_1)$, the interior of C_1,

$$\Phi_{\rho_1}(t) = \hat{g}(t) + \frac{iK}{2M}\eta_1(t) = \Phi_{\rho_2}(t).$$

Whereas if $t \in \mathrm{Int}(C_2)$, $\Phi_{\rho_1}(t) = \hat{g}(t) + \frac{iK}{2M}\eta_2(t)$, which coincides with the complex conjugate, $\overline{\Phi_{\rho_2}}(t)$, of $\Phi_{\rho_2}(t)$ since $\hat{g}$ is real-valued as g is even and η_1, η_2 are also real-valued. We conclude that, if $\theta \in \mathbb{S}^{d-1}$ is such that $\theta \in \mathrm{Int}(C_1)$, then $s\theta : s \in \mathbb{R} \subset \mathrm{Int}(C_1)$ and $\Phi_{\rho_1}(s\theta) = \Phi_{\rho_2}(s\theta)$ for all $s \in \mathbb{R}$, that is, $\theta^\intercal_\sharp \rho_1 = \theta^\intercal_\sharp \rho_2$. If $\theta \in \mathrm{Int}(C_2)$, we obtain similarly that $\theta^\intercal_\sharp \rho_1 = \theta^\intercal_\sharp(-\mathrm{id})_\sharp \rho_2 = (-\mathrm{id})_\sharp \theta^\intercal_\sharp \rho_2$. This means that $\mathsf{IGW}(\theta^\intercal_\sharp \rho_1, \theta^\intercal_\sharp \rho_2) = 0$ for σ_d-almost every θ and so $\underline{\mathsf{IGW}}(\rho_1, \rho_2) = 0$, noting that $C_1 \cup C_2 = \mathbb{R}^d$. We next show that $\mathsf{IGW}(\rho_1, \rho_2) \neq 0$, that is, there does not exist some $\mathbf{\Delta} \in \mathrm{St}(d, d)$ with $\mathbf{\Delta}_\sharp{}^\intercal \rho_1 = \rho_2$. Assume to the contrary that such a $\mathbf{\Delta}$ exists so that $\Phi_{\mathbf{\Delta}_\sharp{}^\intercal \rho_1}(t) = \Phi_{\rho_1}(\mathbf{\Delta}t) = \Phi_{\rho_2}(t)$ for every $t \in \mathbb{R}^d$. Observe that

$$\Phi_{\rho_1}(\mathbf{\Delta}t) = \hat{g}(t) + \frac{iK}{2M}\left(\eta_1(\mathbf{\Delta}t) + \eta_2(\mathbf{\Delta}t)\right)$$

noting that $\mathbf{\Delta}^\intercal\mathbf{\Delta} = \mathrm{id}$ and that g is invariant to orthogonal transformations. It follows that $\Phi_{\mathbf{\Delta}_\sharp{}^\intercal \rho_1}(t) = \Phi_{\rho_2}(t)$ if and only if

$$\eta_1(\mathbf{\Delta}t) + \eta_2(\mathbf{\Delta}t) = \eta_1(t) - \eta_2(t) \quad \text{for all } t \in \mathbb{R}^d.$$

We now illustrate that no such $\mathbf{\Delta}$ exists. Fix $j \in 1, \ldots, d$ and evaluate at $t = v_j$, since v_j lies outside the support of η_2, we have

$$\eta_1(\mathbf{\Delta}v_j) + \eta_2(\mathbf{\Delta}v_j) = \eta_1(v_j) = e^{-1}.$$

Because the supports of the bumps are disjoint and χ attains its maximum value e^{-1} only at the origin, the left-hand side attains e^{-1} only at the bump centers. i.e. $\mathbf{\Delta}v_j \in v_1, \ldots, v_d, e_2$. Since $\mathbf{\Delta}$ is orthogonal, $|\mathbf{\Delta}v_j| = |v_j|$ and $v_1, \ldots, v_d, e_2$ all have distinct length, thus $\mathbf{\Delta}v_j = v_j$. This is true for all $j = 1, \ldots, d$. Since $v_1, \ldots, v_d$ form a basis for $\mathbb{R}^d$, $\mathbf{\Delta} = \mathrm{id}$, which cannot hold as the two measures have different characteristic functions. Therefore, $\mathsf{IGW}(\rho_1, \rho_2) > 0$, proving the claim. □

A.6 Proof of Proposition 4.5

Suppose first that $\mathsf{IGW}(\mu_n, \mu) \to 0$ as $n \to \infty$. By Proposition 2.3, there exists $\mathbf{\Delta}_n \in \mathrm{St}(d, d)$ for which $(\mathbf{\Delta}_n^\intercal)_\sharp \mu_n \xrightarrow{w} \mu$ and $M_2(\mu_n) \to M_2(\mu)$. By Lemma A.4, for any $\theta \in \mathbb{S}^{d-1}$, $(\theta^\intercal \mathbf{\Delta}_n)_\sharp \mu_n \xrightarrow{w} (\theta^\intercal)_\sharp \mu$ and $M_2((\theta^\intercal)_\sharp \mu_n) \to M_2((\theta^\intercal)_\sharp \mu)$ so that $\mathsf{IGW}((\theta^\intercal \mathbf{\Delta}_n)_\sharp \mu_n, (\theta^\intercal)_\sharp \mu) \to 0$. Observe now that

$$0 \leq \mathsf{IGW}((\theta^\intercal \mathbf{\Delta}_n)_\sharp \mu_n, (\theta^\intercal)_\sharp \mu)^2 \leq (\theta^\intercal \mathbf{\Delta}_n \mathbf{R}_{\mu_n} \mathbf{\Delta}_n^\intercal \theta)^2 + (\theta^\intercal \mathbf{R}_\mu \theta)^2 \leq M_2(\mu_n)^4 + M_2(\mu)^4,$$

where we have used the Cauchy-Schwarz inequality to obtain that $|\theta^\intercal \mathbf{\Delta}_n \int zz^\intercal d\mu_n(z) \mathbf{\Delta}_n^\intercal \theta| \leq \int \|z\|^2 d\mu_n(z) \|\mathbf{\Delta}_n^\intercal \theta\|^2 = M_2(\mu_n)^2$ since $\mathrm{St}(d, d)$ is the set of $d \times d$ orthogonal matrices. Since

$M_2(\mu_n) \to M_2(\mu)$, $\mathsf{IGW}((\theta^\intercal \boldsymbol{\Delta}_n)_\sharp \mu_n, (\theta^\intercal)_\sharp \mu)^2$ is dominated by a constant so that we may apply the dominated convergence theorem to obtain that

$$0 \leq \underline{\mathsf{IGW}}(\mu_n, \mu)^2 \leq \int \mathsf{IGW}((\theta^\intercal \boldsymbol{\Delta}_n)_\sharp \mu_n, (\theta^\intercal)_\sharp \mu)^2 d\sigma_d(\theta) \to 0,$$

proving that $\underline{\mathsf{IGW}}(\mu_n, \mu) \to 0$.

On the other hand, if $\underline{\mathsf{IGW}}(\mu_n, \mu) \to 0$ there exists a sequence $(\boldsymbol{\Delta}_n)_{n\in\mathbb{N}} \subset \mathrm{St}(d,d)$ along which $\int \mathsf{IGW}((\theta^\intercal \boldsymbol{\Delta}_n)_\sharp \mu_n, (\theta^\intercal)_\sharp \mu)^2 d\sigma_d(\theta) \to 0$. Consequently, there exists a set A of full σ_d measure for which, after extracting subsequences, $\mathsf{IGW}((\theta^\intercal \boldsymbol{\Delta}_n)_\sharp \mu_n, (\theta^\intercal)_\sharp \mu)^2 \to 0$ for every $\theta \in A$. (To simplify notation, we relabel the subsequence and continue to denote it by the same index)

Now, fix some $\theta \in A$. From Corollary A.3, $\mathsf{W}_2\left((s_n^\theta \,\mathrm{id})_\sharp (\theta^\intercal \boldsymbol{\Delta}_n)_\sharp \mu_n, (\theta^\intercal)_\sharp \mu\right) \to 0$ where

$$s_n^\theta = \begin{cases} 1, & \text{if } \mathsf{W}_2((\theta^\intercal \boldsymbol{\Delta}_n)_\sharp \mu_n, (\theta^\intercal)_\sharp \mu) \leq \mathsf{W}_2((-\mathrm{id})_\sharp (\theta^\intercal \boldsymbol{\Delta}_n)_\sharp \mu_n, (\theta^\intercal)_\sharp \mu), \\ -1, & \text{if } \mathsf{W}_2((\theta^\intercal \boldsymbol{\Delta}_n)_\sharp \mu_n, (\theta^\intercal)_\sharp \mu) > \mathsf{W}_2((-\mathrm{id})_\sharp (\theta^\intercal \boldsymbol{\Delta}_n)_\sharp \mu_n, (\theta^\intercal)_\sharp \mu). \end{cases}$$

Assume that θ is such that $\mathsf{W}_2((\theta^\intercal)_\sharp \mu, (-\mathrm{id})_\sharp (\theta^\intercal)_\sharp \mu) \neq 0$ (we leave the case where it nullifies to the end of the proof). We show that there exists some neighborhood of θ on which $s_n^\vartheta = s_n^\theta$ for all sufficiently large n. We know that

$$\begin{aligned} &\liminf_{n\to\infty} \mathsf{W}_2\left((s_n^\theta \,\mathrm{id})_\sharp (\theta^\intercal \boldsymbol{\Delta}_n)_\sharp \mu_n, (-\mathrm{id})_\sharp (\theta^\intercal)_\sharp \mu\right) \\ &\geq \mathsf{W}_2((\theta^\intercal)_\sharp \mu, (-\mathrm{id})_\sharp (\theta^\intercal)_\sharp \mu) - \lim_{n\to\infty} \mathsf{W}_2\left((s_n^\theta \,\mathrm{id})_\sharp (\theta^\intercal \boldsymbol{\Delta}_n)_\sharp \mu_n, (\theta^\intercal)_\sharp \mu\right) \\ &= \mathsf{W}_2\left((\theta^\intercal)_\sharp \mu, (-\mathrm{id})_\sharp (\theta^\intercal)_\sharp \mu\right). \end{aligned}$$

Hence, for all n sufficiently large, we have that

$$\mathsf{W}_2\left((s_n^\theta \,\mathrm{id})_\sharp (\theta^\intercal \boldsymbol{\Delta}_n)_\sharp \mu_n, (-\mathrm{id})_\sharp (\theta^\intercal)_\sharp \mu\right) > \frac{3}{4} \mathsf{W}_2\left((\theta^\intercal)_\sharp \mu, (-\mathrm{id})_\sharp (\theta^\intercal)_\sharp \mu\right) \text{ and}$$
$$\mathsf{W}_2\left((s_n^\theta \,\mathrm{id})_\sharp (\theta^\intercal \boldsymbol{\Delta}_n)_\sharp \mu_n, (\theta^\intercal)_\sharp \mu\right) < \frac{1}{4} \mathsf{W}_2\left((\theta^\intercal)_\sharp \mu, (-\mathrm{id})_\sharp (\theta^\intercal)_\sharp \mu\right).$$

Moreover, for all such n, we have

$$\begin{aligned} &\mathsf{W}_2\left((s_n^\theta \,\mathrm{id})_\sharp (\vartheta^\intercal \boldsymbol{\Delta}_n)_\sharp \mu_n, (\vartheta^\intercal)_\sharp \mu\right) \\ &\leq \mathsf{W}_2\left((s_n^\theta \,\mathrm{id})_\sharp (\theta^\intercal \boldsymbol{\Delta}_n)_\sharp \mu_n, (\theta^\intercal)_\sharp \mu\right) + \|\theta - \vartheta\| \left(M_2((s_n^\theta \,\mathrm{id})_\sharp (\boldsymbol{\Delta}_n^\intercal)_\sharp \mu_n)^{1/2} + M_2(\mu)^{1/2}\right) \\ &= \mathsf{W}_2\left((s_n^\theta \,\mathrm{id})_\sharp (\theta^\intercal \boldsymbol{\Delta}_n)_\sharp \mu_n, (\theta^\intercal)_\sharp \mu\right) + \|\theta - \vartheta\| \left(M_2(\mu_n)^{1/2} + M_2(\mu)^{1/2}\right) \\ &< \frac{1}{4} \mathsf{W}_2\left((\theta^\intercal)_\sharp \mu, (-\mathrm{id})_\sharp (\theta^\intercal)_\sharp \mu\right) + \|\theta - \vartheta\| \left(M_2(\mu_n)^{1/2} + M_2(\mu)^{1/2}\right), \end{aligned}$$

and, likewise,

$$\begin{aligned} &\mathsf{W}_2\left((s_n^\theta \,\mathrm{id})_\sharp (\vartheta^\intercal \boldsymbol{\Delta}_n)_\sharp \mu_n, (-\mathrm{id})_\sharp (\vartheta^\intercal)_\sharp \mu\right) \\ &\qquad > \frac{3}{4} \mathsf{W}_2\left((\theta^\intercal)_\sharp \mu, (-\mathrm{id})_\sharp (\theta^\intercal)_\sharp \mu\right) - \|\theta - \vartheta\| \left(M_2(\mu_n)^{1/2} + M_2(\mu)^{1/2}\right), \end{aligned}$$

as follows from Proposition 2.2 in [85]. We show later in the proof that $M_2(\mu_n)$ is uniformly bounded so that, for all sufficiently large n, there is a neighborhood N_θ of θ on which

$$\mathsf{W}_2\left((s_n^\theta \,\mathrm{id})_\sharp (\vartheta^\intercal \boldsymbol{\Delta}_n)_\sharp \mu_n, (-\mathrm{id})_\sharp (\vartheta^\intercal)_\sharp \mu\right) > \mathsf{W}_2\left((s_n^\theta \,\mathrm{id})_\sharp (\vartheta^\intercal \boldsymbol{\Delta}_n)_\sharp \mu_n, (\vartheta^\intercal)_\sharp \mu\right) \text{ for every } \vartheta \in N_\theta.$$

By applying Lemma B.1, we obtain that

$$\begin{aligned} \mathsf{W}_2\left((s_n^\theta \,\mathrm{id})_\sharp (\vartheta^\intercal \boldsymbol{\Delta}_n)_\sharp \mu_n, (-\mathrm{id})_\sharp (\vartheta^\intercal)_\sharp \mu\right)^2 &= \inf_{\pi \in \Pi(\mu_n, \mu)} \int \|s_n^\theta \vartheta^\intercal \boldsymbol{\Delta}_n x - (-\vartheta^\intercal y)\|^2 d\pi(x,y) \\ &= \inf_{\pi \in \Pi(\mu_n, \mu)} \int \|(-s_n^\theta \vartheta^\intercal \boldsymbol{\Delta}_n x) - \vartheta^\intercal y\|^2 d\pi(x,y) \\ &= \mathsf{W}_2\left((-s_n^\theta \,\mathrm{id})_\sharp (\vartheta^\intercal \boldsymbol{\Delta}_n)_\sharp \mu_n, (\vartheta^\intercal)_\sharp \mu\right) \end{aligned}$$

so that s_n^{ϑ} agrees with s_n^{θ} at all $\vartheta \in N_\theta$ for all such n sufficiently large. Therefore,

$$\mathsf{W}_2((s_n^{\theta}\,\mathrm{id})_\sharp(\vartheta^\intercal \boldsymbol{\Delta}_n)_\sharp\mu_n, (\vartheta^\intercal)_\sharp\mu) \to 0 \text{ for each } \vartheta \in N_\theta \cap A. \tag{10}$$

It remains to show that $(s_n^{\theta}\,\mathrm{id})_\sharp(\boldsymbol{\Delta}_n)_\sharp\mu_n$ admits a limit in W_2 so that we can apply the reasoning from the proof of Proposition 4.4.

We first show that the operator norms of $\mathbf{R}_n \coloneqq \int zz^\intercal d(\boldsymbol{\Delta}_n)_\sharp\mu_n(z)$ are uniformly bounded. Suppose to the contrary that $\|\mathbf{R}_n\|_{\mathrm{op}}$ is unbounded and let $\lambda_1^n, \ldots, \lambda_d^n$ denote the eigenvalues of $\mathbf{R}_n$ ordered from largest to smallest and $v_1^n, \ldots, v_d^n$ be the corresponding (orthonormal) eigenvectors noting that $\mathbf{R}_n$ is positive semidefinite. As $\|\mathbf{R}_n\|_{\mathrm{op}}$ is unbounded, there exists a subsequence n' along which $\|\mathbf{R}_{n'}\|_{\mathrm{op}} \to \infty$. As $\mathbb{S}^{d-1}$ is compact, $v_1^{n''} \to v \in \mathbb{S}^{d-1}$ along a further subsequence n''. Now, let $S \coloneqq \left\{\theta \in \mathbb{S}^{d-1} : |\theta^\intercal v|^2 > \delta\right\}$ for some $1 > \delta > 0$. As S has positive surface measure, there exists $\theta \in S \cap A$ for which

$$\theta^\intercal \mathbf{R}_{n''}\theta = \sum_{i=1}^d \lambda_i^{n''} \left|\theta^\intercal v_i^{n''}\right|^2 \geq \lambda_1^{n''} \left|\theta^\intercal v_1^{n''}\right|^2.$$

Since $v_1^{n''} \to v$, $\left|\theta^\intercal v_1^{n''}\right|^2 > 0$ for all n'' sufficiently large and, as $\lambda_1^{n''} = \|\mathbf{R}_{n''}\|_{\mathrm{op}}$ it follows that $\theta^\intercal \mathbf{R}_{n''}\theta \to \infty$ as $n'' \to \infty$. However, since $\mathsf{IGW}((\theta^\intercal\boldsymbol{\Delta}_n)_\sharp\mu_n, (\theta^\intercal)_\sharp\mu) \to 0$ for all $\theta \in A$, $M_2((\theta^\intercal\boldsymbol{\Delta}_n)_\sharp\mu_n) = \theta^\intercal\mathbf{R}_n\theta \to \theta^\intercal \int zz^\intercal d\mu(z)\theta$ contradicting the previous line. As such, we must have that $\|\mathbf{R}_n\|_{\mathrm{op}}$ is uniformly bounded.

By Markov's inequality, $\mathbb{P}_{(s_n^{\theta}\,\mathrm{id})_\sharp(\boldsymbol{\Delta}_n)_\sharp\mu_n}(\|X\| > t) \leq \frac{M_2((s_n^{\theta}\,\mathrm{id})_\sharp(\boldsymbol{\Delta}_n)_\sharp\mu_n)}{t^2} \leq \frac{d\|\mathbf{R}_n\|_{\mathrm{op}}}{t^2}$ which can be made arbitrarily small by taking t sufficiently large. Conclude that $((s_n^{\theta}\,\mathrm{id})_\sharp(\boldsymbol{\Delta}_n)_\sharp\mu_n)_{n\in\mathbb{N}}$ is a tight family and hence for any subsequence n' of n there exists a further subsequence, which we denote by n'', along which $(s_{n''}^{\theta}\,\mathrm{id})_\sharp(\boldsymbol{\Delta}_{n''})_\sharp\mu_{n''} \xrightarrow{w} \gamma$ for some probability measure γ by Prokhorov's theorem. By assumption, γ admits an analytic characteristic function and, for any $\vartheta \in N_\theta \cap A$,

$$\Phi_{(s_{n''}^{\theta}\,\mathrm{id})_\sharp(\vartheta^\intercal\boldsymbol{\Delta}_{n''})_\sharp\mu_{n''}}(t) = \Phi_{(s_{n''}^{\theta}\,\mathrm{id})_\sharp(\boldsymbol{\Delta}_{n''})_\sharp\mu_{n''}}(t\vartheta) \to \Phi_\gamma(t\vartheta) \text{ and } \Phi_{(s_{n''}^{\theta}\,\mathrm{id})_\sharp(\vartheta^\intercal\boldsymbol{\Delta}_{n''})_\sharp\mu_{n''}}(t) \to \Phi_\mu(t\vartheta),$$

where the second limit follows from (10). Conclude that, for each $t \in \mathbb{R}$ and $\vartheta \in N_\theta \cap A$, $\Phi_\gamma(t\vartheta) = \Phi_\mu(t\vartheta)$. By continuity of the characteristic functions, this equality extends to $\vartheta \in N_\theta \cap \mathbb{S}^{d-1}$ so that $\gamma = \mu$ by applying the same reasoning as in the proof of Proposition 4.4. With this, $(s_{n''}^{\theta}\,\mathrm{id})_\sharp(\boldsymbol{\Delta}_{n''})_\sharp\mu_{n''} \xrightarrow{w} \mu$. It remains, therefore, to prove that $M_2((\boldsymbol{\Delta}_{n''})_\sharp\mu_{n''}) \to M_2(\mu)$ to conclude that $\mathsf{IGW}(\mu_{n''}, \mu) \to 0$ (recall Proposition 2.3).

Since $(s_{n''}^{\theta}\,\mathrm{id})_\sharp(\boldsymbol{\Delta}_{n''})_\sharp\mu_{n''} \xrightarrow{w} \mu$, we have that $(s_{n''}^{\theta}\,\mathrm{id})_\sharp(\theta^\intercal\boldsymbol{\Delta}_{n''})_\sharp\mu_{n''} \xrightarrow{w} (\theta^\intercal)_\sharp\mu$ for every $\theta \in \mathbb{S}^{d-1}$. As $\mathsf{IGW}((\theta^\intercal\boldsymbol{\Delta}_n)_\sharp\mu_n, (\theta^\intercal)_\sharp\mu) \to 0$ for every $\theta \in A$, $M_2\left((s_n^{\theta}\,\mathrm{id})_\sharp(\theta^\intercal\boldsymbol{\Delta}_n)_\sharp\mu_n\right) = M_2\left((\theta^\intercal\boldsymbol{\Delta}_n)_\sharp\mu_n\right) \to M_2\left((\theta^\intercal)_\sharp\mu\right)$ so that

$$\int \mathsf{W}_2((s_{n''}^{\theta}\,\mathrm{id})_\sharp(\theta^\intercal\boldsymbol{\Delta}_{n''})_\sharp\mu_{n''}, (\theta^\intercal)_\sharp\mu)d\sigma_d(\theta) \to 0.$$

Theorem 2.1 in [85] asserts that $\mathsf{W}_2((s_{n''}^{\theta}\,\mathrm{id})_\sharp(\boldsymbol{\Delta}_{n''})_\sharp\mu_{n''}, \mu) \to 0$ so that $\mathsf{IGW}(\mu_{n''}, \mu) \to 0$. As this limit is independent of the choice of subsequence n'' it follows that $\mathsf{IGW}(\mu_{n''}, \mu) \to 0$.

Finally, assume that $\mathsf{W}_2((\theta^\intercal)_\sharp\mu, (-\,\mathrm{id})_\sharp(\theta^\intercal)_\sharp\mu) = 0$ for every $\theta \in A$. Fix $\theta \in A$ and let $s_n^{\theta} \in \{-1, 1\}$ be as in the first part of the proof so that $\mathsf{W}_2\left((s_n^{\theta}\,\mathrm{id})_\sharp(\theta^\intercal\boldsymbol{\Delta}_n)_\sharp\mu_n, (\theta^\intercal)_\sharp\mu\right) \to 0$. As

$$\mathsf{W}_2\left((s_n^{\theta}\,\mathrm{id})_\sharp(\theta^\intercal\boldsymbol{\Delta}_n)_\sharp\mu_n, (\theta^\intercal)_\sharp\mu\right) = \mathsf{W}_2\left((\theta^\intercal\boldsymbol{\Delta}_n)_\sharp\mu_n, (s_n^{\theta}\,\mathrm{id})_\sharp(\theta^\intercal)_\sharp\mu\right) = \mathsf{W}_2((\theta^\intercal\boldsymbol{\Delta}_n)_\sharp\mu_n, (\theta^\intercal)_\sharp\mu)\,,$$

$\mathsf{W}_2\left((\theta^\intercal\boldsymbol{\Delta}_n)_\sharp\mu_n, (\theta^\intercal)_\sharp\mu\right) \to 0$ for every $\theta \in A$. It follows, as in the previous case, that the sliced Wasserstein distance between $(\boldsymbol{\Delta}_n)_\sharp\mu_n$ and μ converges to 0 and hence that their unsliced Wasserstein distance also converges to 0, proving the claimed result. □

A.7 Proof of Proposition 4.6 and Corollary 4.7

Proof of Proposition 4.6. Applying (3), for each $\theta \in \mathbb{S}^{d_y-1}$ and $\boldsymbol{\Delta} \in \mathrm{St}(d_x, d_y)$,

$$\mathsf{IGW}((\theta^\intercal\boldsymbol{\Delta})_\sharp\mu, (\theta^\intercal)_\sharp\nu)^2 \leq 2\left(M_2((\theta^\intercal\boldsymbol{\Delta})_\sharp\mu) + M_2((\theta^\intercal)_\sharp\nu)\right)\mathsf{W}_2((\theta^\intercal\boldsymbol{\Delta})_\sharp\mu, (\theta^\intercal)_\sharp\nu))^2. \tag{11}$$

Setting $\mathbf{\Delta} = [\mathbf{I}_{d_x}\ \mathbf{0}]^\intercal$, we have that

$$M_2((\theta^\intercal \mathbf{\Delta})_\sharp \mu) = \int (\theta^\intercal \mathbf{\Delta} z)^2 d\mu(z) = \int \left(\sum_{i=1}^{d_x} \theta_i z_i \right)^2 d\mu(z) \leq \|\theta\|^2 \int \|z\|^2 d\mu(z) = M_2(\mu).$$

Similarly, it holds that $M_2((\theta)_\sharp^\intercal \nu) \leq M_2(\nu)$ so that integrating both sides of (11), it holds that

$$\underline{\mathsf{IGW}}(\mu,\nu)^2 \leq \int \mathsf{IGW}((\theta^\intercal \mathbf{\Delta})_\sharp \mu, (\theta^\intercal)_\sharp \nu)^2 d\sigma_{d_y}(\theta) \leq 2\left(M_2(\mu) + M_2(\nu)\right) \underline{\mathsf{W}}_2((\mathbf{\Delta})_\sharp \mu, \nu)^2,$$

which proves the claimed result. □

Proof of Corollary 4.7. We have from Proposition 4.6 that

$$\underline{\mathsf{IGW}}(\hat{\mu}_n, \mu) \leq \sqrt{2}\sqrt{M_2(\hat{\mu}_n) + M_2(\mu)}\, \underline{\mathsf{W}}_2(\hat{\mu}_n, \mu),$$

noting that $\hat{\mu}_n$ and μ are distributions on the same Euclidean space and hence the particular alignment, $\mathbf{\Delta}$, from Proposition 4.6 is the identity matrix. Taking expectations on both sides of the display and applying the Cauchy-Schwarz inequality, it holds that

$$\begin{aligned}
\mathbb{E}\Big[\underline{\mathsf{IGW}}(\hat{\mu}_n, \mu)\Big] &\leq \sqrt{2}\mathbb{E}\Big[\sqrt{M_2(\hat{\mu}_n) + M_2(\mu)}\, \underline{\mathsf{W}}_2(\hat{\mu}_n, \mu)\Big] \\
&\leq \sqrt{2}\sqrt{\mathbb{E}\Big[M_2(\hat{\mu}_n) + M_2(\mu)\Big]}\sqrt{\mathbb{E}\Big[\underline{\mathsf{W}}_2(\hat{\mu}_n, \mu)^2\Big]}.
\end{aligned}$$

Note that $\mathbb{E}\Big[M_2(\hat{\mu}_n)\Big] = \mathbb{E}\Big[\frac{1}{n}\sum_{i=1}^n \|X_i\|^2\Big] = \mathbb{E}[\|X_1\|^2] = M_2(\mu)$ since the samples are i.i.d. from μ. Conclude that $\mathbb{E}\Big[\underline{\mathsf{IGW}}(\hat{\mu}_n, \mu)\Big] \leq 2\sqrt{M_2(\mu)}\sqrt{\mathbb{E}\Big[\underline{\mathsf{W}}_2(\hat{\mu}_n, \mu)^2\Big]}$.

It remains to bound $\mathbb{E}\Big[\underline{\mathsf{W}}_2(\hat{\mu}_n, \mu)^2\Big]$ using bounds from the literature. For instance, if μ has finite q-th moment for $q > 2$, Equation (S40) in [20] asserts that there exists a constant $C_q > 0$ depending only on q for which

$$\mathbb{E}\Big[\underline{\mathsf{W}}_2(\hat{\mu}_n, \mu)^2\Big] \leq C_q M_q^{2/q}(\mu) \begin{cases} n^{-1/2}, & \text{if } q > 4 \\ n^{-1/2}\log n, & \text{if } q = 4 \\ n^{-(q-2)/q}, & \text{if } q \in (2,4), \end{cases}$$

which proves the one-sample result upon taking the square root. The two-sample case follows from the one-sample setting via the triangle inequality;

$$\begin{aligned}
\underline{\mathsf{IGW}}(\hat{\mu}_n, \hat{\nu}_n) - \underline{\mathsf{IGW}}(\mu,\nu) &\leq \underline{\mathsf{IGW}}(\hat{\mu}_n, \mu) + \underline{\mathsf{IGW}}(\hat{\nu}_n, \nu) + \underline{\mathsf{IGW}}(\mu,\nu) - \underline{\mathsf{IGW}}(\mu,\nu) \\
\underline{\mathsf{IGW}}(\hat{\mu}_n, \hat{\nu}_n) - \underline{\mathsf{IGW}}(\mu,\nu) &\geq \underline{\mathsf{IGW}}(\hat{\mu}_n, \hat{\nu}_n) - \underline{\mathsf{IGW}}(\hat{\nu}_n, \nu) - \underline{\mathsf{IGW}}(\hat{\mu}_n, \mu) - \underline{\mathsf{IGW}}(\hat{\mu}_n, \hat{\nu}_n)
\end{aligned}$$

so that $\left|\underline{\mathsf{IGW}}(\hat{\mu}_n, \hat{\nu}_n) - \underline{\mathsf{IGW}}(\mu,\nu)\right| \leq \underline{\mathsf{IGW}}(\hat{\mu}_n, \mu) + \underline{\mathsf{IGW}}(\hat{\nu}_n, \nu)$. □

A.8 Proof of Proposition 4.8

We start by noting the following Lipschitz continuity properties, see Appendix B.5 for a proof.

Lemma A.5 (Lipschitz continuity of projected IGW)**.** *Let $\mu \in \mathcal{P}(\mathbb{R}^{d_x}), \nu \in \mathcal{P}(\mathbb{R}^{d_y})$ have finite second moment. Then,*

1. *For any fixed $\mathbf{\Delta} \in \mathrm{St}(d_x, d_y)$, the function $\theta \in \mathbb{S}^{d_y - 1} \mapsto \mathsf{IGW}\left((\theta^\intercal \mathbf{\Delta})_\sharp \mu, (\theta^\intercal)_\sharp \nu\right)^2$ is Lipschitz continuous with constant at most $L_{\mu,\nu} \coloneqq 4(M_2(\mu) + M_2(\nu))^2$.*
2. *For any fixed $\theta \in \mathbb{S}^{d_y - 1}$, the function $\mathbf{\Delta} \in \mathrm{St}(d_x, d_y) \mapsto \mathsf{IGW}\left((\theta^\intercal \mathbf{\Delta})_\sharp \mu, (\theta^\intercal)_\sharp \nu\right)^2$ is Lipschitz continuous with respect to $\|\cdot\|_{\mathrm{F}}$ with constant at most $L'_{\mu,\nu} \coloneqq 4M_2(\mu)^2 + 4M_2(\mu)M_2(\nu)$.*

Proof of Proposition 4.8. Fix $\mathbf{\Delta} \in \mathrm{St}(d_x, d_y)$ and write $f(\theta) \coloneqq g_{\mathbf{\Delta}}(\theta)$, so that by Lemma A.5 we have that

$$|f(\theta) - f(\theta')| \leq L_{\mu,\nu}\|\theta - \theta'\|, \text{ for each } \theta, \theta' \in \mathbb{S}^{d_y - 1}.$$

It follows that $f/L_{\mu,\nu}$ is 1-Lipschitz so that Equation 3.41 in [86] yields that

$$\mathbb{P}\left(|f(\Theta)-\mathbb{E}[f(\Theta)]|^2 \geq t\right) = \mathbb{P}\left(\frac{|f(\Theta)-\mathbb{E}[f(\Theta)]|}{L_{\mu,\nu}} \geq \frac{\sqrt{t}}{L_{\mu,\nu}}\right) \leq 2\sqrt{2\pi}e^{-\frac{d_y t}{L^2_{\mu,\nu}}}.$$

Integrating the above inequality with respect to t, it holds that

$$\mathrm{Var}\left(f(\Theta)\right) \leq 2\sqrt{2\pi}\int_0^\infty e^{-\frac{d_y t}{L^2_{\mu,\nu}}}\,dt = 2\sqrt{2\pi}\frac{L^2_{\mu,\nu}}{d_y}. \tag{12}$$

Since $\Theta_1,\ldots,\Theta_m$ are i.i.d. with law given by Θ,

$$\widehat{\underline{\mathsf{IGW}}}^2_{\mathbf{\Delta},m}(\mu,\nu) - \underline{\mathsf{IGW}}^2_{\mathbf{\Delta}}(\mu,\nu) = \frac{1}{m}\sum_{j=1}^m \Big(f(\Theta_j)-\mathbb{E}[f(\Theta)]\Big),$$

so that $\widehat{\underline{\mathsf{IGW}}}^2_{\mathbf{\Delta},m}(\mu,\nu) - \underline{\mathsf{IGW}}^2_{\mathbf{\Delta}}(\mu,\nu)$ is mean-zero and, applying (12), we obtain that

$$\mathrm{Var}\Big(\widehat{\underline{\mathsf{IGW}}}^2_{\mathbf{\Delta},m}(\mu,\nu) - \underline{\mathsf{IGW}}^2_{\mathbf{\Delta}}(\mu,\nu)\Big) = \frac{1}{m}\mathrm{Var}\big(f(\Theta)\big) \leq \frac{2\sqrt{2\pi}L^2_{\mu,\nu}}{md_y}.$$

the result for fixed $\mathbf{\Delta}$ thus follows by applying Jensen's inequality as below,

$$\mathbb{E}\Big[\big|\widehat{\underline{\mathsf{IGW}}}^2_{\mathbf{\Delta},m}(\mu,\nu) - \underline{\mathsf{IGW}}^2_{\mathbf{\Delta}}(\mu,\nu)\big|\Big] \leq \sqrt{\mathbb{E}\Big[\big(\widehat{\underline{\mathsf{IGW}}}^2_{\mathbf{\Delta},m}(\mu,\nu) - \underline{\mathsf{IGW}}^2_{\mathbf{\Delta}}(\mu,\nu)\big)^2\Big]} \leq \frac{\sqrt{2}(2\pi)^{1/4}L_{\mu,\nu}}{\sqrt{md_y}}.$$

To prove the result with the infimum, we apply a Rademacher symmetrization argument. Let $\{\sigma_j\}_j$ be i.i.d. Rademacher random variables that take values ± 1 with equal probability which are independent of the Θ_j's. Then,

$$\begin{aligned}
\mathbb{E}\Big[\Big|\widehat{\underline{\mathsf{IGW}}}_m(\mu,\nu)^2 - \underline{\mathsf{IGW}}(\mu,\nu)^2\Big|\Big] &= \mathbb{E}_\Theta\left[\left|\inf_{\mathbf{\Delta}_1\in\mathrm{St}(d_x,d_y)}\frac{1}{m}\sum_{j=1}^m g_{\mathbf{\Delta}_1}(\Theta_j) - \inf_{\mathbf{\Delta}_2\in\mathrm{St}(d_x,d_y)}\mathbb{E}[g_{\mathbf{\Delta}_2}(\Theta_j)]\right|\right] \\
&\leq \mathbb{E}_\Theta\left[\sup_{\mathbf{\Delta}\in\mathrm{St}(d_x,d_y)}\left|\frac{1}{m}\sum_{j=1}^m g_{\mathbf{\Delta}}(\Theta_j) - \mathbb{E}[g_{\mathbf{\Delta}}(\Theta_j)]\right|\right] \\
&\leq 2\mathbb{E}_{\Theta,\sigma}\left[\sup_{\mathbf{\Delta}\in\mathrm{St}(d_x,d_y)}\left|\frac{1}{m}\sum_{j=1}^m \sigma_j g_{\mathbf{\Delta}}(\Theta_j)\right|\right] \\
&\leq 2\mathbb{E}_{\Theta,\sigma}\left[\sup_{f\in\mathcal{F}}\frac{1}{m}\sum_{j=1}^m \sigma_j f(\Theta_j)\right],
\end{aligned}$$

where $\mathcal{F} = \{\pm g_{\mathbf{\Delta}} : \mathbf{\Delta}\in\mathrm{St}(d_x,d_y)\}\cup\{0\}$. For any vector $\vartheta \overset{d}{=} (\Theta_1,\ldots,\Theta_m)$, define the metric $\|f-g\|_{p,\vartheta} = \left(\frac{1}{m}\sum_{j=1}^m |f(\Theta_j)-g(\Theta_j)|^p\right)^{1/p}$. For each fixed ϑ, equip $\mathcal{F}$ with the metric $\|\cdot\|_{2,\vartheta}$. Let $X^\vartheta_f = \frac{1}{m}\sum_{j=1}^m \sigma_j f(\Theta_j)$ and notice that the random process $\{X^\vartheta_f\}_{f\in\mathcal{F}}$ has zero mean and sub-Gaussian increments. In particular, we have that

$$\mathbb{E}[e^{\lambda(X^\vartheta_f - X^\vartheta_g)}] = \mathbb{E}\left[e^{\frac{\lambda}{m}\sum_{j=1}^m \sigma_j(f(\Theta_j)-g(\Theta_j))}\right] \leq e^{\frac{\lambda^2}{m^2}\sum_{j=1}^m \frac{|f(\Theta_j)-g(\Theta_j)|^2}{2}} = e^{\frac{\lambda^2}{2m}\|f-g\|^2_{2,\vartheta}},$$

as follows from the proof of Theorem 2.2.1 in [87].

Next, we consider an ϵ-net for the metric space $(\mathcal{F},\|\cdot\|_{2,\vartheta})$; a set $\mathcal{F}_\epsilon\subseteq\mathcal{F}$ is called an ϵ-net for $\mathcal{F}$ if, for every $f\in\mathcal{F}$, there exists $f_\epsilon\in\mathcal{F}_\epsilon$ such that $\|f-f_\epsilon\|_{2,\vartheta}\leq\epsilon$. The ϵ-covering number is then

$$\mathcal{N}(\mathcal{F},\|\cdot\|_{2,\vartheta},\epsilon) = \inf\{\mathrm{card}\,(\mathcal{F}_\epsilon) : \mathcal{F}_\epsilon \text{ is an } \epsilon\text{-net of } \mathcal{F}\}.$$

Recall that $0 \in \mathcal{F}$ and, applying Dudley's entropy integral bound with the pseudometric being $\rho(f,g) = \frac{1}{\sqrt{m}}\|f-g\|_{2,\vartheta}$ [86, Theorem 5.22], we obtain that

$$\begin{aligned}
\mathbb{E}_\sigma\left[\sup_{f\in\mathcal{F}} X_f^\vartheta\right] &\le 2\mathbb{E}_\sigma\left[\sup_{f,g\in\mathcal{F},\rho(f,g)\le\epsilon}\left|X_f^\vartheta - X_g^\vartheta\right|\right] + 32\int_{\epsilon/4}^{\sup_{f,g\in\mathcal{F}}\rho(f,g)}\sqrt{\log\mathcal{N}(\mathcal{F},\rho,\delta)}\,d\delta \\
&\le 2\mathbb{E}_\sigma\left[\sup_{f,g\in\mathcal{F},\|f-g\|_{2,\vartheta}\le\epsilon\sqrt{m}}\frac{1}{m}\sum_{j=1}^m|\sigma_j||f(\Theta_j)-g(\Theta_j)|\right] \\
&\qquad + 32\int_{\epsilon/4}^{\sup_{f,g\in\mathcal{F}}\frac{1}{\sqrt{m}}\|f-g\|_{2,\vartheta}}\sqrt{\log\mathcal{N}\left(\mathcal{F},\frac{1}{\sqrt{m}}\|\cdot\|_{2,\vartheta},\delta\right)}\,d\delta \\
&\le 2\epsilon\sqrt{m} + \frac{32}{\sqrt{m}}\int_{\sqrt{m}\epsilon/4}^{\sup_{f,g\in\mathcal{F}}\|f-g\|_{2,\vartheta}}\sqrt{\log\mathcal{N}(\mathcal{F},\|\cdot\|_{2,\vartheta},\delta)}\,d\delta \\
&\le 2\epsilon\sqrt{m} + 64m^{-1/2}\sup_{f\in\mathcal{F}}\|f\|_{2,\vartheta}\sqrt{\log\mathcal{N}(\mathcal{F},\|\cdot\|_{2,\vartheta},\sqrt{m}\epsilon/4)}.
\end{aligned}$$

Since ϵ is arbitrary, we may scale ϵ by $\frac{4}{\sqrt{m}}$ to obtain that

$$\mathbb{E}_\sigma\left[\sup_{f\in\mathcal{F}} X_f^\vartheta\right] \le \inf_{\epsilon>0}\left\{64m^{-1/2}\sup_{f\in\mathcal{F}}\|f\|_{2,\vartheta}\sqrt{\log\mathcal{N}(\mathcal{F},\|\cdot\|_{2,\vartheta},\epsilon)} + 8\epsilon\right\}.$$

From the expression for univariate IGW in Theorem 3.1, we know that, for any θ,

$$g_{\boldsymbol{\Delta}}(\theta) = \mathsf{IGW}\big((\theta^\intercal\boldsymbol{\Delta})_\sharp\mu, (\theta^\intercal)_\sharp\nu\big)^2 \le M_2((\theta^\intercal\boldsymbol{\Delta})_\sharp\mu)^2 + M_2((\theta^\intercal)_\sharp\nu)^2 \le M_2(\mu)^2 + M_2(\nu)^2.$$

Therefore,

$$\sup_{f\in\mathcal{F}}\|f\|_{2,\vartheta} = \sup_{\boldsymbol{\Delta}\in\mathrm{St}(d_x,d_y)}\|g_{\boldsymbol{\Delta}}\|_{2,\vartheta} = \sup_{\boldsymbol{\Delta}\in\mathrm{St}(d_x,d_y)}\left(\frac{1}{m}\sum_{j=1}^m g_{\boldsymbol{\Delta}}(\Theta_j)^2\right)^{1/2} \le M_2(\mu)^2 + M_2(\nu)^2.$$

Combined, we have that

$$\mathbb{E}_{\Theta,\sigma}\left[\sup_{f\in\mathcal{F}}\frac{1}{m}\sum_{j=1}^m\sigma_j f(\Theta_j)\right] \le \inf_{\epsilon>0}\left\{8\epsilon + \mathbb{E}_\Theta\left[\frac{64(M_2(\mu)^2+M_2(\nu)^2)}{\sqrt{m}}\sqrt{\log\mathcal{N}(\mathcal{F},\|\cdot\|_{2,\vartheta},\epsilon)}\right]\right\}. \tag{13}$$

We now bound the covering number $\mathcal{N}(\mathcal{F},\|\cdot\|_{2,\vartheta},\epsilon)$. By Lemma A.5 the function $\boldsymbol{\Delta}\mapsto g_{\boldsymbol{\Delta}}(\theta)$ is $L'_{\mu,\nu}$-Lipschitz uniformly in θ so that, for each $\boldsymbol{\Delta}_1,\boldsymbol{\Delta}_2\in\mathrm{St}(d_x,d_y)$,

$$\|g_{\boldsymbol{\Delta}_1} - g_{\boldsymbol{\Delta}_2}\|_{2,\vartheta}^2 = \frac{1}{m}\sum_{j=1}^m(g_{\boldsymbol{\Delta}_1}(\Theta_j) - g_{\boldsymbol{\Delta}_2}(\Theta_j))^2 \le (L'_{\mu,\nu})^2\|\boldsymbol{\Delta}_1-\boldsymbol{\Delta}_2\|_{\mathrm{F}}^2.$$

Thus, $\mathcal{N}(\mathcal{F},\|\cdot\|_{2,\vartheta},\epsilon) \le 2\mathcal{N}(\mathrm{St}(d_x,d_y),\|\cdot\|_{\mathrm{F}},\epsilon/L'_{\mu,\nu}) + 1$ provided that $L'_{\mu,\nu}\neq 0$ and it suffices to bound the covering number of the Stiefel manifold. If $L'_{\mu,\nu}=0$, the covering number of $\mathcal{F}$ under $\|\cdot\|_{2,\vartheta}$ is at most 3 and the proof is complete. To bound $\mathcal{N}(\mathrm{St}(d_x,d_y),\|\cdot\|_{\mathrm{F}},\epsilon/L'_{\mu,\nu})$, we construct a $\epsilon/L'_{\mu,\nu}$-net column by column. We begin with the first column, which lies in the unit sphere $\mathbb{S}^{d_y-1}$. Let $\delta_1>0$ and choose a δ_1-net of $\mathbb{S}^{d_y-1}$. By Lemma 5.2 in [88], $\mathcal{N}(\mathbb{S}^{d_y-1},\|\cdot\|_2,\delta_1)\le 3^{d_y}\delta_1^{-d_y}$ for each $\delta_1<1$. For each vector in this δ_1-net, we pick a δ_2-net for the set vectors orthogonal to it, which is isometric to $\mathbb{S}^{d_y-2}$. This can result in at most $\delta_1+\delta_2$ error in the second vector by the triangle inequality. Following the same procedure, the error in the i^{th} column is at most $\sum_{j=1}^i\delta_j$. To obtain a $\epsilon/L'_{\mu,\nu}$-net under $\|\cdot\|_{\mathrm{F}}$, we need

$$\sum_{i=1}^{d_x}\left(\sum_{j=1}^i\delta_j\right)^2 \le \left(\epsilon/L'_{\mu,\nu}\right)^2, \tag{14}$$

so that the covering number will be bounded by

$$\mathcal{N}(\mathrm{St}(d_x,d_y),\|\cdot\|_{\mathrm{F}},\epsilon/L'_{\mu,\nu}) \leq \prod_{i=1}^{d_x}(\delta_i/3)^{-(d_y-i+1)}. \tag{15}$$

Setting $\delta_i = \delta M^{i-1}$ for some $0 < M < 1$, we bound the left hand side of (14) as

$$\sum_{i=1}^{d_x}\left(\sum_{j=1}^{i}\delta_j\right)^2 = \sum_{i=1}^{d_x}\delta^2\left(\sum_{j=1}^{i}M^{j-1}\right)^2 = \sum_{i=1}^{d_x}\delta^2\left(\frac{1-M^i}{1-M}\right)^2 \leq \delta^2 d_x\left(\frac{1}{1-M}\right)^2$$

So that the choice $\delta_i = \delta M^{i-1}$ is valid for (14) for the choice $\delta = d_x^{-1/2}(1-M)\epsilon/L'_{\mu,\nu}$, where $M \in (0,1)$ is arbitrary. With this choice of δ_i's, (15) can be bounded as

$$\begin{aligned}\mathcal{N}\left(\mathrm{St}(d_x,d_y),\|\cdot\|_{\mathrm{F}},\epsilon/L'_{\mu,\nu}\right) &\leq \prod_{i=1}^{d_x}\left(d_x^{-1/2}\frac{(1-M)M^{i-1}\epsilon}{3L'_{\mu,\nu}}\right)^{-(d_y-i+1)} \\ &\leq C_{d_x,d_y}\left(\epsilon/L'_{\mu,\nu}\right)^{-\left(d_xd_y-\frac{d_x(d_x-1)}{2}\right)}.\end{aligned}$$

Therefore, plugging this expression into (13), we have for any $\epsilon > 0$,

$$\begin{aligned}&\mathbb{E}_{\Theta,\sigma}\left[\sup_{\boldsymbol{\Delta}\in\mathrm{St}(d_x,d_y)}\left|\frac{1}{m}\sum_{j=1}^{m}\sigma_j g_{\boldsymbol{\Delta}}(\Theta_j)\right|\right] \\ &\leq 8\epsilon + \frac{64(M_2(\mu)^2+M_2(\nu)^2)}{\sqrt{m}}\sqrt{\left(d_xd_y-\frac{d_x(d_x-1)}{2}\right)\log\left(L'_{\mu,\nu}/\epsilon\right)+\log 4C_{d_x,d_y}}.\end{aligned}$$

Recall that $L'_{\mu,\nu} = 4M_2(\mu)^2 + 4M_2(\mu)M_2(\nu) \leq 6(M_2(\mu)^2+M_2(\nu)^2)$. Picking $\epsilon \sim L'_{\mu,\nu}m^{-1/2}$, it holds that

$$\mathbb{E}\left[\left|\widehat{\underline{\mathsf{IGW}}}_m(\mu,\nu)^2 - \underline{\mathsf{IGW}}(\mu,\nu)^2\right|\right] \lesssim \frac{(M_2(\mu)^2+M_2(\nu)^2)}{\sqrt{m}}\sqrt{\left(d_xd_y-\frac{d_x(d_x-1)}{2}\right)\log\left(m^{1/2}\right)}.$$

This proves the claim. □

A.9 Proof of Proposition 4.10

We begin with the proof of the first point. Note that G^θ is the optimal value function of a concave minimization problem over the polyhedron $\Pi(\mu,\nu)$, recalling that μ and ν are finitely supported so that the marginal constraints simplify to $2N$ equality constraints. Letting $\mathcal{V} = \{\pi_1,\ldots,\pi_{|\mathcal{V}|}\}$ denote the set of vertices of $\Pi(\mu,\nu)$, concavity of the objective implies that

$$\inf_{\pi\in\Pi(\mu,\nu)} -2\left(\int \theta^\intercal\boldsymbol{\Delta}xy^\intercal\theta d\pi(x,y)\right)^2 = \min_{i=1}^{|\mathcal{V}|} -2\left(\int \theta^\intercal\boldsymbol{\Delta}xy^\intercal\theta d\pi_i(x,y)\right)^2 \coloneqq \min_{i=1}^{|\mathcal{V}|}\ell_{i,\theta}(\boldsymbol{\Delta}),$$

where $\ell_{i,\theta}(\boldsymbol{\Delta}) = -2\left(\left\langle\left(\int xy^\intercal\theta\theta^\intercal d\pi_i(x,y)\right)^\intercal,\boldsymbol{\Delta}\right\rangle_{\mathrm{F}}\right)^2$, it follows from the chain rule that

$$D_{[\boldsymbol{\Delta}]}\ell_{i,\theta}(\Xi) = -4\left\langle\left(\int xy^\intercal\theta\theta^\intercal d\pi_i(x,y)\right)^\intercal,\boldsymbol{\Delta}\right\rangle_{\mathrm{F}}\left\langle\left(\int xy^\intercal\theta\theta^\intercal d\pi_i(x,y)\right)^\intercal,\Xi\right\rangle_{\mathrm{F}},$$

for each $i \in \{1,\ldots,|\mathcal{V}|\}$ and each $\Xi \in \mathbb{R}^{d_y\times d_x}$. Local Lipschitz continuity and the formula for the subdifferential of G^θ then follows from Proposition 2.3.12 and the surrounding discussion in [71] (see also Proposition 2.3.6 in the same reference).

The second point is a consequence of the sum rule for the Clarke subdifferential (see Corollary 2 and 3 on p.39-40 as well as Proposition 2.3.12 of [71]) noting that the function $\boldsymbol{\Delta} \mapsto M_2^2((\theta^\intercal\boldsymbol{\Delta})_\sharp\mu)$ is Fréchet differentiable at $\boldsymbol{\Delta} \in \mathbb{R}^{d_y\times d_x}$ with derivative

$$\Xi\in\mathbb{R}^{d_y\times d_x} \mapsto 2\theta^\intercal\boldsymbol{\Delta}\mathbf{M}_\mu\boldsymbol{\Delta}^\intercal\theta\left(\theta^\intercal\Xi\mathbf{M}_\mu\boldsymbol{\Delta}^\intercal\theta+\theta^\intercal\boldsymbol{\Delta}\mathbf{M}_\mu\Xi^\intercal\theta\right) = 4\theta^\intercal\boldsymbol{\Delta}\mathbf{M}_\mu\boldsymbol{\Delta}^\intercal\theta\left\langle(\mathbf{M}_\mu\boldsymbol{\Delta}^\intercal\theta\theta^\intercal)^\intercal,\Xi\right\rangle_{\mathrm{F}}.$$

□

B Proofs of Auxiliary Results

We begin this section with a useful technical lemma; the remainder of the section compiles the proofs of the lemmas used to prove the main results.

Lemma B.1. *Given measurable functions $T : \mathbb{R}^{d_0} \to \mathbb{R}^{d_0'}$ and $S : \mathbb{R}^{d_1} \to \mathbb{R}^{d_1'}$,*

$$\Pi(T_\sharp \mu, S_\sharp \nu) = \{(T, S)_\sharp \pi : \pi \in \Pi(\mu, \nu)\}.$$

Proof. It is clear that, for any $\pi' \in \Pi(\mu, \nu)$, $(T, S)_\sharp \pi' \in \Pi(T_\sharp \mu, S_\sharp \nu)$. On the other hand, if we fix some $\pi \in \Pi(T_\sharp \mu, S_\sharp \nu)$, we may disintegrate μ and ν as $(\mu_{x'})_{x' \in \mathbb{R}^{d_0'}} \subset \mathcal{P}(\mathbb{R}^{d_0})$ and $(\nu_{y'})_{y' \in \mathbb{R}^{d_1'}} \subset \mathcal{P}(\mathbb{R}_1^d)$ for which

$$\mu(A) = \int_{\mathbb{R}^{d_0'}} \left(\int_{T^{-1}(x')} \mathbb{1}_A(x) d\mu_{x'}(x) \right) d\left(T_\sharp \mu\right)(x'),$$
$$\nu(B) = \int_{\mathbb{R}^{d_1'}} \left(\int_{S^{-1}(y')} \mathbb{1}_B(y) d\nu_{y'}(y) \right) d\left(S_\sharp \nu\right)(y'),$$

for any Borel sets $A \subset \mathbb{R}^{d_0}, B \subset \mathbb{R}^{d_1}$ and

$$\mu_{x'}\left(T^{-1}(x')\right) = 1 \quad \text{for } T_\sharp \mu\text{-a.e. } x', \quad \nu_{y'}\left(S^{-1}(y')\right) = 1 \quad \text{for } S_\sharp \nu\text{-a.e. } y',$$

see Theorem 5.3.1 in [89]. Define $\pi' \in \mathcal{P}(\mathbb{R}^{d_0} \times \mathbb{R}^{d_1})$ via

$$\pi'(A \times B) = \int_{\mathbb{R}^{d_0'} \times \mathbb{R}^{d_1'}} \left(\int_{T^{-1}(x')} \int_{S^{-1}(y')} \mathbb{1}_{A \times B}(x, y) d\nu_{y'}(y) d\mu_{x'}(x) \right) d\pi(x', y'),$$

for each pair of Borel sets $A \subset \mathbb{R}^{d_0}, B \subset \mathbb{R}^{d_1}$. Then,

$$\pi'(A \times \mathbb{R}^{d_1}) = \int_{\mathbb{R}^{d_0'} \times \mathbb{R}^{d_1'}} \int_{T^{-1}(x')} \mathbb{1}_A(x) d\mu_{x'}(x) \nu_{y'}(S^{-1}(y')) d\pi(x', y') = \mu(A),$$

so that π' has first marginal given by μ; a similar argument shows that its second marginal is ν. Finally, for any Borel sets $C \subset \mathbb{R}^{d_0'}, D \subset \mathbb{R}^{d_1'}$,

$$\begin{aligned}(T, S)_\sharp \pi'(C \times D) &= \int_{\mathbb{R}^{d_0'} \times \mathbb{R}^{d_1'}} \mu_{x'}\left(T^{-1}(x') \cap T^{-1}(C)\right) \nu_{y'}\left(S^{-1}(y') \cap S^{-1}(D)\right) d\pi(x', y') \\ &= \int_{\mathbb{R}^{d_0'} \times \mathbb{R}^{d_1'}} \mathbb{1}_{C \times D}(x', y') d\pi(x', y') = \pi(C \times D),\end{aligned}$$

proving the desired claim. □

B.1 Proof of Lemma A.1

Let $\pi^\star$ be an arbitrary solution of $\mathsf{IGW}(\mu, \nu)$ and consider the SVD of the matrix $\int xy^\intercal d\pi^\star(x, y) = \mathbf{P}\mathbf{\Lambda}_{\pi^\star}\mathbf{Q}^\intercal$ where $\mathbf{P} \in \mathbb{R}^{d_x \times d_x}$ and $\mathbf{Q} \in \mathbb{R}^{d_y \times d_y}$ are orthogonal and $\mathbf{\Lambda}_{\pi^\star} \in \mathbb{R}^{d_x \times d_y}$ contains the singular values ordered from largest to smallest. Now, let $\mathbf{\Delta} = [\mathbf{I}_{d_x} \quad \mathbf{0}]^\intercal \in \mathrm{St}(d_x, d_y) \subset \mathbb{R}^{d_y \times d_x}$. Since IGW is invariant to orthogonal transformations, $\mathsf{IGW}((\mathbf{\Delta}\mathbf{P}^\intercal)_\sharp \mu, (\mathbf{Q}^\intercal)_\sharp \nu) = \mathsf{IGW}(\mu, \nu)$ and, by virtue of Lemma B.1, $\tilde{\pi} := (\mathbf{\Delta}\mathbf{P}^\intercal, \mathbf{Q}^\intercal)_\sharp \pi^\star \in \Pi(\mathbf{\Delta}\mathbf{P}^\intercal)_\sharp \mu, (\mathbf{Q})_\sharp^\intercal \nu)$ with

$$\begin{aligned}\mathsf{IGW}((\mathbf{\Delta}\mathbf{P}^\intercal)_\sharp \mu, (\mathbf{Q})_\sharp^\intercal \nu)^2 &\leq \iint |\langle x, x' \rangle - \langle y, y' \rangle|^2 d\tilde{\pi} \otimes \tilde{\pi}(x, y, x', y') \\ &= \iint |\langle \mathbf{\Delta}\mathbf{P}^\intercal x, \mathbf{\Delta}\mathbf{P}^\intercal x' \rangle - \langle \mathbf{Q}^\intercal y, \mathbf{Q}^\intercal y' \rangle|^2 d\pi^\star \otimes \pi^\star(x, y, x', y') \\ &= \iint |\langle x, x' \rangle - \langle y, y' \rangle|^2 d\pi^\star \otimes \pi^\star(x, y, x', y') = \mathsf{IGW}(\mu, \nu)^2,\end{aligned}$$

where we have used the fact that $\mathbf{\Delta}\mathbf{P}^\intercal$ and $\mathbf{Q}^\intercal$ are orthogonal matrices. It follows that $\tilde{\pi}$ solves $\mathsf{IGW}((\mathbf{\Delta}\mathbf{P}^\intercal)_\sharp \mu, (\mathbf{Q})_\sharp^\intercal \nu)$.

Observe that

$$\mathsf{W}_2((\boldsymbol{\Delta}\mathbf{P}^\intercal)_\sharp\mu,(\mathbf{Q}^\intercal)_\sharp\nu)^2 \leq \int \|x-y\|^2 d\tilde{\pi}(x,y) = \int \|x\|^2 - 2\langle x,y\rangle + \|y\|^2 d\tilde{\pi}(x,y)$$
$$= M_2((\boldsymbol{\Delta}\mathbf{P}^\intercal)_\sharp\mu) + M_2((\mathbf{Q}^\intercal)_\sharp\nu) - 2\mathrm{Tr}\left(\int xy^\intercal d\tilde{\pi}(x,y)\right).$$

On the other hand, setting $\tilde{\mu} = (\boldsymbol{\Delta}\mathbf{P}^\intercal)_\sharp\mu$ and $\tilde{\nu} = (\mathbf{Q}^\intercal)_\sharp\nu$

$$\mathsf{IGW}(\mu,\nu)^2 = \mathsf{IGW}(\tilde{\mu},\tilde{\nu})^2 = \left\|\int xx^\intercal d\tilde{\mu}(x)\right\|_{\mathrm{F}}^2 + \left\|\int yy^\intercal d\tilde{\nu}(y)\right\|_{\mathrm{F}}^2 - 2\left\|\int xy^\intercal d\tilde{\pi}(x,y)\right\|_{\mathrm{F}}^2.$$

Note that

$$\int xy^\intercal d\tilde{\pi}(x,y) = \boldsymbol{\Delta}\mathbf{P}^\intercal \int xy^\intercal d\pi^\star(x,y)\mathbf{Q} = \boldsymbol{\Delta}\boldsymbol{\Lambda}_{\pi^\star}$$

and observe that $\boldsymbol{\Delta}\boldsymbol{\Lambda}_{\pi^\star} = \mathrm{diag}\left((\boldsymbol{\Lambda}_{\pi^\star})_{11} \quad (\boldsymbol{\Lambda}_{\pi^\star})_{22} \quad \cdots \quad (\boldsymbol{\Lambda}_{\pi^\star})_{d_xd_x} \quad 0 \quad \cdots \quad 0\right) \in \mathbb{R}^{d_y\times d_y}$ so that $\int xy^\intercal d\tilde{\pi}(x,y)$ is a diagonal matrix. Conclude that

$$\left\|\int xy^\intercal d\tilde{\pi}(x,y)\right\|_{\mathrm{F}}^2 = \sum_{i=1}^{d_x}(\boldsymbol{\Lambda}_{\pi^\star})_{ii}^2 \text{ and } \mathrm{Tr}\left(\int xy^\intercal d\tilde{\pi}(x,y)\right) = \sum_{i=1}^{d_x}(\boldsymbol{\Lambda}_{\pi^\star})_{ii}$$

and from the above, $(\boldsymbol{\Lambda}_{\pi^\star})_{ii} = \int x_iy_i d\tilde{\pi}(x,y)$, where we observe that

$$0 \leq \int (x_i-y_i)^2 d\tilde{\pi}(x,y) = \int x_i^2 d\tilde{\mu}(x) + \int y_i^2 d\tilde{\nu}(y) - 2\int x_iy_i d\tilde{\pi}(x,y). \tag{16}$$

Conclude that $0 \leq 2\left(\boldsymbol{\Lambda}_{\pi^\star}\right)_{ii} = 2\int x_iy_i d\tilde{\pi}(x,y) \leq M_2(\tilde{\mu})_i + M_2(\tilde{\nu})_i$, where we have set $M_2(\tilde{\mu})_i = \int x_i^2 d\tilde{\mu}(x)$ for $i = 1,\ldots,d_x$ and similarly for $\tilde{\nu}$ so that, for each $i = 1,\ldots,d_x$,

$$-2\left(\int x_iy_i d\tilde{\pi}(x,y)\right)^2 \geq -\left(\int x_iy_i d\tilde{\pi}(x,y)\right)\left(M_2(\tilde{\mu})_i + M_2(\tilde{\nu})_i\right). \tag{17}$$

By the Cauchy-Schwarz inequality, we further have that

$$0 \leq \int x_iy_i d\tilde{\pi}(x,y) \leq \sqrt{M_2(\tilde{\mu})_iM_2(\tilde{\nu})_i} \tag{18}$$

so that $\int x_iy_i d\tilde{\pi}(x,y) = 0$ at each $i \in \mathcal{I}^c$ where $\mathcal{I} \coloneqq \{i = 1,\ldots,d_x : M_2(\tilde{\mu})_i \text{ and } M_2(\tilde{\nu})_i \neq 0\}$ which we assume to be nonempty. Setting $\mathcal{J} \coloneqq \{j = 1,\ldots,d_y : M_2(\tilde{\nu})_j \neq 0\} \cap \mathcal{I}^c$ and $\mathcal{K} \coloneqq \{k = 1,\ldots,d_y : M_2(\tilde{\mu})_k \neq 0\} \cap \mathcal{I}^c$,

$$\begin{aligned}
\mathsf{IGW}(\mu,\nu)^2 &= \sum_{i,j=1}^{d_y}\left(\int x_ix_j d\tilde{\mu}(x)\right)^2 + \sum_{i,j=1}^{d_y}\left(\int y_iy_j d\tilde{\nu}(y)\right)^2 - 2\sum_{i=1}^{d_x}\left(\int x_iy_i d\tilde{\pi}(x,y)\right)^2 \\
&\geq \sum_{i\in\mathcal{I}}\left(M_2(\tilde{\mu})_i^2 + M_2(\tilde{\nu})_i^2 - 2\left(\int x_iy_i d\tilde{\pi}(x,y)\right)^2\right) + \sum_{j\in\mathcal{J}}M_2(\tilde{\nu})_j^2 + \sum_{k\in\mathcal{K}}M_2(\tilde{\mu})_k^2 \\
&\geq \frac{1}{2}\sum_{i\in\mathcal{I}}\left(M_2(\tilde{\mu})_i + M_2(\tilde{\nu})_i\right)\left(M_2(\tilde{\mu})_i + M_2(\tilde{\nu})_i - 2\int x_iy_i d\tilde{\pi}(x,y)\right) \\
&\quad + \sum_{j\in\mathcal{J}}M_2(\tilde{\nu})_j^2 + \sum_{k\in\mathcal{K}}M_2(\tilde{\mu})_k^2,
\end{aligned} \tag{19}$$

where we have used the standard inequality $(a+b)^2 \leq 2(a^2+b^2)$ and (17) in the final lower bound. By (16) we further have that

$$\begin{aligned}
&\sum_{i\in\mathcal{I}}\left(M_2(\tilde{\mu})_i + M_2(\tilde{\nu})_i\right)\underbrace{\left(M_2(\tilde{\mu})_i + M_2(\tilde{\nu})_i - 2\int x_iy_i d\tilde{\pi}(x,y)\right)}_{\geq 0} \\
&\qquad \geq \sum_{i\in\mathcal{I}}\min_{i\in\mathcal{I}}\left\{M_2(\tilde{\mu})_i + M_2(\tilde{\nu})_i\right\}\left(M_2(\tilde{\mu})_i + M_2(\tilde{\nu})_i - 2\int x_iy_i d\tilde{\pi}(x,y)\right) \\
&\qquad = \iota\sum_{i\in\mathcal{I}}\left(M_2(\tilde{\mu})_i + M_2(\tilde{\nu})_i - 2\int x_iy_i d\tilde{\pi}(x,y)\right),
\end{aligned}$$

for $\iota \coloneqq \min_{i\in\mathcal{I}}\{M_2(\tilde{\mu})_i + M_2(\tilde{\nu})_i\}$. Applying this bound yields

$$\begin{aligned}
&\mathsf{W}_2((\mathbf{\Delta}\mathbf{P}^\intercal)_\sharp\mu, (\mathbf{Q}^\intercal)_\sharp\nu)^2 \\
&\quad\leq \sum_{i=1}^{d_y} M_2(\tilde{\mu})_i + \sum_{j=1}^{d_y} M_2(\tilde{\nu})_j - 2\sum_{i=1}^{d_x}\int x_i y_i d\tilde{\pi}(x,y) \\
&\quad= \sum_{i\in\mathcal{I}}\left(M_2(\tilde{\mu})_i + M_2(\tilde{\nu})_i - 2\int x_i y_i d\tilde{\pi}(x,y)\right) + \sum_{j\in\mathcal{J}} M_2(\tilde{\nu})_j + \sum_{k\in\mathcal{K}} M_2(\tilde{\mu})_k \\
&\quad\leq \frac{2}{\iota}\left(\mathsf{IGW}(\mu,\nu)^2 - \sum_{j\in\mathcal{J}} M_2(\tilde{\nu})_j^2 - \sum_{k\in\mathcal{K}} M_2(\tilde{\mu})_k^2\right) + \sum_{j\in\mathcal{J}} M_2(\tilde{\nu})_j + \sum_{k\in\mathcal{K}} M_2(\tilde{\mu})_k,
\end{aligned}$$

proving the claim.

B.2 Proof of Lemma A.2

We assume the notation of the proof of Proposition 2.2 above, letting $\pi^\star$ solve $\mathsf{IGW}(\mu,\nu)$ and $\tilde{\pi}$ solve $\mathsf{IGW}(\tilde{\mu},\tilde{\nu})$. From (18), if $\mathcal{I}=\emptyset$, then $\int x_i y_i d\tilde{\pi}(x,y) = 0$ for every $i \in \{1,\dots,d_x\}$ so that all singular values of $\int xy^\intercal d\pi^\star(x,y)$ are 0, i.e., $\int xy^\intercal d\pi^\star(x,y)$ is the 0 matrix. From the decomposition

$$\mathsf{IGW}(\mu,\nu)^2 = \left\|\int xx^\intercal d\mu(x)\right\|_{\mathrm{F}}^2 + \left\|\int yy^\intercal d\nu(y)\right\|_{\mathrm{F}}^2 + \inf_{\pi\in\Pi(\mu,\nu)} -2\left\|\int xy^\intercal d\pi(x,y)\right\|_{\mathrm{F}}^2,$$

we see that every coupling $\pi\in\Pi(\mu,\nu)$ must, therefore, satisfy $\int xy^\intercal d\pi(x,y)=0$.

Applying Lemma B.1 it follows that, for any $a\in\mathbb{R}^{d_0}, b\in\mathbb{R}^{d_1}$, every coupling, $\pi_{a,b}$, between $(a^\intercal)_\sharp\mu$ and $(b^\intercal)_\sharp\nu$ satisfies $\int xyd\pi_{a,b}(x,y)=0$. Assuming to the contrary that μ and ν are not point masses, there exist a,b for which $\mu_a = (a^\intercal)_\sharp\mu$ and $\nu_b = (b^\intercal)_\sharp\nu$ are not point masses. Hence, there exist $x_1<x_2$ and $y_1<y_2$ for which $\mu_a((-\infty,x_1]), \mu_a([x_2,\infty)), \nu_b((-\infty,y_1]), \nu_b([y_2,\infty)) > 0$. Fix some $\epsilon>0$ and consider the measure

$$\begin{aligned}
\pi' = \mu_a\otimes\nu_b + \epsilon\big(&\mu_a|_{(-\infty,x_1]}\otimes\nu_b|_{[y_2,\infty)} + \mu_a|_{[x_2,\infty)}\otimes\nu_b|_{(-\infty,y_1]} \\
&-\mu_a|_{(-\infty,x_1]}\otimes\nu_b|_{(-\infty,y_1]} - \mu_a|_{[x_2,\infty)}\otimes\nu_b|_{[y_2,\infty)}\big),
\end{aligned}$$

where $\mu_a|_{(-\infty,x_1])}$ is the restriction of μ_a to $(-\infty,x_1]$ normalized by $\mu_a((-\infty,x_1])$ and so on. It is easy to see that π' is a coupling of (μ_a,ν_b) and, for any $A\subset(-\infty,x_1], B\subset(-\infty,y_1]$,

$$\begin{aligned}
\pi'(A\times B) &= \mu_a(A)\nu_b(B) - \epsilon\mu_a|_{(-\infty,x_1]}(A)\,\nu_b|_{(-\infty,y_1]}(B) \\
&= \mu_a(A)\nu_b(B)\left(1 - \frac{\epsilon}{\mu_a((-\infty,x_1])\nu_b((-\infty,y_1])}\right)
\end{aligned}$$

so that we set $0<\epsilon<\mu_a((-\infty,x_1])\nu_b((-\infty,y_1])$ to ensure that $\pi'(A\times B)\geq 0$. Similarly, if $C\subset[x_2,\infty)$ and $D\subset[y_2,\infty)$,

$$\pi'(C\times D) = \mu_a(C)\nu_b(D)\left(1 - \frac{\epsilon}{\mu_a([x_2,\infty))\nu_b([y_2,\infty))}\right) \geq 0$$

if we set $0<\epsilon<\mu_a([x_2,\infty))\nu_b([y_2,\infty))$. We conclude that π' is a nonnegative measure with marginals μ_a and ν_b once ϵ is sufficiently small and hence is an element of $\Pi(\mu_a,\nu_b)$. Finally,

$$\begin{aligned}
\int xyd\pi'(x,y) =& \mathbb{E}_{\mu_a}[X]\mathbb{E}_{\nu_b}[Y] + \epsilon\mathbb{E}_{\mu_a|_{[x_2,\infty)}}[X]\mathbb{E}_{\nu_b|_{(-\infty,y_1]}}[Y] + \epsilon\mathbb{E}_{\mu_a|_{(-\infty,x_1]}}[X]\mathbb{E}_{\nu_b|_{[y_2,\infty)}}[Y] \\
&- \epsilon\mathbb{E}_{\mu_a|_{(-\infty,x_1]}}[X]\mathbb{E}_{\nu_b|_{(-\infty,y_1]}}[Y] - \epsilon\mathbb{E}_{\mu_a|_{[x_2,\infty)}}[X]\mathbb{E}_{\nu_b|_{[y_2,\infty)}}[Y] \\
=& \mathbb{E}_{\mu_a}[X]\mathbb{E}_{\nu_b}[Y] + \epsilon\Big(\mathbb{E}_{\mu_a|_{[x_2,\infty)}}[X] - \mathbb{E}_{\mu_a|_{(-\infty,x_1]}}[X]\Big)\Big(\mathbb{E}_{\nu_b|_{(-\infty,y_1]}}[Y] - \mathbb{E}_{\nu_b|_{[y_2,\infty)}}[Y]\Big).
\end{aligned}$$

Evidently, $\mathbb{E}_{\mu_a|_{[x_2,\infty)}}[X] > \mathbb{E}_{\mu_a|_{(-\infty,x_1]}}[X]$ and $\mathbb{E}_{\nu_b|_{(-\infty,y_1]}}[Y] < \mathbb{E}_{\nu_b|_{[y_2,\infty)}}[Y]$ so that $\int xyd\pi'(x,y) < \int xyd\mu_a\otimes\nu_b(x,y)$, contradicting the fact that all couplings must have 0 cross-correlation.

B.3 Proof of Corollary A.3

If neither μ nor ν is a point mass, (8) follows from Proposition 2.2. Suppose that μ is a point mass and ν is arbitrary. In this case, $\mu \otimes \nu$ is the unique coupling of μ and ν so that

$$\begin{aligned}
\mathsf{W}_2(\mu,\nu)^2 &= M_2(\mu) + M_2(\nu) - 2\mathbb{E}_\mu[X]\mathbb{E}_\nu[Y] \\
\mathsf{W}_2(-(\mathrm{id})_\sharp\mu,\nu)^2 &= M_2(\mu) + M_2(\nu) + 2\mathbb{E}_\mu[X]\mathbb{E}_\nu[Y] \\
\mathsf{IGW}(\mu,\nu)^2 &= M_2(\mu)^2 + M_2(\nu)^2 - 2\left(\mathbb{E}_\mu[X]\mathbb{E}_\nu[Y]\right)^2.
\end{aligned}$$

If $\mathbb{E}_\mu[X]\mathbb{E}_\nu[Y] \geq 0$, $\mathsf{W}_2(\mu,\nu) \leq \mathsf{W}_2((-\mathrm{id})_\sharp\mu,\nu)$ and

$$0 \leq \int (x-y)^2 d\mu\otimes\nu(x,y) = M_2(\mu) + M_2(\nu) - 2\mathbb{E}_\mu[X]\mathbb{E}_\nu[Y]$$

so that $0 \leq 2\mathbb{E}_\mu[X]\mathbb{E}_\nu[Y] \leq M_2(\mu) + M_2(\nu)$ whereby

$$\begin{aligned}
\mathsf{IGW}(\mu,\nu)^2 &\geq M_2(\mu)^2 + M_2(\nu)^2 - \mathbb{E}_\mu[X]\mathbb{E}_\nu[Y](M_2(\mu)+M_2(\nu)) \\
&= M_2(\mu)^2 + M_2(\nu)^2 + \frac{1}{2}\left(\mathsf{W}_2(\mu,\nu)^2 - M_2(\mu) - M_2(\nu)\right)(M_2(\mu)+M_2(\nu)) \\
&= \underbrace{M_2(\mu)^2 + M_2(\nu)^2 - \frac{1}{2}\left(M_2(\mu)+M_2(\nu)\right)^2}_{=\frac{1}{2}(M_2(\mu)-M_2(\nu))^2\geq 0} + \frac{1}{2}\mathsf{W}_2(\mu,\nu)^2(M_2(\mu)+M_2(\nu))
\end{aligned}$$

proving the claim. If $\mathbb{E}_\mu[X]\mathbb{E}_\nu[Y] \leq 0$, then we work with the expression for $\mathsf{W}_2((-\mathrm{id})_\sharp\mu,\nu)$ and the argument follows analogously.

With (8) in hand, if $\mathsf{IGW}(\mu_n,\mu) \to 0$, then

$$\frac{1}{2}(M_2(\mu_n)+M_2(\mu))\min\left\{\mathsf{W}_2(\mu_n,\mu)^2, \mathsf{W}_2((-\mathrm{id})_\sharp\mu_n,\mu)^2\right\} \to 0$$

so that either $M_2(\mu_n)+M_2(\mu) \to 0$ or $\min\left\{\mathsf{W}_2(\mu_n,\mu)^2, \mathsf{W}_2((-\mathrm{id})_\sharp\mu_n,\mu)^2\right\} \to 0$. To see this, note that if $M_2(\mu) > 0$ the minimum of Wasserstein distances converges to 0. If $M_2(\mu) = 0$, μ is a point mass at 0 so that the Wasserstein distances are both simply $M_2(\mu_n)^{1/2}$. Therefore, the product converges to 0 if and only if $M_2(\mu_n) \to 0$ and so both sequences converge to 0.

If the second moments converge to 0, μ_n converges weakly to a point mass at 0 and μ is a point mass at 0. In the other case, we see that $\mathsf{W}_2\left((s_n\,\mathrm{id})_\sharp\mu_n,\mu\right) \to 0$ proving the claim. The final assertion is a direct consequence of Proposition 2.3. □

B.4 Proof of Lemma A.4

Let $a \in \mathbb{R}^{d_y}$ be arbitrary and $(a_n)_{n\in\mathbb{N}} \subset \mathbb{R}^{d_y}$ be a sequence converging to a. It is easy to see that the functions $g_n : x \in \mathbb{R}^{d_x} \mapsto a_n^\intercal \boldsymbol{\Delta} x$ converge uniformly on any compact set to $g : x \in \mathbb{R}^{d_x} \mapsto a^\intercal\boldsymbol{\Delta}x$ so that Lemma 5.2.1 in [89] yields that $((a_n)^\intercal\boldsymbol{\Delta})_\sharp\mu \xrightarrow{w} (a^\intercal\boldsymbol{\Delta})_\sharp\mu$. Similarly, we obtain that $(-\mathrm{id})_\sharp((a_n)^\intercal\boldsymbol{\Delta})_\sharp\mu \xrightarrow{w} (-\mathrm{id})_\sharp(a^\intercal\boldsymbol{\Delta})_\sharp\mu$ and that $((a_n)^\intercal)_\sharp\nu \xrightarrow{w} (a^\intercal)_\sharp\nu$. Furthermore,

$$M_2(((a_n)^\intercal\boldsymbol{\Delta})_\sharp\mu) = a_n^\intercal\boldsymbol{\Delta}\int xx^\intercal d\mu(x)\boldsymbol{\Delta}^\intercal a_n \to a^\intercal\boldsymbol{\Delta}\int xx^\intercal d\mu(x)\boldsymbol{\Delta}^\intercal a = M_2((a^\intercal\boldsymbol{\Delta})_\sharp\mu),$$

and the second moments of the other distributions converge similarly. Conclude that $\mathsf{W}_2(((a_n)^\intercal\boldsymbol{\Delta})_\sharp\mu, ((a_n)^\intercal)_\sharp\nu) \to \mathsf{W}_2((a^\intercal\boldsymbol{\Delta})_\sharp\mu, (a^\intercal)_\sharp\nu)$ and, by the same logic, $\mathsf{W}_2((-\mathrm{id})_\sharp((a_n)^\intercal\boldsymbol{\Delta})_\sharp\mu, ((a_n)^\intercal)_\sharp\nu) \to \mathsf{W}_2((-\mathrm{id})_\sharp(a^\intercal\boldsymbol{\Delta})_\sharp\mu, (a^\intercal)_\sharp\nu)$. As the choice of a and sequence was arbitrary, the result follows. □

B.5 Proof of Lemma A.5

Recall that $\mathsf{IGW}(\mu,\nu)^2 = \mathsf{F}_1(\mu,\nu) + \mathsf{F}_2(\mu,\nu)$, where

$$\begin{aligned}
\mathsf{F}_1(\mu,\nu) &= \int |\langle x,x'\rangle|^2 d\mu\otimes\mu(x,x') + \int |\langle y,y'\rangle|^2 d\nu\otimes\nu(y,y'), \\
\mathsf{F}_2(\mu,\nu) &= \inf_{\pi\in\Pi(\mu,\nu)} -2\int \langle x,x'\rangle\langle y,y'\rangle\, d\pi\otimes\pi(x,y,x',y').
\end{aligned}$$

We begin with Lipschitz continuity in θ. For any $\theta, \vartheta \in \mathbb{S}^{d_y-1}$, we have that

$$\begin{aligned}
&|\mathsf{IGW}\big((\theta^\intercal \boldsymbol{\Delta})_\sharp \mu, (\theta^\intercal)_\sharp \nu\big)^2 - \mathsf{IGW}\big((\vartheta^\intercal \boldsymbol{\Delta})_\sharp \mu, (\vartheta^\intercal)_\sharp \nu\big)^2 \\
&\leq |\mathsf{F}_1((\theta^\intercal \boldsymbol{\Delta})_\sharp \mu, (\theta^\intercal)_\sharp \nu) - \mathsf{F}_1((\vartheta^\intercal \boldsymbol{\Delta})_\sharp \mu, (\vartheta^\intercal)_\sharp \nu)| \\
&+ |\mathsf{F}_2((\theta^\intercal \boldsymbol{\Delta})_\sharp \mu, (\theta^\intercal)_\sharp \nu) - \mathsf{F}_2((\vartheta^\intercal \boldsymbol{\Delta})_\sharp \mu, (\vartheta^\intercal)_\sharp \nu)| \,.
\end{aligned}$$

We bound the two terms separately starting with the second,

$$\begin{aligned}
&|\mathsf{F}_2((\theta^\intercal \boldsymbol{\Delta})_\sharp \mu, (\theta^\intercal)_\sharp \nu) - \mathsf{F}_2((\vartheta^\intercal \boldsymbol{\Delta})_\sharp \mu, (\vartheta^\intercal)_\sharp \nu)| \\
&= \left| \inf_{\pi \in \Pi(\mu,\nu)} -2 \left(\int \theta^\intercal \boldsymbol{\Delta} x y^\intercal \theta d\pi(x,y) \right)^2 - \inf_{\pi \in \Pi(\mu,\nu)} -2 \left(\int \vartheta^\intercal \boldsymbol{\Delta} x y^\intercal \vartheta d\pi(x,y) \right)^2 \right| \\
&\leq 2 \sup_{\pi \in \Pi(\mu,\nu)} \left| \left(\int \theta^\intercal \boldsymbol{\Delta} x y^\intercal \theta d\pi(x,y) \right)^2 - \left(\int \vartheta^\intercal \boldsymbol{\Delta} x y^\intercal \vartheta d\pi(x,y) \right)^2 \right| \\
&= 2 \sup_{\pi \in \Pi(\mu,\nu)} \left| \int \theta^\intercal \boldsymbol{\Delta} x y^\intercal \theta d\pi(x,y) - \int \vartheta^\intercal \boldsymbol{\Delta} x y^\intercal \vartheta d\pi(x,y) \right| \times \\
&\qquad\qquad \left| \int \theta^\intercal \boldsymbol{\Delta} x y^\intercal \theta d\pi(x,y) + \int \vartheta^\intercal \boldsymbol{\Delta} x y^\intercal \vartheta d\pi(x,y) \right| .
\end{aligned}$$

Observe that

$$\theta^\intercal \boldsymbol{\Delta} x y^\intercal \theta - \vartheta^\intercal \boldsymbol{\Delta} x y^\intercal \vartheta = (\theta - \vartheta)^\intercal \boldsymbol{\Delta} x y^\intercal \theta + \vartheta^\intercal \boldsymbol{\Delta} x y^\intercal (\theta - \vartheta),$$

so that

$$\begin{aligned}
&\left| \int \theta^\intercal \boldsymbol{\Delta} x y^\intercal \theta - \vartheta^\intercal \boldsymbol{\Delta} x y^\intercal \vartheta d\pi(x,y) \right| \\
&\leq \int |(\theta - \vartheta)^\intercal \boldsymbol{\Delta} x y^\intercal \theta| \, d\pi(x,y) + \int |\vartheta^\intercal \boldsymbol{\Delta} x y^\intercal (\theta - \vartheta)| \, d\pi(x,y) \\
&\leq \int \|\theta - \vartheta\| \|x\| \|y\| d\pi(x,y) + \int \|x\| \|y\| \|\theta - \vartheta\| d\pi(x,y) \\
&\leq 2\|\theta - \vartheta\| \sqrt{M_2(\mu) M_2(\nu)},
\end{aligned}$$

Applying a similar approach, we have that

$$\left| \int \theta^\intercal \boldsymbol{\Delta} x y^\intercal \theta d\pi(x,y) + \int \vartheta^\intercal \boldsymbol{\Delta} x y^\intercal \vartheta d\pi(x,y) \right| \leq 2\sqrt{M_2(\mu) M_2(\nu)}.$$

Overall,

$$|\mathsf{F}_2((\theta^\intercal \boldsymbol{\Delta})_\sharp \mu, (\theta^\intercal)_\sharp \nu) - \mathsf{F}_2((\vartheta^\intercal \boldsymbol{\Delta})_\sharp \mu, (\vartheta^\intercal)_\sharp \nu)| \leq 8 M_2(\mu) M_2(\nu) \|\theta - \vartheta\|.$$

We now bound the remaining term using a similar approach,

$$\begin{aligned}
&|\mathsf{F}_1((\theta^\intercal \boldsymbol{\Delta})_\sharp \mu, (\theta^\intercal)_\sharp \nu) - \mathsf{F}_1((\vartheta^\intercal \boldsymbol{\Delta})_\sharp \mu, (\vartheta^\intercal)_\sharp \nu)| \\
&\qquad = \left| \int \left(\theta^\intercal \boldsymbol{\Delta} x' x^\intercal \boldsymbol{\Delta}^\intercal \theta\right)^2 d\mu \otimes \mu(x,x') - \int \left(\vartheta^\intercal \boldsymbol{\Delta} x' x^\intercal \boldsymbol{\Delta}^\intercal \vartheta\right)^2 d\mu \otimes \mu(x,x') \right| \\
&\qquad + \left| \int \left(\theta^\intercal y' y^\intercal \theta\right)^2 d\nu \otimes \nu(y,y') - \int \left(\vartheta^\intercal y' y^\intercal \vartheta\right)^2 d\nu \otimes \nu(y,y') \right| .
\end{aligned}$$

Observe that

$$\begin{aligned}
&\left| \left(\theta^\intercal \boldsymbol{\Delta} x' x^\intercal \boldsymbol{\Delta}^\intercal \theta\right)^2 - \left(\vartheta^\intercal \boldsymbol{\Delta} x' x^\intercal \boldsymbol{\Delta}^\intercal \vartheta\right)^2 \right| \\
&= \left| \left(\theta^\intercal \boldsymbol{\Delta} x' x^\intercal \boldsymbol{\Delta}^\intercal \theta\right)^2 - \left(\vartheta^\intercal \boldsymbol{\Delta} x' x^\intercal \boldsymbol{\Delta}^\intercal \vartheta\right)^2 \right| \\
&= |(\theta - \vartheta)^\intercal \boldsymbol{\Delta} x' x^\intercal \boldsymbol{\Delta}^\intercal \theta + \vartheta^\intercal \boldsymbol{\Delta} x' x^\intercal \boldsymbol{\Delta}^\intercal (\theta - \vartheta)| \, |\theta^\intercal \boldsymbol{\Delta} x' x^\intercal \boldsymbol{\Delta}^\intercal \theta + \vartheta^\intercal \boldsymbol{\Delta} x' x^\intercal \boldsymbol{\Delta}^\intercal \vartheta| \,,
\end{aligned}$$

so that $\left| \left(\theta^\intercal \boldsymbol{\Delta} x' x^\intercal \boldsymbol{\Delta}^\intercal \theta\right)^2 - \left(\vartheta^\intercal \boldsymbol{\Delta} x' x^\intercal \boldsymbol{\Delta}^\intercal \vartheta\right)^2 \right| \leq \|x\|^2 \|x'\|^2 \left(\|\theta\|^2 + \|\vartheta\|^2\right) \left(\|\theta\| + \|\vartheta\|\right) \|\theta - \vartheta\|$,
and $\left| \int \left(\theta^\intercal \boldsymbol{\Delta} x' x^\intercal \boldsymbol{\Delta}^\intercal \theta\right)^2 d\mu \otimes \mu(x,x') - \int \left(\vartheta^\intercal \boldsymbol{\Delta} x' x^\intercal \boldsymbol{\Delta}^\intercal \vartheta\right)^2 d\mu \otimes \mu(x,x') \right| \leq 4\|\theta - \vartheta\| M_2(\mu)^2$.
Similarly, $\left| \int \left(\theta^\intercal y' y^\intercal \theta\right)^2 d\nu \otimes \nu(y,y') - \int \left(\vartheta^\intercal y' y^\intercal \vartheta\right)^2 d\nu \otimes \nu(y,y') \right| \leq 4\|\theta - \vartheta\| M_2(\nu)^2$, so that,

$$|\mathsf{F}_1((\theta^\intercal \boldsymbol{\Delta})_\sharp \mu, (\theta^\intercal)_\sharp \nu) - \mathsf{F}_1((\vartheta^\intercal \boldsymbol{\Delta})_\sharp \mu, (\vartheta^\intercal)_\sharp \nu)| \leq 4\|\theta - \vartheta\| (M_2(\mu)^2 + M_2(\nu)^2).$$

By compiling these results, we obtain the desired Lipschitz constant.

For the second part of the statement, we have similarly that, for any $\mathbf{\Delta}_1, \mathbf{\Delta}_2 \in \mathbb{R}^{d_y \times d_x}$,

$$\begin{aligned}
&|\mathsf{IGW}\big((\theta^\intercal \mathbf{\Delta}_1)_\sharp \mu, (\theta^\intercal)_\sharp \nu\big)^2 - \mathsf{IGW}\big((\theta^\intercal \mathbf{\Delta}_2)_\sharp \mu, (\theta^\intercal)_\sharp \nu\big)^2 \\
&\leq |\mathsf{F}_1((\theta^\intercal \mathbf{\Delta}_1)_\sharp \mu, (\theta^\intercal)_\sharp \nu) - \mathsf{F}_1((\theta^\intercal \mathbf{\Delta}_2)_\sharp \mu, (\theta^\intercal)_\sharp \nu)| \\
&+ |\mathsf{F}_2((\theta^\intercal \mathbf{\Delta}_1)_\sharp \mu, (\theta^\intercal)_\sharp \nu) - \mathsf{F}_2((\theta^\intercal \mathbf{\Delta}_2)_\sharp \mu, (\theta^\intercal)_\sharp \nu)| \,.
\end{aligned}$$

Following a similar approach, we can bound

$$\begin{aligned}
&|\mathsf{F}_2((\theta^\intercal \mathbf{\Delta}_1)_\sharp \mu, (\theta^\intercal)_\sharp \nu) - \mathsf{F}_2((\theta^\intercal \mathbf{\Delta}_2)_\sharp \mu, (\theta^\intercal)_\sharp \nu)| \\
&\leq 2 \sup_{\pi \in \Pi(\mu,\nu)} \left| \left(\int \theta^\intercal \mathbf{\Delta}_1 x y^\intercal \theta d\pi(x,y) \right)^2 - \left(\int \theta^\intercal \mathbf{\Delta}_2 x y^\intercal \theta d\pi(x,y) \right)^2 \right| \\
&= 2 \sup_{\pi \in \Pi(\mu,\nu)} \left| \int \theta^\intercal (\mathbf{\Delta}_1 - \mathbf{\Delta}_2) x y^\intercal \theta d\pi(x,y) \right| \left| \int \theta^\intercal (\mathbf{\Delta}_1 + \mathbf{\Delta}_2) x y^\intercal \theta d\pi(x,y) \right| \\
&\leq 2 \|\mathbf{\Delta}_1 - \mathbf{\Delta}_2\|_{\mathrm{op}} \|\mathbf{\Delta}_1 + \mathbf{\Delta}_2\|_{\mathrm{op}} M_2(\mu) M_2(\nu),
\end{aligned}$$

and,

$$\begin{aligned}
&|\mathsf{F}_1((\theta^\intercal \mathbf{\Delta}_1)_\sharp \mu, (\theta^\intercal)_\sharp \nu) - \mathsf{F}_1((\theta^\intercal \mathbf{\Delta}_2)_\sharp \mu, (\theta^\intercal)_\sharp \nu)| \\
&= \left| \int \left(\theta^\intercal \mathbf{\Delta}_1 x' x^\intercal \mathbf{\Delta}_1^\intercal \theta\right)^2 - \left(\theta^\intercal \mathbf{\Delta}_2 x' x^\intercal \mathbf{\Delta}_2^\intercal \theta\right)^2 d\mu \otimes \mu(x,x') \right| \\
&= \left| \int \left(\theta^\intercal \mathbf{\Delta}_1 x' x^\intercal \mathbf{\Delta}_1^\intercal \theta + \theta^\intercal \mathbf{\Delta}_2 x' x^\intercal \mathbf{\Delta}_2^\intercal \theta\right) \left(\theta^\intercal \mathbf{\Delta}_1 x' x^\intercal \mathbf{\Delta}_1^\intercal \theta - \theta^\intercal \mathbf{\Delta}_2 x' x^\intercal \mathbf{\Delta}_2^\intercal \theta\right) d\mu \otimes \mu(x,x') \right|,
\end{aligned}$$

where we observe that

$$\begin{aligned}
&|\left(\theta^\intercal \mathbf{\Delta}_1 x' x^\intercal \mathbf{\Delta}_1^\intercal \theta + \theta^\intercal \mathbf{\Delta}_2 x' x^\intercal \mathbf{\Delta}_2^\intercal \theta\right) \left(\theta^\intercal \mathbf{\Delta}_1 x' x^\intercal \mathbf{\Delta}_1^\intercal \theta - \theta^\intercal \mathbf{\Delta}_2 x' x^\intercal \mathbf{\Delta}_2^\intercal \theta\right)| \\
&\leq \left(\|\mathbf{\Delta}_1\|_{\mathrm{op}}^2 + \|\mathbf{\Delta}_2\|_{\mathrm{op}}^2\right) \|x\|^2 \|x'\|^2 \|\mathbf{\Delta}_1 + \mathbf{\Delta}_2\|_{\mathrm{op}} \|\mathbf{\Delta}_1 - \mathbf{\Delta}_2\|_{\mathrm{op}},
\end{aligned}$$

so that, when $\mathbf{\Delta}_1, \mathbf{\Delta}_2 \in \mathrm{St}(d_x, d_y)$,

$$|\mathsf{F}_1((\theta^\intercal \mathbf{\Delta}_1)_\sharp \mu, (\theta^\intercal)_\sharp \nu) - \mathsf{F}_1((\theta^\intercal \mathbf{\Delta}_2)_\sharp \mu, (\theta^\intercal)_\sharp \nu)| \leq 2 \left(\|\mathbf{\Delta}_1\|_{\mathrm{op}}^2 + \|\mathbf{\Delta}_2\|_{\mathrm{op}}^2\right) M_2(\mu)^2 \|\mathbf{\Delta}_1 - \mathbf{\Delta}_2\|_{\mathrm{op}}.$$

In sum, for $\mathbf{\Delta}_1, \mathbf{\Delta}_2 \in \mathrm{St}(d_x, d_y)$, we readily obtain that

$$\begin{aligned}
&|\mathsf{IGW}\big((\theta^\intercal \mathbf{\Delta}_1)_\sharp \mu, (\theta^\intercal)_\sharp \nu\big)^2 - \mathsf{IGW}\big((\theta^\intercal \mathbf{\Delta}_2)_\sharp \mu, (\theta^\intercal)_\sharp \nu\big)^2 \\
&\leq (4M_2(\mu)^2 + 4M_2(\mu)M_2(\nu)) \|\mathbf{\Delta}_1 - \mathbf{\Delta}_2\|_{\mathrm{op}} \leq (4M_2(\mu)^2 + 4M_2(\mu)M_2(\nu)) \|\mathbf{\Delta}_1 - \mathbf{\Delta}_2\|_{\mathrm{F}}.
\end{aligned}$$

□

C Formula for IGW between Gaussians

To derive the claimed formula, we first observe that, for any $\mathbf{\Delta} \in \mathrm{St}(d_x, d_y)$ and $\theta \in \mathbb{S}^{d_y - 1}$,

$$\mathsf{IGW}\left((\theta^\intercal \mathbf{\Delta})_\sharp \mu, (\theta^\intercal)_\sharp \nu\right)^2 = (\theta^\intercal \mathbf{\Delta} \mathbf{\Sigma}_\mu \mathbf{\Delta}^\intercal \theta)^2 + (\theta^\intercal \mathbf{\Sigma}_\nu \theta)^2 - 2\theta^\intercal \mathbf{\Delta} \mathbf{\Sigma}_\mu \mathbf{\Delta}^\intercal \theta \theta^\intercal \mathbf{\Sigma}_\nu \theta,$$

as follows from Corollary 3.2 in [35]. To integrate this expression over $\mathbb{S}^{d_y - 1}$, we underscore that each term is of the form $\theta^\intercal \mathbf{A} \theta \theta^\intercal \mathbf{B} \theta$ for some PSD matrices $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{d_y \times d_y}$. As $\int_{\mathbb{S}^{d_y-1}} \theta^\intercal \mathbf{A} \theta \theta^\intercal \mathbf{B} \theta \, d\sigma_{d_y}(\theta) = \sum_{i,j,k,l=1}^{d_y} A_{ij} B_{kl} \int_{\mathbb{S}^{d_y-1}} \theta_i \theta_j \theta_k \theta_l \, d\sigma_{d_y}(\theta)$, we apply the following formula for integrating polynomials over the sphere [90, p.447],

$$\int_{\mathbb{S}^{d_y-1}} \theta_i \theta_j \theta_k \theta_l \, d\sigma_{d_y} = \begin{cases} \frac{2\Gamma(1/2)^{d_y-1}\Gamma(5/2)}{\Gamma(2+d_y/2)|\mathbb{S}^{d_y-1}|} = \frac{3}{d_y(d_y+2)} & \text{if } i=j=k=l \\ \frac{2\Gamma(1/2)^{d_y-2}\Gamma(3/2)^2}{\Gamma(2+d_y/2)|\mathbb{S}^{d_y-1}|} = \frac{1}{d_y(d_y+2)} & \text{if } i=j\neq k=l \text{ up to permuting } (i,j,k,l), \\ 0 & \text{otherwise,} \end{cases}$$

whereby,

$$\begin{aligned}
&\int_{\mathbb{S}^{d_y-1}} \theta^\intercal \mathbf{A}\theta\theta^\intercal \mathbf{B}\theta \, d\sigma_{d_y}(\theta) \\
&\qquad = \frac{3}{d_y(d_y+2)} \sum_{i=1}^{d_y} A_{ii}B_{ii} + \frac{1}{d_y(d_y+2)} \left(\sum_{i=1}^{d_y} \sum_{\substack{j=1 \\ j\neq i}}^{d_y} A_{ii}B_{jj} + A_{ij}B_{ij} + A_{ij}B_{ji} \right) \\
&\qquad = \frac{1}{d_y(d_y+2)} \left(\sum_{i,j=1}^{d_y} A_{ii}B_{jj} + A_{ij}B_{ij} + A_{ij}B_{ji} \right) \\
&\qquad = \frac{1}{d_y(d_y+2)} \big(\mathrm{Tr}(\mathbf{A})\mathrm{Tr}(\mathbf{B}) + \mathrm{Tr}(\mathbf{A}^\intercal\mathbf{B}) + \mathrm{Tr}(\mathbf{AB}) \big).
\end{aligned}$$

Using the above formula and accounting for the minimization over $\mathrm{St}(d_x, d_y)$, we have that

$$\begin{aligned}
\underline{\mathsf{IGW}}(\mu,\nu)^2 = &\inf_{\boldsymbol{\Delta}\in\mathrm{St}(d_x,d_y)} \frac{1}{d_y(d_y+2)} \big(\mathrm{Tr}(\boldsymbol{\Delta}\boldsymbol{\Sigma}_\mu\boldsymbol{\Delta}^\intercal)^2 + 2\mathrm{Tr}((\boldsymbol{\Delta}\boldsymbol{\Sigma}_\mu\boldsymbol{\Delta}^\intercal)^2) + \mathrm{Tr}(\boldsymbol{\Sigma}_\nu)^2 + 2\mathrm{Tr}(\boldsymbol{\Sigma}_\nu^2) \big) \\
&- \frac{2}{d_y(d_y+2)} \left(\mathrm{Tr}(\boldsymbol{\Delta}\boldsymbol{\Sigma}_\mu\boldsymbol{\Delta}^\intercal)\mathrm{Tr}(\boldsymbol{\Sigma}_\nu) + 2\mathrm{Tr}(\boldsymbol{\Delta}\boldsymbol{\Sigma}_\mu\boldsymbol{\Delta}^\intercal\boldsymbol{\Sigma}_\nu) \right) \\
&= \frac{1}{d_y(d_y+2)} \big(\mathrm{Tr}(\boldsymbol{\Sigma}_\mu)^2 + 2\mathrm{Tr}(\boldsymbol{\Sigma}_\mu^2) + \mathrm{Tr}(\boldsymbol{\Sigma}_\nu)^2 + 2\mathrm{Tr}(\boldsymbol{\Sigma}_\nu^2) - 2\mathrm{Tr}(\boldsymbol{\Sigma}_\mu)\mathrm{Tr}(\boldsymbol{\Sigma}_\nu) \big) \\
&- \frac{4}{d_y(d_y+2)} \sup_{\boldsymbol{\Delta}\in\mathrm{St}(d_x,d_y)} \mathrm{Tr}(\boldsymbol{\Delta}\boldsymbol{\Sigma}_\mu\boldsymbol{\Delta}^\intercal\boldsymbol{\Sigma}_\nu),
\end{aligned}$$

where we have used the cyclic property of the trace and the fact that $\boldsymbol{\Delta}^\intercal\boldsymbol{\Delta} = \mathbf{I}_{d_x}$. Now, by Von Neumann's trace inequality [91], $\mathrm{Tr}(\boldsymbol{\Delta}\boldsymbol{\Sigma}_\mu\boldsymbol{\Delta}^\intercal\boldsymbol{\Sigma}_\nu) \leq \sum_{i=1}^{d_x} \lambda_i(\boldsymbol{\Delta}\boldsymbol{\Sigma}_\mu\boldsymbol{\Delta}^\intercal)\lambda_i(\boldsymbol{\Sigma}_\nu) = \sum_{i=1}^{d_x} \lambda_i(\boldsymbol{\Sigma}_\mu)\lambda_i(\boldsymbol{\Sigma}_\nu)$. Diagonalizing both covariance matrices as $\boldsymbol{\Sigma}_\mu = \mathbf{U}\mathbf{D}_\mu\mathbf{U}^\intercal$ and $\boldsymbol{\Sigma}_\nu = \mathbf{V}\mathbf{D}_\nu\mathbf{V}^\intercal$, we see that the trace inequality is saturated for $\boldsymbol{\Delta} = \mathbf{V}\mathbf{J}\mathbf{U}^\intercal$ where $\mathbf{J} = [\mathbf{I}_{d_x}\ \mathbf{0}]^\intercal$. Inserting this formula in the above display proves the assertion.

D Proof of Equivalence for Distributions on Lattices

Let ν be supported in a full rank lattice $\Lambda \coloneqq \{\mathbf{B}x : x \in \mathbb{Z}^d\}$ where $\mathbf{B} \in \mathbb{R}^{d\times d}$ is invertible and that $\underline{\mathsf{IGW}}(\mu,\nu) = 0$. We remove the full-rank assumption at the end of the proof. Following the proof of Proposition 4.4, there exists $\boldsymbol{\Delta} \in \mathrm{St}(d,d)$ and a set $C \subseteq \mathbb{R}^d$ satisfying the property that if $x \in C$, then $tx \in C$ for each $t \in \mathbb{R}$ for which $\Phi_{(\boldsymbol{\Delta})_\sharp\mu} = \Phi_\nu$ on C, $\Phi_{(\boldsymbol{\Delta})_\sharp\mu} = \overline{\Phi_\nu}$ on C^c. One can check $C^c \cup \{0\}$ also satisfies the property that if $x \in C^c \cup \{0\}$, then $tx \in C^c \cup \{0\}$ for each $t \in \mathbb{R}$.

If either $C = \emptyset$ or $C^c = \emptyset$, the conclusion is immediate. Suppose that $C \neq \emptyset$ and $C^c \neq \emptyset$ and fix $a \in C$ for which the imaginary part of $\Phi_\nu(a)$ is not zero. If such an a does not exist, $\Phi_\nu(x)$ is real for each $x \in C$ and $\Phi_\nu(x) = \overline{\Phi_\nu(x)}$ for each $x \in C$ whereby $\Phi_{(\boldsymbol{\Delta})_\sharp\mu} = \overline{\Phi_\nu}$ and the proof is complete. The existence of such a also implies C contains a cone with an open base (i.e. there exists U open in $\mathbb{S}^{d-1}$ such that $\{tx : t \in \mathbb{R}, x \in U\} \subseteq C$), as in proof of Proposition 4.4. By a similar argument, $C^c \cup \{0\}$ also contains such a cone provided that there is a point in C^c where Φ_ν has nonzero imaginary part. If no such point exists, then the characteristic functions coincide and the proof is done.

We now prove that there exists a $b \in C^c$ with $\Phi_\nu(b) = 1$. To this end, let Λ^* be the dual lattice to Λ; $\Lambda^* = \{y \in \mathrm{span}(\mathbf{B}\mathbb{Z}^d) : y^\intercal x \in \mathbb{Z}, \text{ for all } x \in \Lambda\} = \{y \in \mathrm{span}(\mathbf{B}\mathbb{Z}^d) : y^\intercal\mathbf{B}z \in \mathbb{Z}, \text{ for all } z \in \mathbb{Z}^d\}$ since Λ is full rank. Thus $\Lambda^* = \mathbf{B}^{-\intercal}\mathbb{Z}^d$ is also a nonempty full-rank lattice and, for each $b \in 2\pi\Lambda^*$,

$$\Phi_\nu(b) = \int_\Lambda e^{ib^\intercal x} d\nu(x) = \nu(\Lambda) = 1,$$

since $b^\intercal x \in 2\pi\mathbb{Z}$ for each $x \in \Lambda$. It remains to show that $\Lambda^* \cap C^c \neq \emptyset$. To this end, recall that $C^c \cup \{0\}$ contains a cone K. Since $\mathbf{B}$ is invertible $\mathbf{B}^\intercal K$ is a nonempty open cone in $\mathbb{R}^d$

and so there exists $q \in \mathbf{B}^\intercal K \cap \mathbb{Q}^d$. Thus, there exists $N \in \mathbb{N}$ such that $Nq \in \mathbb{Z}^d \cap B^\intercal K$ and so $\lambda = B^{-\intercal} Nq \in \Lambda^\star \cap K$ which is nonzero and λ is an interior point of K. Thus $k\lambda \in \Lambda^*$ for any $k \in \mathbb{Z}$ so that $2\pi k\lambda \in 2\pi\Lambda^* \cap C^c$ for each $k \in \mathbb{Z}\backslash\{0\}$. Since $\lambda \in (K \cup (-K))\backslash\{0\}$ lies in the interior of the cone, there exists an open ball of radius $r > 0$ centered at $2\pi\lambda$, denoted $B(2\pi\lambda, r)$, which is contained in C^c. Thanks to the scaling property, for any positive integer k, $2\pi k\lambda \in C^c$, and $B(2\pi k\lambda, kr) \subset C^c$ as if $\|l - 2\pi k\lambda\| < kr$, then $\|l/k - 2\pi\lambda\| < r$ so that $l/k \in B(2\pi\lambda, r) \subset C^c$. Letting k be a positive integer greater than $\|a\|/r$ and setting $b = 2\pi k\lambda$ it holds that $a \in C, b \in C^c, a + b \in C^c$, $\Phi_\nu(b) = 1$, and $\Phi_\nu(a)$ has non-zero imaginary part.

With this, we consider the Gram matrix of $\Phi_{(\boldsymbol{\Delta})_\sharp\mu}$ at the points $0, a, a + b$, given by

$$\begin{pmatrix} \Phi_{(\boldsymbol{\Delta})_\sharp\mu}(0) & \Phi_{(\boldsymbol{\Delta})_\sharp\mu}(-a) & \Phi_{(\boldsymbol{\Delta})_\sharp\mu}(-(a+b)) \\ \Phi_{(\boldsymbol{\Delta})_\sharp\mu}(a) & \Phi_{(\boldsymbol{\Delta})_\sharp\mu}(0) & \Phi_{(\boldsymbol{\Delta})_\sharp\mu}(-b) \\ \Phi_{(\boldsymbol{\Delta})_\sharp\mu}(a+b) & \Phi_{(\boldsymbol{\Delta})_\sharp\mu}(b) & \Phi_{(\boldsymbol{\Delta})_\sharp\mu}(0) \end{pmatrix} = \begin{pmatrix} 1 & \Phi_\nu(-a) & \overline{\Phi_\nu(-(a+b))} \\ \Phi_\nu(a) & 1 & 1 \\ \overline{\Phi_\nu(a+b)} & 1 & 1 \end{pmatrix},$$

where we note that

$$\Phi_\nu(a+b) = \int_\Lambda e^{ia^\intercal x} e^{ib^\intercal x} d\nu(x) = \int_\Lambda e^{ia^\intercal x} d\nu(x) = \Phi_\nu(a)$$

since $b^\intercal x \in 2\pi\mathbb{Z}$ for each $x \in \Lambda$ by construction. Similarly, $\Phi_\nu(-(a+b)) = \Phi_\nu(-a)$. It follows that the determinant of the above matrix is

$$\begin{aligned} &1 + \Phi_\nu(-a)\overline{\Phi_\nu(a)} + \Phi_\nu(a)\overline{\Phi_\nu(-a)} - \overline{\Phi_\nu(-a)\Phi_\nu(a)} - \Phi_\nu(-a)\Phi_\nu(a) - 1 \\ &= \Phi_\nu(-a)^2 + \Phi_\nu(a)^2 - 2\Phi_\nu(-a)\Phi_\nu(a) = \left(\Phi_\nu(-a) - \Phi_\nu(a)\right)^2, \end{aligned}$$

where we have used the fact that $\Phi_\nu(-t) = \overline{\Phi_\nu(t)}$ for any $t \in \mathbb{R}^d$. Since $\Phi_\nu(a) = c + di$ with $d \neq 0$ by construction, $\left(\Phi_\nu(-a) - \Phi_\nu(a)\right)^2 = (-2di)^2 = -4d^2 < 0$. However, Bochner's theorem (cf. e.g., Theorem 7.13.1 [92]) asserts that the characteristic function of a measure must be positive definite, i.e., each of its Gram matrices are positive semidefinite. As the above Gram matrix is not positive semidefinite, $\Phi_{(\boldsymbol{\Delta})_\sharp\mu}$ is not a characteristic function unless $\Phi_{(\boldsymbol{\Delta})_\sharp\mu} = \Phi_\nu$ or $\Phi_{(\boldsymbol{\Delta})_\sharp\mu} = \overline{\Phi_\nu}$ on the entire space.

Now we return to the case when the lattice Λ does not have full rank. In this case, let the rank be $d' < d$, and basis $\mathbf{B} = [\mathbf{b}_1, \ldots, \mathbf{b}_{d'}]$ with linearly independent columns. Since ν is supported on $\Lambda = \{\mathbf{B}x : x \in \mathbb{Z}^{d'}\}$,

$$\Phi_\nu(x) = \int_\Lambda e^{i\langle x, y\rangle}\, d\nu(y) = \int_\Lambda e^{i\langle \mathrm{proj}_{\mathrm{span}(\mathbf{B})} x,\, y\rangle}\, d\nu(y) = \Phi_\nu\big(\mathrm{proj}_{\mathrm{span}(\mathbf{B})} x\big).$$

Therefore, the characteristic function Φ_ν is completely determined by its values on $\mathrm{span}(\mathbf{B})$. Since $\underline{\mathsf{IGW}}(\mu, \nu) = 0$, $\mathsf{IGW}((\theta^\intercal\Delta)_\sharp\mu, (\theta^\intercal)_\sharp\nu) = 0$ for almost every $\theta \in \mathbb{S}^{d_y - 1}$. In particular, $\theta^\intercal \boldsymbol{\Delta}\mathbf{R}_\mu\boldsymbol{\Delta}^\intercal\theta = \theta^\intercal\mathbf{R}_\nu\theta$ for every $\theta \in \mathbb{S}^{d_y-1}$ by continuity. Thus $\boldsymbol{\Delta}\mathbf{R}_\mu\boldsymbol{\Delta}^\intercal = \mathbf{R}_\nu$. Since ν is supported on Λ, if $b \in \mathrm{span}(\mathbf{B})^\perp$, $0 = \int_\Lambda (b^\intercal y)^2 d\nu(y) = b^\intercal\mathbf{R}_\nu b$ whereby, running the same argument in reverse, $(\boldsymbol{\Delta})_\sharp\mu$ is supported in $\mathrm{span}(\mathbb{B})$. Conclude that we can restrict the analysis to the subspace $\mathrm{span}(\mathbf{B})$.

Notice that for the general case, the dual lattice can be written as $\Lambda^* = \mathbf{B}(\mathbf{B}^\intercal\mathbf{B})^{-1}\mathbb{Z}^{d'}$, which is also a full rank lattice when restricted to $\mathrm{span}(\mathbf{B})$. The same argument as the full rank case then applies, with C and C^c being replaced by $C \cap \mathrm{span}(\mathbf{B})$ and $\mathrm{span}(\mathbf{B}) \setminus C$, and with the notion of interior and openness interpreted relative to $\mathrm{span}(\mathbf{B})$.

E Projected and Riemannian Subdifferentials

The Riemannian Clarke subdifferential of F at $\boldsymbol{\Delta} \in \mathrm{St}(d_x, d_y)$ is defined as [72, Equation 3.6]

$$\{\chi \in T_{\boldsymbol{\Delta}}(\mathrm{St}(d_x, d_y)) : \langle \chi, \psi\rangle_g \leq F'(\boldsymbol{\Delta}; \psi) \text{ for every } \psi \in T_{\boldsymbol{\Delta}}(\mathrm{St}(d_x, d_y))\},$$

where $T_{\mathbf{\Delta}}(\mathrm{St}(d_x, d_y))$ is the tangent space to $\mathrm{St}(d_x, d_y)$, the inner product is inherited from the metric tensor, and

$$F'(\mathbf{\Delta}; \psi) = \limsup_{\mathrm{St}(d_x,d_y) \ni \mathbf{\Delta}' \to \mathbf{\Delta}, t \downarrow 0} \frac{F(p_{\mathrm{St}(d_x,d_y)}(\mathbf{\Delta}' + t\psi)) - F(\mathbf{\Delta}')}{t}$$

where $p_{S_t(d_x,d_y)}$ denotes the projection onto the Stiefel manifold.

As noted in Section 4, the Riemannian Clarke subdifferential of F at $\mathbf{\Delta}$ coincides with the collection of projections of all elements of $\partial F(\mathbf{\Delta})$ onto the tangent space to $\mathrm{St}(d_x, d_y)$ at $\mathbf{\Delta}$,

$$P_{F,\mathbf{\Delta}} = \left\{ \mathbf{Z} - \frac{1}{2}\mathbf{\Delta}(\mathbf{\Delta}^\intercal \mathbf{Z} + \mathbf{Z}^\intercal \mathbf{\Delta}) : \mathbf{Z} \in \partial F(\mathbf{\Delta}) \right\}.$$

However, as noted in [72, Theorem 5.1], we generally only have that $\partial F|_{\mathrm{St}(d_x,d_y)} \subset P_{F,\mathbf{\Delta}}$. It remains, therefore, to prove the opposite inclusion. Since

$$F(\mathbf{\Delta}) = \frac{1}{R}\sum_{i=1}^{R} \left((\theta^{(i)})^\intercal \mathbf{\Delta} \mathbf{R}_\mu \mathbf{\Delta}^\intercal \theta^{(i)} \right)^2 + \left((\theta^{(i)})^\intercal \mathbf{R}_\nu \theta^{(i)} \right)^2 + \min_{j=1}^{|\mathcal{V}|} \ell_{j,\theta^{(i)}}(\mathbf{\Delta})$$

recalling the proof of Proposition 4.10 where $\mathcal{V} = \{\pi_1, \ldots, \pi_{|\mathcal{V}|}\}$ is the set of vertices of $\Pi(\mu, \nu)$ and $\ell_{j,\theta^{(i)}} = -2\left(\int (\theta^{(i)})^\intercal \mathbf{\Delta} x y^\intercal \theta^{(i)} d\pi_j(\mu, \nu)\right)^2$, we can write

$$\begin{aligned} F(\mathbf{\Delta}) &= \min_{j_1,\ldots,j_R=1}^{|\mathcal{V}|} \left\{ \frac{1}{R}\sum_{i=1}^{R} \left(\left((\theta^{(i)})^\intercal \mathbf{\Delta} \mathbf{R}_\mu \mathbf{\Delta}^\intercal \theta^{(i)} \right)^2 + \left((\theta^{(i)})^\intercal \mathbf{R}_\nu \theta^{(i)} \right)^2 + \ell_{j_i,\theta^{(i)}}(\mathbf{\Delta}) \right) \right\} \\ &\coloneqq \min_{j_1,\ldots,j_R=1}^{|\mathcal{V}|} F_{(j_1,\ldots,j_R)}(\mathbf{\Delta}). \end{aligned}$$

For a fixed $\mathbf{\Delta} \in \mathrm{St}(d_x, d_y)$, let $\bar{j}(\mathbf{\Delta}) \in \{1, \ldots, |\mathcal{V}|\}^R$ be a choice of indices achieving the minimum above, that is, $F(\mathbf{\Delta}) = F_{\bar{j}(\mathbf{\Delta})}(\mathbf{\Delta})$ and let $\bar{J}(\mathbf{\Delta})$ be the set of all such $\bar{j}(\mathbf{\Delta})$.

With these preparations in hand, let (U, φ) be a smooth chart of $\mathrm{St}(d_x, d_y)$ containing $\mathbf{\Delta}$ (i.e., U is an open subset of $\mathrm{St}(d_x, d_y)$ and φ is a diffeomorphism between U and a subset of a Euclidean space). Indeed, $D\varphi$ is an isomorphism between the tangent space and a Euclidean space, see the proof of Proposition 3.10 in [93]. Letting $u_{\mathbf{\Delta}}$ be such that $\varphi^{-1}(u_{\mathbf{\Delta}}) = \mathbf{\Delta}$ we have that $F|_{\mathrm{St}(d_x,d_y)}(\chi) = F \circ \varphi^{-1}(u_\chi)$ for each $\chi \in U$. Consider the function

$$S : u \in \varphi(U) \mapsto \min_{j_1,\ldots,j_R=1}^{|\mathcal{V}|} F_{(j_1,\ldots,j_R)}(\varphi^{-1}(u))$$

which is such that $F_{(j_1,\ldots,j_R)} \circ \varphi^{-1}$ is a finite collection of smooth functions so that, following the logic of the proof of Proposition 4.10,

$$\begin{aligned} \partial S(u_{\mathbf{\Delta}}) &= \mathrm{conv}\left(\left\{ \nabla \left(F_{\bar{j}(\mathbf{\Delta})} \circ \varphi^{-1} \right)(u_{\mathbf{\Delta}}) : \bar{j}(\mathbf{\Delta}) \in \bar{J}(\mathbf{\Delta}) \right\} \right) \\ &= \mathrm{conv}\left(\left\{ (D\varphi^{-1}(u_{\mathbf{\Delta}}))^* D(F_{\bar{j}(\mathbf{\Delta})})_{[\mathbf{\Delta}]} : \bar{j}(\mathbf{\Delta}) \in \bar{J}(\mathbf{\Delta}) \right\} \right) \\ &= \mathrm{conv}\left(\left\{ ((D\varphi(\mathbf{\Delta}))^{-1})^* D(F_{\bar{j}(\mathbf{\Delta})})_{[\mathbf{\Delta}]} : \bar{j}(\mathbf{\Delta}) \in \bar{J}(\mathbf{\Delta}) \right\} \right) \\ &= ((D\varphi(\mathbf{\Delta}))^{-1})^* \,\mathrm{conv}\left(\left\{ D(F_{\bar{j}(\mathbf{\Delta})})_{[\mathbf{\Delta}]} : \bar{j}(\mathbf{\Delta}) \in \bar{J}(\mathbf{\Delta}) \right\} \right) \end{aligned}$$

where we have used the notation $D\varphi^{-1}$ to denote the derivative of φ^{-1}. By [72, Proposition 3.1]

$$\partial_R F(\mathbf{\Delta}) = \left\{ (D\varphi(\mathbf{\Delta}))^{-1} \left(G_{u_{\mathbf{\Delta}}}^{-1} ((D\varphi(\mathbf{\Delta}))^{-1})^* g \right) : g \in \mathrm{conv}\left(\left\{ D(F_{\bar{j}(\mathbf{\Delta})})_{[\mathbf{\Delta}]} : \bar{j}(\mathbf{\Delta}) \in \bar{J}(\mathbf{\Delta}) \right\} \right) \right\}$$

where $G_{u_\Delta} = ((D\varphi(\mathbf{\Delta}))^{-1})^* D\varphi(\mathbf{\Delta})^{-1}$. Thus $(D\varphi(\mathbf{\Delta}))^{-1} G_{u_{\mathbf{\Delta}}}^{-1} (D\varphi(\mathbf{\Delta})^*)^{-1}$ is the orthogonal projection onto the range of $(D\varphi(\mathbf{\Delta}))^{-1}$ which is simply the tangent space to $\mathrm{St}(d_x, d_y)$ at $\mathbf{\Delta}$. The display above therefore shows that $\partial_R F(\mathbf{\Delta})$ coincides with the collection of projections of subgradients of F at $\mathbf{\Delta}$ onto the tangent space of $\mathrm{St}(d_x, d_y)$ as desired.

F Convergence of Algorithm 2

Given that the proof of Theorem 4.11 requires a number of auxiliary results, we provide additional exposition in this section.

With Proposition 4.10 in hand, it is relatively straightforward to derive bounds on the (local) Lipschitz constant of F. It will be helpful, in the sequel, to establish estimates on this constant over the set of all points $\mathbf{\Delta} \in \mathbb{R}^{d_y \times d_x}$ which are close to the Stiefel manifold in the sense that $\|\mathbf{\Delta}^\intercal \mathbf{\Delta} - \mathbf{I}_{d_x}\|_{\mathrm{F}} \leq R$ for some fixed $R > 0$.

Corollary F.1 (Subgradient bound)**.** *In the setting of Proposition 4.10, for any* $\mathbf{\Delta} \in \mathbb{R}^{d_y \times d_x}$,

$$\sup_{\mathbf{S} \in \partial F(\mathbf{\Delta})} \|\mathbf{S}\|_{\mathrm{F}} \leq 4M_2(\mu)M_2(\nu)\|\mathbf{\Delta}\|_{\mathrm{op}} + 4M_2(\mu)^2\|\mathbf{\Delta}\|_{\mathrm{op}}^3 \text{ and } \|\mathbf{\Delta}\|_{\mathrm{op}} \leq \sqrt{1 + \|\mathbf{\Delta}^\intercal \mathbf{\Delta} - \mathbf{I}_{d_x}\|_{\mathrm{F}}}.$$

If $\mathbf{\Delta}, \bar{\mathbf{\Delta}} \in \mathbb{R}^{d_y \times d_x}$ *are such that* $\max\{\|\mathbf{\Delta}^\intercal \mathbf{\Delta} - \mathbf{I}_{d_x}\|_{\mathrm{F}}, \|\bar{\mathbf{\Delta}}^\intercal \bar{\mathbf{\Delta}} - \mathbf{I}_{d_x}\|_{\mathrm{F}}\} \leq R$*, then*

$$\left|F(\mathbf{\Delta}) - F(\bar{\mathbf{\Delta}})\right| \leq L_R \|\mathbf{\Delta} - \bar{\mathbf{\Delta}}\|_{\mathrm{F}},$$

for $L_R \coloneqq 4M_2(\mu)\sqrt{1+R}\left(M_2(\nu) + M_2(\mu)(1+R)\right)$.

Proof. We first establish the inequality $\|\mathbf{\Delta}\|_{\mathrm{op}} \leq \sqrt{1 + \|\mathbf{\Delta}^\intercal \mathbf{\Delta} - \mathbf{I}_{d_x}\|_{\mathrm{F}}}$ as follows,

$$\|\mathbf{\Delta}^\intercal \mathbf{\Delta} - \mathbf{I}_{d_x}\|_{\mathrm{F}}^2 = \sum_{i=1}^{d_x} \left(\lambda_i(\mathbf{\Delta}^\intercal \mathbf{\Delta}) - 1\right)^2 \geq \left(\lambda_1(\mathbf{\Delta}^\intercal \mathbf{\Delta}) - 1\right)^2 = \left(\|\mathbf{\Delta}\|_{\mathrm{op}}^2 - 1\right)^2.$$

The bound on the norm of the subgradients is a direct consequence of Proposition 2.1.2 in [71] and the proof of Lemma A.5, see Appendix B.5. The final claim also follows directly from the proof of Lemma A.5. □

Proposition F.2 (Propositions 3.1, 3.2, and 3.6 in [69])**.** *Let* $H : \mathbf{\Delta} \in \mathbb{R}^{d_y \times d_x} \mapsto \mathbb{R}$ *be given as*

$$H(\mathbf{\Delta}) = F(A(\mathbf{\Delta})) + \frac{\beta}{4}\|\mathbf{\Delta}^\intercal \mathbf{\Delta} - \mathbf{I}_{d_x}\|_{\mathrm{F}}^2, \text{ where } A(\mathbf{\Delta}) = \frac{1}{8}\mathbf{\Delta}\left(15\mathbf{I}_{d_x} - 10\mathbf{\Delta}^\intercal \mathbf{\Delta} + 3(\mathbf{\Delta}^\intercal \mathbf{\Delta})^2\right),$$

and $\beta > 0$ *is a constant. Then, for any* $\mathbf{\Delta} \in \mathbb{R}^{d_y \times d_x}$,

1. $\|\mathbf{\Delta}^\intercal \mathbf{\Delta} - \mathbf{I}_{d_x}\|_{\mathrm{F}}^3 \geq \|A(\mathbf{\Delta})^\intercal A(\mathbf{\Delta}) - \mathbf{I}_{d_x}\|_{\mathrm{F}}$ *provided that* $\|\mathbf{\Delta}^\intercal \mathbf{\Delta} - \mathbf{I}_{d_x}\|_{\mathrm{F}} \leq 1$,
2. $\partial H(\mathbf{\Delta}) = DA_{[\mathbf{\Delta}]}(\partial F(A(\mathbf{\Delta}))) + \beta\mathbf{\Delta}(\mathbf{\Delta}^\intercal \mathbf{\Delta} - \mathbf{I}_{d_x})$,
3. $DA_{[\mathbf{\Delta}]}(\mathbf{\Xi}) = \frac{1}{8}\mathbf{\Xi}\left(15\mathbf{I}_{d_x} - 10\mathbf{\Delta}^\intercal \mathbf{\Delta} + 3(\mathbf{\Delta}^\intercal \mathbf{\Delta})^2\right) - \mathbf{\Delta}\Phi\left(\mathbf{\Delta}^\intercal \mathbf{\Xi}\right) + \frac{3}{2}\mathbf{\Delta}\Phi\left(\Phi(\mathbf{\Delta}^\intercal \mathbf{\Xi})(\mathbf{\Delta}^\intercal \mathbf{\Delta} - \mathbf{I}_{d_x})\right)$ *where, for a square matrix* $\mathbf{M}$, $\Phi(\mathbf{M}) = \frac{1}{2}(\mathbf{M} + \mathbf{M}^\intercal)$.

The main insight in [69], which we exposit below, is that applying unconstrained optimization procedures to H can yield, in certain cases, a (Riemannian) Clarke critical point for F and even local minimizers of F over $\mathrm{St}(d_x, d_y)$. Proposition F.2 endows us with a formula for the subdifferential of H at $\mathbf{\Delta}$ in terms of the subdifferential for F at $A(\mathbf{\Delta})$ so that subgradient methods can easily be implemented and shows that, if $\mathbf{\Delta} \in \mathbb{R}^{d_y \times d_x}$ $\|\mathbf{\Delta}^\intercal \mathbf{\Delta} - \mathbf{I}_{d_x}\|_{\mathrm{F}} < 1$, then $A(\mathbf{\Delta})$ violates the Stiefel manifold constraint less than $\mathbf{\Delta}$ so that $A(\mathbf{\Delta})$ can be thought of as a correction term whereas $\frac{\beta}{4}\|\mathbf{\Delta}^\intercal \mathbf{\Delta} - \mathbf{I}_{d_x}\|_{\mathrm{F}}^2$ serves as a penalty term. The following result establishes that, under certain conditions, a critical point of H is in fact a Riemannian critical for F over the Stiefel manifold.

Proposition F.3 (Theorems 3.10 and 3.14 and Remark 3.7 in [69])**.** *Suppose that* $\mathbf{\Delta} \in \mathbb{R}^{d_y \times d_x}$ *satisfies* $\|\mathbf{\Delta}^\intercal \mathbf{\Delta} - \mathbf{I}_{d_x}\|_{\mathrm{F}} \leq \frac{\beta}{2\beta + 8L_1}$ *for* L_1 *as defined in Corollary F.1, then*

$$\inf_{\mathbf{W} \in \partial H(\mathbf{\Delta})} \|\mathbf{W}\|_{\mathrm{F}} \geq \frac{\beta}{4}\|\mathbf{\Delta}^\intercal \mathbf{\Delta} - \mathbf{I}_{d_x}\|_{\mathrm{F}} \geq \frac{\beta}{4}\|A(\mathbf{\Delta})^\intercal A(\mathbf{\Delta}) - \mathbf{I}_{d_x}\|_{\mathrm{F}}^{1/3}.$$

Moreover, if $\mathbf{\Delta} \in \mathrm{St}(d_x, d_y)$*, then* $A(\mathbf{\Delta}) = \mathbf{\Delta}$ *and* $\partial H(\mathbf{\Delta}) = \partial_R F(A(\mathbf{\Delta})) = \partial_R F(\mathbf{\Delta})$*. Finally, if* $\mathbf{\Delta}^\star$ *is a local minimizer of* H *satisfying* $\|(\mathbf{\Delta}^\star)^\intercal \mathbf{\Delta}^\star - \mathbf{I}_{d_x}\|_{\mathrm{F}} \leq \frac{\beta}{2\beta + 8L_1}$*, then* $\mathbf{\Delta}^\star$ *is a local minimizer of* F *over* $\mathrm{St}(d_x, d_y)$.

Proposition F.3 implies, in particular, that if H admits a critical point, $\mathbf{\Delta} \in \mathbb{R}^{d_y \times d_x}$, which violates the Stiefel manifold condition by less than a prescribed tolerance, $\mathbf{\Delta}$ must actually be an element of the Stiefel manifold and is critical in the Riemannian sense. Local minimality also transfers over in a similar manner.

As Proposition F.3 requires a certain *a priori* bound on how much $\mathbf{\Delta}$ violates the Stiefel manifold constraint, it is convenient to establish sufficient conditions for the iterates of a minimization procedure applied to H to satisfy these conditions. Precisely, we consider iterates generated by Algorithm 2 which can be shown to satisfy $\|\mathbf{\Delta}_k^\intercal \mathbf{\Delta}_k - \mathbf{I}_{d_x}\| \leq \frac{1}{6}$ for all $k \in \mathbb{N}$ provided that the algorithm is initialized at a point, $\mathbf{\Delta}_1 \in \mathbb{R}^{d_y \times d_x}$, satisfying the same constraint.

Proposition F.4 (Stiefel violation)**.** *Fix* $\alpha = 4M_2(\mu)\sqrt{\frac{217}{216}}\left(M_2(\nu) + \frac{217}{216}M_2(\mu)\right), \beta \geq 162\alpha$, *and let* $\eta_k \leq \frac{1}{2\beta}$ *for all* $k \in \mathbb{N}$. *Then, the iterates of Algorithm 2 satisfy* $\|\mathbf{\Delta}_k^\intercal \mathbf{\Delta}_k - \mathbf{I}_{d_x}\|_{\mathrm{F}} \leq 1/6$.

Proof. The proof of Proposition F.4 follows directly from Lemma 4.12 in [69] once we establish that, at any $\mathbf{\Delta}_k \in \mathbb{R}^{d_y \times d_x}$ with $\|\mathbf{\Delta}_k^\intercal \mathbf{\Delta}_k - \mathbf{I}_{d_x}\|_{\mathrm{F}} \leq \frac{1}{6}$ it holds that $\|DA_{[\mathbf{\Delta}_k]}(\mathbf{S}_k)\|_{\mathrm{F}} \leq 2.7\alpha$. First, we have from Corollary F.1 that for any $\mathbf{S}_k \in \partial F(A(\mathbf{\Delta}_k))$,

$$\|\mathbf{S}_k\|_{\mathrm{F}} \leq 4M_2(\mu)\sqrt{1 + \|A(\mathbf{\Delta}_k)^\intercal A(\mathbf{\Delta}_k) - \mathbf{I}_{d_x}\|_{\mathrm{F}}} \times \\ \left(M_2(\nu) + M_2(\mu)\left(1 + \|A(\mathbf{\Delta}_k)^\intercal A(\mathbf{\Delta}_k) - \mathbf{I}_{d_x}\|_{\mathrm{F}}\right)\right).$$

By Proposition F.2, if $\|\mathbf{\Delta}_k^\intercal \mathbf{\Delta}_k - \mathbf{I}_{d_x}\|_{\mathrm{F}} \leq \frac{1}{6}$, $\|A(\mathbf{\Delta}_k)^\intercal A(\mathbf{\Delta}_k) - \mathbf{I}_{d_x}\|_{\mathrm{F}} \leq \|\mathbf{\Delta}_k^\intercal \mathbf{\Delta}_k - \mathbf{I}_{d_x}\|_{\mathrm{F}}^3 \leq \frac{1}{216}$ so that $\|\mathbf{S}_k\|_{\mathrm{F}} \leq \alpha$. Now, recall that

$$DA_{[\mathbf{\Delta}]}(\mathbf{\Xi}) = \frac{1}{8}\mathbf{\Xi}\left(15\mathbf{I}_{d_x} - 10\mathbf{\Delta}^\intercal\mathbf{\Delta} + 3(\mathbf{\Delta}^\intercal\mathbf{\Delta})^2\right) - \mathbf{\Delta}\Phi\left(\mathbf{\Delta}^\intercal\mathbf{\Xi}\right) + \frac{3}{2}\mathbf{\Delta}\Phi\left(\Phi(\mathbf{\Delta}^\intercal\mathbf{\Xi})\left(\mathbf{\Delta}^\intercal\mathbf{\Delta} - \mathbf{I}_{d_x}\right)\right)$$

for $\Phi(\mathbf{M}) = \frac{1}{2}(\mathbf{M} + \mathbf{M}^\intercal)$. We now simplify the various expressions. First,

$$\begin{aligned}
&\frac{1}{8}\mathbf{\Xi}\left(15\mathbf{I}_{d_x} - 10\mathbf{\Delta}^\intercal\mathbf{\Delta} + 3(\mathbf{\Delta}^\intercal\mathbf{\Delta})^2\right) \\
&= \frac{1}{8}\mathbf{\Xi}\left(8\mathbf{I}_{d_x} + 7\left(\mathbf{I}_{d_x} - \mathbf{\Delta}^\intercal\mathbf{\Delta}\right) + 3(\mathbf{\Delta}^\intercal\mathbf{\Delta})\left(\mathbf{\Delta}^\intercal\mathbf{\Delta} - \mathbf{I}_{d_x}\right)\right) \\
&= \mathbf{\Xi} + \frac{7}{8}\mathbf{\Xi}\left(\mathbf{I}_{d_x} - \mathbf{\Delta}^\intercal\mathbf{\Delta}\right) + \frac{3}{8}\mathbf{\Xi}\left(\mathbf{\Delta}^\intercal\mathbf{\Delta} - \mathbf{I}_{d_x}\right)^2 + \frac{3}{8}\mathbf{\Xi}\left(\mathbf{\Delta}^\intercal\mathbf{\Delta} - \mathbf{I}_{d_x}\right),
\end{aligned}$$

whereby $\frac{1}{8}\mathbf{S}_k\left(15\mathbf{I}_{d_x} - 10\mathbf{\Delta}_k^\intercal\mathbf{\Delta}_k + 3(\mathbf{\Delta}_k^\intercal\mathbf{\Delta}_k)^2\right) \leq \alpha + \frac{10}{48}\alpha + \frac{3}{288}\alpha = \frac{39}{32}\alpha$.

For the next term, it holds that

$$\|\mathbf{\Delta}_k^\intercal\mathbf{S}_k\|_{\mathrm{F}} \leq \|\mathbf{\Delta}_k^\intercal\|_{\mathrm{op}}\|\mathbf{S}_k\|_{\mathrm{F}} \leq \alpha\sqrt{1 + \|\mathbf{\Delta}_k^\intercal\mathbf{\Delta} - \mathbf{I}_{d_x}\|_{\mathrm{F}}} \leq \alpha\sqrt{\frac{7}{6}}$$

so that $\|\mathbf{\Delta}_k\Phi(\mathbf{\Delta}_k^\intercal\mathbf{S}_k)\|_{\mathrm{F}} \leq \|\mathbf{\Delta}_k\|_{\mathrm{op}}\|\Phi(\mathbf{\Delta}_k^\intercal\mathbf{S}_k)\|_{\mathrm{F}} \leq \frac{7}{6}\alpha$ noting that, for square matrix $\mathbf{M}$, $\|\Phi(\mathbf{M})\|_{\mathrm{F}} \leq \frac{1}{2}(\|\mathbf{M}^\intercal\|_{\mathrm{F}} + \|\mathbf{M}\|_{\mathrm{F}}) = \|\mathbf{M}\|_{\mathrm{F}}$.

Using this logic for the final term we have

$$\begin{aligned}
\frac{3}{2}\|\mathbf{\Delta}_k\Phi\left(\Phi(\mathbf{\Delta}_k^\intercal\mathbf{S}_k)\left(\mathbf{\Delta}_k^\intercal\mathbf{\Delta}_k - \mathbf{I}_{d_x}\right)\right)\|_{\mathrm{F}} &\leq \frac{3}{2}\|\mathbf{\Delta}_k\|_{\mathrm{op}}\|\Phi(\mathbf{\Delta}_k^\intercal\mathbf{S}_k)\left(\mathbf{\Delta}_k^\intercal\mathbf{\Delta}_k - \mathbf{I}_{d_x}\right)\|_{\mathrm{F}} \\
&\leq \frac{3}{2}\|\mathbf{\Delta}_k\|_{\mathrm{op}}\|\mathbf{\Delta}_k^\intercal\mathbf{S}_k\|_{\mathrm{F}}\|\mathbf{\Delta}_k^\intercal\mathbf{\Delta}_k - \mathbf{I}_{d_x}\|_{\mathrm{F}} \leq \frac{7}{24}\alpha.
\end{aligned}$$

Applying the triangle inequality, we obtain that $\|DA_{[\mathbf{\Delta}_k]}(\mathbf{S}_k)\|_{\mathrm{F}} \leq \left(\frac{39}{32} + \frac{7}{6} + \frac{7}{24}\right)\alpha = \frac{257}{96}\alpha \leq 2.7\alpha$, so that we may apply Lemma 4.12 in [69] which asserts that if $\beta \geq 60 \cdot 2.7\alpha$ and $\eta_k \leq \frac{1}{2\beta}$, the subgradient method generates iterates $\mathbf{\Delta}_k$ satisfying $\|\mathbf{\Delta}_k^\intercal\mathbf{\Delta}_k - \mathbf{I}_{d_x}\|_{\mathrm{F}} \leq 1/6$ provided that $\mathbf{\Delta}_1$ also satisfies this condition. □

With Proposition F.4 in hand, it remains to show that the iterates generated by Algorithm 2 converge in a suitable sense. To this end, we leverage Corollary 5.9 in [70] (see also Corollary 4.16 in [69]) for which it suffices to prove that H is a semialgebraic function, that is, the graph of H, $\mathrm{grp}(H) \coloneqq \left\{(\mathrm{vec}(\mathbf{\Delta}), y) \in \mathbb{R}^{d_x d_y} \times \mathbb{R} : H(\mathbf{\Delta}) = y\right\}$, can be written as a finite union of sets of the form $\left\{(\mathrm{vec}(\mathbf{\Delta}), y) \in \mathbb{R}^{d_x d_y} \times \mathbb{R} : p_i(\mathrm{vec}(\mathbf{\Delta}), y) \leq 0 \text{ for } i = 1, \ldots, \ell\right\}$ where $(p_i)_{i=1}^\ell$ is a finite collection of polynomials; any set of this form is said to be semialgebraic.

Lemma F.5 (Semialgebraicity of H). *The function $\mathbf{\Delta} \in \mathbb{R}^{d_y \times d_x} \mapsto H(\mathbf{\Delta})$ is semialgebraic.*

Proof. We first compile some useful properties of semialgebraic maps. The first two points can be found on p.29 in [94] and the last two in Corollary 2.9 of the same reference,

(1) any map with polynomial coordinates is semialgebraic,

(2) the absolute value of a semialgebraic function is semialgebraic,

(3) the composition of two semialgebraic mappings is semialgebraic,

(4) semialgebraic functions are closed under addition and scalar multiplication.

As a direct corollary to items (2) and (4), we obtain that the pointwise minimum of two semialgebraic functions $f, g : \mathbb{R}^{d_y \times d_x} \to \mathbb{R}$ is semialgebraic as $\min\{f(x), g(x)\} = \frac{1}{2}\left(f(x) + g(x) - |f(x) - g(x)|\right)$. Functions defined as the pointwise minimum of finitely many semialgebraic functions can be seen to be semialgebraic by iteratively applying the above result.

With these points in mind, we note that A is a semialgebraic mapping and that

$$
\begin{aligned}
P &: \mathbf{\Delta} \in \mathbb{R}^{d_y \times d_x} \mapsto \frac{\beta}{4}\|\mathbf{\Delta}^\intercal \mathbf{\Delta} - \mathbf{I}_{d_x}\|_{\mathrm{F}}^2, \\
U_i &: \mathbf{\Delta} \in \mathbb{R}^{d_y \times d_x} \mapsto \left(\left(\theta^{(i)}\right)^\intercal \mathbf{\Delta} \int z z^\intercal d\mu(z) \mathbf{\Delta}^\intercal \theta^{(i)}\right)^2, \ (i = 1, \dots, R) \\
V_{i,\pi} &: \mathbf{\Delta} \in \mathbb{R}^{d_y \times d_x} \mapsto -2\left(\left(\theta^{(i)}\right)^\intercal \mathbf{\Delta} \int x y^\intercal d\pi(x, y) \theta^{(i)}\right)^2, \ (i = 1, \dots, R, \pi \in \Pi(\mu, \nu))
\end{aligned}
$$

are all polynomials of $\mathrm{vec}(\mathbf{\Delta})$ and are hence semialgebraic. Since

$$
F(\mathbf{\Delta}) = \frac{1}{R}\sum_{i=1}^{R}\left(U_i(\mathbf{\Delta}) + M_2(((\theta^{(i)})^\intercal)_\sharp \nu) + \min_{j=1}^{|\mathcal{V}|} V_{i,\pi_j}(\mathbf{\Delta})\right),
$$

where $\mathcal{V}$ is the set of vertices of $\Pi(\mu, \nu)$ (recall the proof of Proposition 4.10), F is semialgebraic as follows from the previous deliberations. Now, $H(\mathbf{\Delta}) = F(A(\mathbf{\Delta})) + P(\mathbf{\Delta})$, proving that H is semialgebraic. □

With these preparations in hand, we readily obtain the convergence of Algorithm 2.

Proof of Theorem 4.11. Given the previous deliberations Theorem 4.11 is a direct consequence of Corollary 5.9 in [70]. Notably, while their result is stated for locally Lipschitz Whitney stratifiable functions, Lemma F.5 proves that H is a semialgebraic function and hence satisfies the conditions of Corollary 5.9 in [70]; see the discussion following that result.

It remains to show that Assumption C in the same reference holds. Given the choice of step sequence and the fact that the subgradients are not stochastic, it suffices to show that $\sup_{k \in \mathbb{N}} \|\mathbf{\Delta}_k\|_{\mathrm{F}} < \infty$. However, since $\eta_k \leq \frac{1}{2\beta}$, Proposition F.4 yields that $\|\mathbf{\Delta}_k\|_{\mathrm{op}} \leq \sqrt{1 + \|\mathbf{\Delta}_k^\intercal \mathbf{\Delta}_k - \mathbf{I}_{d_x}\|_{\mathrm{F}}} \leq \sqrt{\frac{7}{6}}$, as all norms on $\mathbb{R}^{d_y \times d_x}$ are equivalent, this proviso is met.

Applying Corollary 5.9 in [70], we obtain that the iterates of Algorithm 2 admit cluster points which are critical points of H. Finally, since $\beta \geq 2L_1$, Proposition F.3 asserts that any critical point $\bar{\mathbf{\Delta}}$ of H satisfying $\|\bar{\mathbf{\Delta}}^\intercal \bar{\mathbf{\Delta}} - \mathbf{I}_{d_x}\|_{\mathrm{F}} \leq \frac{1}{6} \leq \frac{\beta}{2\beta + 8L_1}$ lies on the Stiefel manifold and is a Riemannian critical point for F over the Stiefel manifold. However, all iterates of Algorithm 2 satisfy $\|\mathbf{\Delta}_k^\intercal \mathbf{\Delta}_k - \mathbf{I}_{d_x}\|_{\mathrm{F}} \leq \frac{1}{6}$, see Proposition F.4. □

G Additional Details on Experiments

The code used to generate the experimental results is available in the supplement to the submission.

G.1 Monte-Carlo Error and Empirical Convergence Validation

We provide additional details for the validation experiments from Section 5. Recall that we fix $d_x = 5$ and $d_y = 10$, and consider centered Gaussian distributions $\mu = \mathcal{N}(0, \mathbf{\Sigma}_\mu)$ and $\nu = \mathcal{N}(0, \mathbf{\Sigma}_\nu)$. The covariance matrices are generated as

$$\mathbf{\Sigma}_\mu = \mathbf{S}_\mu^\intercal \mathbf{S}_\mu, \qquad \mathbf{\Sigma}_\nu = \mathbf{S}_\nu^\intercal \mathbf{S}_\nu,$$

where the entries of $\mathbf{S}_\mu \in \mathbb{R}^{d_x \times d_x}$ and $\mathbf{S}_\nu \in \mathbb{R}^{d_y \times d_y}$ are sampled uniformly from $[0, 1]$ and are given respectively by

$$\begin{pmatrix} 0.3236 & 0.2191 & 0.6560 & 0.5031 & 0.8479 \\ 0.9789 & 0.9872 & 0.9049 & 0.8046 & 0.1269 \\ 0.0338 & 0.0987 & 0.1037 & 0.5406 & 0.2284 \\ 0.5851 & 0.4209 & 0.8828 & 0.3423 & 0.2496 \\ 0.7108 & 0.2808 & 0.6270 & 0.4175 & 0.8096 \end{pmatrix}$$

and

$$\begin{pmatrix} 0.9161 & 0.2135 & 0.7142 & 0.8724 & 0.2980 & 0.6841 & 0.4130 & 0.5256 & 0.2159 & 0.3184 \\ 0.7531 & 0.5333 & 0.7351 & 0.6972 & 0.1939 & 0.8199 & 0.8531 & 0.6292 & 0.6706 & 0.8047 \\ 0.9606 & 0.8050 & 0.5413 & 0.1567 & 0.3258 & 0.7775 & 0.0108 & 0.0359 & 0.5499 & 0.3627 \\ 0.6861 & 0.3847 & 0.6097 & 0.7144 & 0.2869 & 0.7862 & 0.0350 & 0.9600 & 0.5924 & 0.4226 \\ 0.9395 & 0.2172 & 0.7135 & 0.4205 & 0.5785 & 0.1037 & 0.5993 & 0.9797 & 0.8764 & 0.0343 \\ 0.9353 & 0.6064 & 0.1920 & 0.2301 & 0.8376 & 0.5580 & 0.3768 & 0.5624 & 0.8435 & 0.3506 \\ 0.7907 & 0.2041 & 0.4826 & 0.1531 & 0.5318 & 0.4466 & 0.8603 & 0.7347 & 0.2604 & 0.3680 \\ 0.2773 & 0.0817 & 0.3731 & 0.1265 & 0.8599 & 0.5523 & 0.4136 & 0.7552 & 0.1519 & 0.4664 \\ 0.4340 & 0.9845 & 0.7846 & 0.2799 & 0.7939 & 0.6587 & 0.5673 & 0.2789 & 0.7801 & 0.0522 \\ 0.2094 & 0.5340 & 0.9646 & 0.4347 & 0.9174 & 0.6073 & 0.9301 & 0.0570 & 0.4485 & 0.8742 \end{pmatrix}$$

where the displayed values are truncated to four digits after the decimal point.

The value of $\underline{\mathsf{IGW}}(\mu, \nu)$ and the corresponding population optimizer over $\mathrm{St}(d_x, d_y)$ are provided in closed form in Example 4.3. Furthermore, the IGW cost for the projected distributions can be evaluated directly from the covariance matrices. Specifically, for each $\mathbf{\Delta} \in \mathbb{R}^{d_y \times d_x}$ and $\theta \in \mathbb{S}^{d_y - 1}$,

$$\mathsf{IGW}\left((\theta^\intercal \mathbf{\Delta})_\sharp \mu, (\theta^\intercal)_\sharp \nu\right)^2 = (\theta^\intercal \mathbf{\Delta} \mathbf{\Sigma}_\mu \mathbf{\Delta}^\intercal \theta)^2 + (\theta^\intercal \mathbf{\Sigma}_\nu \theta)^2 - 2\theta^\intercal \mathbf{\Delta} \mathbf{\Sigma}_\mu \mathbf{\Delta}^\intercal \theta \, \theta^\intercal \mathbf{\Sigma}_\nu \theta,$$

by [35, Corollary 3.2]. Thus, the MC validation experiment does not require discretizing μ and ν. In this experiment, we set $\beta = 100$, a maximum number of iterations of 2500, and take the stepsize sequence $\eta_k = \min\{.01/\max\{\|\mathbf{G}_k\|, 1\}, 5000/(k+1)\}$ for the standard subgradient method Algorithm 2. While this choice of β and stepsize do not fall within the regime required to apply Algorithm 2, they are chosen as they were seen to lead to a good tradeoff between performance and compute time. For the Riemannian subgradient method, Algorithm 3, we set a maximum number of iterations of 500, a gradient tolerance of 5×10^{-6}, and a maximum number of backtracking steps of 12. The average runtime for this experiment over the 25 runs as a function of the number of slices with the different choices of initialization and algorithm is provided in Fig. 5.

For the validation of the sample complexity, we compute the sliced IGW distance between empirical measures. The complexity of computing one subgradient of F in this setting is as follows.

Assume first that, $\theta \in \mathbb{S}^{d_y - 1}$ and $\mathbf{\Delta} \in \mathrm{St}(d_x, d_y)$ are fixed and that μ and ν are both uniformly distributed on N points $(X_i)_{i=1}^N$ and $(Y_j)_{j=1}^N$, respectively. Following Theorem 3.1, the optimal coupling is obtained by performing the following routine under the assumption that the projected distributions are still uniformly supported on the same number of points.

1. Compute the projected support points for μ, $x_i^\theta = \theta^\intercal \mathbf{\Delta} X_i$ for $_{i=1}$ to N: $O(d_x d_y + N d_x)$,
2. Compute the projected support points for ν, $y_i^\theta = \theta^\intercal Y_i$ for $i = 1$ to N: $O(N d_y)$,
3. Sort the projected support points x_i^θ and y_i^θ in ascending order: $O(N \log(N))$,
4. Compute $v_{\mathrm{id}} = \left(\sum_{i=1}^N x_i^\theta y_i^\theta\right)^2$ and $v_{\check{\mathrm{id}}} = \left(\sum_{i=1}^N x_i^\theta y_{N+1-i}^\theta\right)^2$: $O(N)$,
5. If $v_{\mathrm{id}} \geq v_{\check{\mathrm{id}}}$, $\pi^\star = \pi_{\mathrm{id}}$ is optimal for $\mathsf{IGW}((\theta^\intercal \mathbf{\Delta})_\sharp \mu, (\theta^\intercal)_\sharp \nu)$, else $\pi^\star = \pi_{\check{\mathrm{id}}}$ is optimal (see Theorem 3.1)

6. Compute an element of the subdifferential of G^θ: $\mathbf{T}_\theta = -4\theta^\intercal \mathbf{\Delta} \mathbf{C}_{\pi^\star} \theta \left(\mathbf{C}_{\pi^\star} \theta\theta^\intercal\right)^\intercal$, where $\mathbf{C}_{\pi^\star} = \int xy^\intercal d\pi^\star(x,y)$, $\theta^\intercal \mathbf{\Delta} \mathbf{C}_{\pi^\star} \theta$ can be computed in $O(N)$ whereas $\left(\mathbf{C}_{\pi^\star} \theta\theta^\intercal\right)^\intercal$ can be computed in $O(Nd_x + d_x d_y)$
7. Compute $\mathbf{L}_\theta = 4\left((\theta)^\intercal \mathbf{\Delta} \mathbf{R}_\mu \mathbf{\Delta}^\intercal \theta\right)\left(\mathbf{R}_\mu \mathbf{\Delta}^\intercal \theta(\theta)^\intercal\right)^\intercal$ where $\mathbf{R}_\mu = \int zz^\intercal d\mu(z)$. Again, the constant $\left((\theta)^\intercal \mathbf{\Delta} \mathbf{R}_\mu \mathbf{\Delta}^\intercal \theta\right)$ can be computed in $O(N)$ whereas $\left(\mathbf{R}_\mu \mathbf{\Delta}^\intercal \theta(\theta)^\intercal\right)^\intercal$ can be computed in $O(Nd_x + d_x d_y)$.

In sum, the overall complexity of computing $\mathbf{T}_\theta + \mathbf{R}_\theta$ is $O(N(d_x + d_y) + d_x d_y + N\log(N))$ and, since $d_x \leq d_y$ and $\partial F(\mathbf{\Delta}) = \frac{1}{R}\sum_{i=1}^R (\mathbf{T}_{\theta^{(i)}} + \mathbf{R}_{\theta^{(i)}})$ for some collection of unit vectors $(\theta^{(i)})_{i=1}^R$, the overall per-iteration complexity for computing one subgradient is given by $O(R\left(N(d_y + \log(N)) + d_x d_y\right))$. With this, computing a subgradient of H at $\mathbf{\Delta}$ via Proposition 4.10 requires first computing $A(\mathbf{\Delta}) = \frac{1}{8}\mathbf{\Delta}\left(15\mathbf{I}_{d_x} - 10\mathbf{\Delta}^\intercal\mathbf{\Delta} + 3(\mathbf{\Delta}^\intercal\mathbf{\Delta})^2\right)$ whose complexity, $O(d_x^2 d_y)$, is dominated by forming $\mathbf{\Delta}^\intercal\mathbf{\Delta}$ (since $d_x \leq d_y$), then computing a subgradient of F at $A(\mathbf{\Delta})$ which we have just analyzed and composing that matrix with $DA_{[\mathbf{\Delta}]}$ (which consists of multiplying matrices of shape $d_y \times d_x$ or $d_x \times d_y$). Finally, the term $\beta\mathbf{\Delta}(\mathbf{\Delta}^\intercal\mathbf{\Delta} - \mathbf{I}_{d_x})$ can be computed in $O(d_y d_x^2)$ as argued above. Altogether, the complexity for computing the subgradient of H for distributions on N points satisfying the conditions described previously based on R slices is given by $O(R(N(d_y + \log(N)) + d_x d_y) + d_x^2 d_y)$.

In the case that μ, ν are such that the projected measures are supported on N, M points with, possibly, $N \neq M$ or the projected measures non-uniform weights, we instead obtain the optimal coupling, steps 1-5 above, using the North-West corner rule. Notably, the optimal coupling from this rule has at most $N + M - 1$ nonzero entries. This requires:

1. Computing the projected support points for μ and ν and sorting them: $O(d_x d_y + Nd_x + M_d y + N\log(N) + M\log(M))$,
2. Running the North-West corner rule Algorithm 1 to obtain $\pi_{\mathrm{NW}} = \pi_{\mathrm{NW}}((\theta^\intercal\mathbf{\Delta})_\sharp\mu, (\theta^\intercal)_\sharp\nu)$ and $\pi^-_{\mathrm{NW}} = (-\mathrm{id}, \mathrm{id})_\sharp \pi_{\mathrm{NW}}((-\mathrm{id})_\sharp(\theta^\intercal\mathbf{\Delta})_\sharp\mu, (\theta^\intercal)_\sharp\nu)$: $O(N+M)$
3. Compute
$$v_{\mathrm{NW}} = \left(\sum_{i=1}^N\sum_{j=1}^M x_i^\theta y_j^\theta \pi_{\mathrm{NW}}(\{(x_i^\theta, y_j^\theta)\})\right)^2, v^-_{\mathrm{NW}} = \left(\sum_{i=1}^N\sum_{j=1}^M x_i^\theta y_j^\theta \pi^-_{\mathrm{NW}}(\{(x_i^\theta, y_j^\theta)\})\right)^2$$
which can be done in $O(N+M)$ since both couplings have at most $N+M-1$ nonzero entries,
4. If $v_{\mathrm{NW}} \geq v^-_{\mathrm{NW}}$, $\pi^\star = \pi_{\mathrm{NW}}$ is optimal for $\mathsf{IGW}((\theta^\intercal\mathbf{\Delta})_\sharp\mu, (\theta^\intercal)_\sharp\nu)$, else $\pi^\star = \pi^-_{\mathrm{NW}}$ is optimal (see Theorem 3.1)
5. Compute an element of the subdifferential of G^θ: $\mathbf{T}_\theta = -4\theta^\intercal \mathbf{\Delta} \mathbf{C}_{\pi^\star} \theta \left(\mathbf{C}_{\pi^\star} \theta\theta^\intercal\right)^\intercal$, where $\mathbf{C}_{\pi^\star} = \int xy^\intercal d\pi^\star(x,y)$, $\theta^\intercal \mathbf{\Delta} \mathbf{C}_{\pi^\star} \theta$ can be computed in $O(N+M)$ (due to the structure of optimal couplings) whereas $\left(\mathbf{C}_{\pi^\star} \theta\theta^\intercal\right)^\intercal$ can be computed in $O(N + M + Nd_x + d_x d_y)$
6. Compute $\mathbf{L}_\theta = 4\left((\theta)^\intercal \mathbf{\Delta} \mathbf{R}_\mu \mathbf{\Delta}^\intercal \theta\right)\left(\mathbf{R}_\mu \mathbf{\Delta}^\intercal \theta(\theta)^\intercal\right)^\intercal$ where $\mathbf{R}_\mu = \int zz^\intercal d\mu(z)$. Again, the constant $\left((\theta)^\intercal \mathbf{\Delta} \mathbf{R}_\mu \mathbf{\Delta}^\intercal \theta\right)$ can be computed in $O(N)$ whereas $\left(\mathbf{R}_\mu \mathbf{\Delta}^\intercal \theta(\theta)^\intercal\right)^\intercal$ can be computed in $O(Nd_x + d_x d_y)$.

For the sample complexity experiment, we compute the IGW distance between the one-dimensional projections of the empirical measures. If $\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ where $X_1, \ldots, X_n \overset{i.i.d.}{\sim} \mu$, then for any $\theta \in \mathbb{S}^{d_y - 1}$ and $\mathbf{\Delta} \in \mathrm{St}(d_x, d_y)$, $(\theta^\intercal\mathbf{\Delta})_\sharp\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^n \delta_{\theta^\intercal\mathbf{\Delta}X_i}$. Thus, as long $\theta^\intercal\mathbf{\Delta}X_i \neq \theta^\intercal\mathbf{\Delta}X_j$ for each $i \neq j$ and similarly $\theta^\intercal Y_i \neq \theta^\intercal Y_j$ for each $i \neq j$ for the samples $Y_1, \ldots, Y_n \overset{i.i.d.}{\sim} \nu$, $\mathsf{IGW}\left((\theta^\intercal\mathbf{\Delta})_\sharp\hat{\mu}_n, (\theta^\intercal)_\sharp\hat{\nu}_n\right)$ can be computed using the simple sorting method as stated above. In our numerical implementation, we implement a check to see if the sorted values are pairwise distinct and warn the user if they are not – in which case multiple optimal couplings may exist. In our experiments this check always passes.

We use the same hyperparameters as the MC validation experiment for the validation of the sample complexity with a fixed number of slices, $m = 3000$. The average runtime for this experiment over the 25 runs as a function of the number of samples with the different choices of initialization and

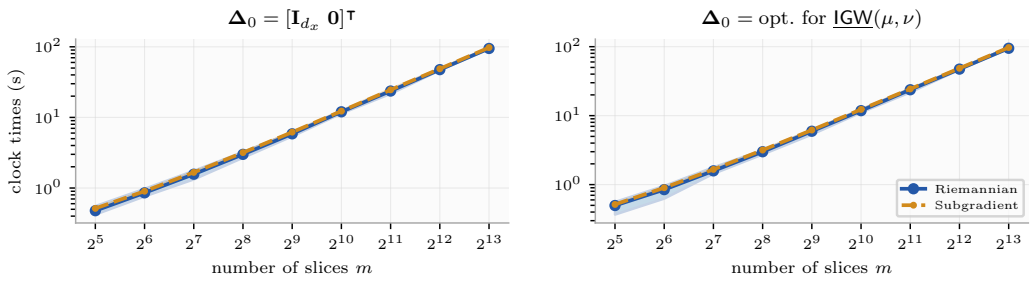


Figure 5: Average time required for the 25 runs used to generate the data for the Monte Carlo error plot Fig. 1 along with a shaded region for the maximum and minimum runtime over the 25 runs. The total compute time required to generate this data is 19243 seconds on a laptop with 64 GB of ram and an Intel i7-13620H processor.

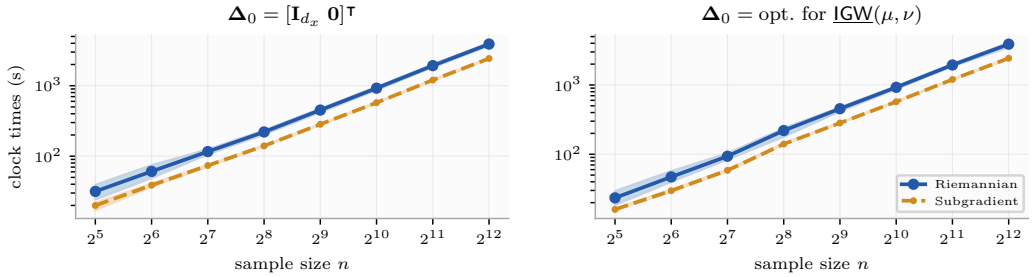


Figure 6: Average time required for the 25 runs used to generate the data for the sample complexity plot Fig. 1 along with a shaded region for the maximum and minimum runtime over the 25 runs. The total compute time required to generate this data is 617932 seconds on a laptop with 32 GB of ram and a Ryzen 7 PRO 4750U processor.

algorithm is provided in Fig. 6. For practical purposes, the data was generated by running 5 instances of the code in parallel, each running 5 runs with different seeds to generate the samples.

G.2 Riemannian Subgradient Method

The practical Riemannian subgradient method used in the experiments is presented in Algorithm 3. Neglecting the additional backtracking step, the per-iteration complexity of this method scales in the same way as the standard subgradient method in d_x, d_y, N, and R; see the discussion in the previous section and note that the QR decomposition can be implemented in $O(d_y d_x^2)$. Though the convergence of this algorithm is not guaranteed, it is seen to yield similar values to Algorithm 2.

The operation $\mathrm{proj}_{\mathrm{St}(d_x,d_y)}$ takes a matrix of the form $\mathbf{\Delta} + \mathbf{P}$, where $\mathbf{\Delta} \in \mathrm{St}(d_x, d_y)$ and $\mathbf{P}$ is in the tangent space to $\mathrm{St}(d_x, d_y)$ at $\mathbf{\Delta}$, and returns an element of $\mathrm{St}(d_x, d_y)$. This corresponds to the standard QR-based retraction onto the Stiefel manifold; see, for instance, Example 4.1.3 in [73]. The implementation used in our experiments is given in Algorithm 4.

G.3 LLM Representation Comparison

The representation-comparison experiment loads the two text datasets, `ag_news` and `amazon_polarity`, from the Hugging Face Datasets package for Python [95]. The former dataset includes no licensing information on Hugging Face, but is stated to be for non-commercial use only on the original dataset's website [96]. The latter dataset is listed under the apache-2.0 license.

As noted in the text, we use `bert-base-uncased` as the teacher model and the 23 distilled BERT student models from [77] as provided in the Hugging Face transformers package for Python [97]. All of these models are listed under the apache-2.0 license on Hugging Face. For each model and each dataset, we embed the same collection of 800 text passages. Texts are tokenized using the

Algorithm 3 Riemannian subgradient method for sliced IGW

Fix initialization $\mathbf{\Delta}_1 \in \mathrm{St}(d_x, d_y)$, total number of iterations $0 \leq K \in \mathbb{N}$, maximum number of backtracking steps $0 \leq L \in \mathbb{N}$, backtracking parameter $\alpha > 0$, and tolerance $\epsilon > 0$.

for $k = 1, \ldots, K$ **do**
  Choose $\mathbf{S}_k \in \partial F(\mathbf{\Delta}_k)$
  $\mathbf{G}_k \leftarrow \mathbf{S}_k - \frac{1}{2}\mathbf{\Delta}_k(\mathbf{\Delta}_k^\intercal \mathbf{S}_k + \mathbf{S}_k^\intercal \mathbf{\Delta}_k)$ (element of $\partial_R F(\mathbf{\Delta}_k)$)
  if $\|\mathbf{G}_k\|_{\mathrm{F}} < \epsilon$ **then**
    return Riemannian ϵ-critical point $\mathbf{\Delta}_k$
  $\eta \leftarrow 1$ and accepted $\leftarrow$ false
  for $l = 1, \ldots, L$ **do**
    $\tilde{\mathbf{\Delta}} \leftarrow \mathrm{proj}_{\mathrm{St}(d_x,d_y)}(\mathbf{\Delta}_k - \eta \mathbf{G}_k)$
    if $F(\tilde{\mathbf{\Delta}}) \leq F(\mathbf{\Delta}_k) - \alpha\eta\|\mathbf{G}_k\|_{\mathrm{F}}^2$ **then**
      accepted $\leftarrow$ true
      break
    else
      $\eta \leftarrow \eta/2$
  if accepted $=$ false **then**
    return Optimizer stuck at non-critical point $\mathbf{\Delta}_k$
  $\mathbf{\Delta}_{k+1} \leftarrow \tilde{\mathbf{\Delta}}$
return $\mathbf{\Delta}_{K+1}$

Algorithm 4 $\mathrm{proj}_{\mathrm{St}(d_x,d_y)}(\mathbf{\Delta} + \mathbf{P})$

Compute a QR decomposition $\mathbf{\Delta} + \mathbf{P} = \mathbf{QR}$, with $\mathbf{R}$ having strictly positive diagonal
return $\mathbf{Q}$

corresponding model tokenizer and represented by attention-masked mean pooling of the last hidden layer. The resulting embeddings are centered before computing the pairwise distances.

For a student-teacher pair, let X^S and Y^T denote the centered embedding matrices, whose rows correspond to the same text passages embedded by the student and teacher models, respectively. We let μ^S and ν^T denote the empirical distributions over the rows of X^S and Y^T. The empirical sliced IGW value $\underline{\mathsf{IGW}}(\mu^S, \nu^T)$ is computed using $m = 500$ slices via the Riemannian subgradient method described above. We choose this method over the standard subgradient method from the text, Algorithm 2, as it does not require a choice of regularization parameter β and does not require a choice of stepsize. The algorithm uses the same implementation for the gradient and distance computations as described in Appendix G.1 with $m = 500$ slices and a maximum number of iterations of 200. The initialization is the Gaussian alignment from Example 4.3, applied to the empirical covariance matrices $\mathbf{\Sigma}_S$ and $\mathbf{\Sigma}_T$.

We compare the empirical sliced distance with $\underline{\mathsf{IGW}}(\mathcal{N}(0, \mathbf{\Sigma}_S), \mathcal{N}(0, \mathbf{\Sigma}_T))$ and $\mathsf{IGW}(\mathcal{N}(0, \mathbf{\Sigma}_S), \mathcal{N}(0, \mathbf{\Sigma}_T))$, which admit the closed form expressions

$$\underline{\mathsf{IGW}}(\mathcal{N}(0, \mathbf{\Sigma}_S), \mathcal{N}(0, \mathbf{\Sigma}_T))^2 = \frac{\big(\mathrm{Tr}(\mathbf{\Sigma}_S) - \mathrm{Tr}(\mathbf{\Sigma}_T)\big)^2 + 2\sum_{i=1}^{d_y}\big(\lambda_i(\mathbf{\Sigma}_S) - \lambda_i(\mathbf{\Sigma}_T)\big)^2}{d_y(d_y+2)},$$

see Example 4.3, and

$$\mathsf{IGW}(\mathcal{N}(0, \mathbf{\Sigma}_S), \mathcal{N}(0, \mathbf{\Sigma}_T))^2 = \sum_{i=1}^{d_y}\big(\lambda_i(\mathbf{\Sigma}_S) - \lambda_i(\mathbf{\Sigma}_T)\big)^2,$$

see Corollary 3.2 in [35]. We recall that, by convention, $\lambda_i(\mathbf{\Sigma}_\mu) = 0$ for each $i > d_x$ above. We underscore that $\mathbf{\Sigma}_S$ and $\mathbf{\Sigma}_T$ are the empirical covariance matrices from the data and so work well when the datasets are approximately normal. We also compare with the centered kernel alignment,

$$\mathrm{CKA}(X^S, Y^T) = \frac{\|(Y^T)^\intercal X^S\|_{\mathrm{F}}^2}{\|(X^S)^\intercal X^S\|_{\mathrm{F}}\|(Y^T)^\intercal Y^T\|_{\mathrm{F}}}, \tag{20}$$

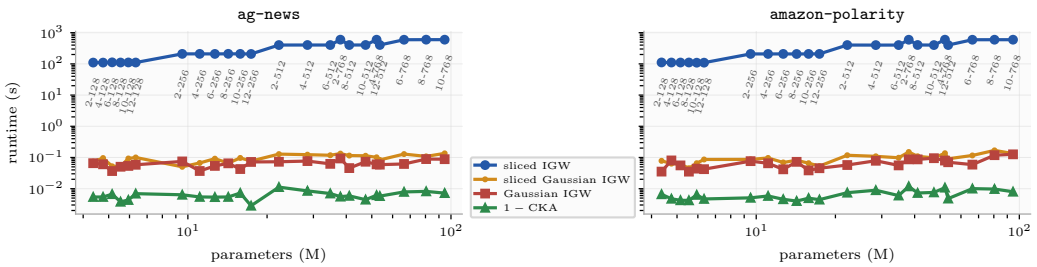


Figure 7: Time required to compute the various metrics between the student and teacher embeddings to generate the data presented in Fig. 2. The total compute time required to generate this data is 15643 seconds on a laptop with 64 GB of ram and an Intel i7-13620H processor; generating the embeddings took 1016 seconds whereas computing the values of the metrics took 14627 seconds in total.

and report $1 - \mathrm{CKA}(X^S, Y^T)$ so that smaller values correspond to stronger agreement with the teacher representation. We report the runtime of each metric for each student-teacher pair and for each choice of dataset in Fig. 7.

G.4 Heterogeneous User Clustering

In addition to the `ag_news` dataset considered previously, this experiment also uses the `imdb` and `arXiv-summarization` text datasets loaded using the Hugging Face Datasets package for Python [95]. The `arXiv-summarization` dataset entry on Hugging Face lists no license details whereas the `imdb` dataset entry lists its license as "other". We note that the original repo for `arXiv-summarization` and the accompanying paper [81], `https://github.com/armancohan/long-summarization`, downloads the datasets directly from the arXiv OpenAccess repository and that all metadata from arXiv (including the abstracts in the dataset) have the CC0 1.0 license. As for the `imdb` dataset, the original paper [82] states that "In the interest of providing a benchmark for future work in this area, we release this dataset to the public"; we understand our use of the dataset to be in line with this declaration.

Regarding the models used to generate the embeddings, we utilize `all-MiniLM-L6-v2` and `all-mpnet-base-v2` from the Hugging Face transformers Python package [97], which generate 384 and 768-dimensional embeddings respectively. These models are both listed under apache-2.0 licenses on Hugging Face. The embeddings are generated in the same manner as described in Appendix G.3 following the homogeneous and heterogeneous settings described in the text.

We underscore that the styles are ordered according to an index $0, 1, 2$ and, to form the users for the style at index 0, say, the primary-style documents are assigned from a seeded random permutation without replacement. Then, a number, k, between 0 and 10 is selected uniformly at random which corresponds to the number of samples from the non-primary style 1, and finally $10 - k$ documents are selected from the non-primary style 2.

With the simulated users in hand, we compute the pairwise distances between the embedded datasets again using the same implementation of the Riemannian subgradient method described before, but with $m = 200$ slices and a maximum number of iterations of 200 to speed up computation. This enables us to construct a pairwise distance matrix $\mathbf{D} \in \mathbb{R}^{24\times 24}$ to which we apply self-tuning spectral clustering as described in the main text to cluster the data. To measure the accuracy of the clustering, we report the adjusted Rand index [84] which is a standard chance-corrected agreement score between the predicted and ground-truth partitions. We also report purity which is the fraction of users assigned to the majority ground-truth label within their predicted cluster summed over predicted clusters. The MDS coordinates in Fig. 3 are used only for visualization and are computed from the sliced-IGW distance matrix in each setting. Clustering is performed from each method's own pairwise distance matrix, not from the MDS coordinates. Prediction labels are permutation-aligned to the ground-truth style labels for display.

We perform the same procedure using the pairwise distances generated by the sliced IGW between Gaussians whose matrices are given by the empirical covariance matrices by complete analogy with Appendix G.3.

	$\phi = 0°$	$\phi = 15°$	$\phi = 30°$	$\phi = 45°$	$\phi = 60°$	$\phi = 90°$
$v_{\mathrm{id},\phi}$	0	0.03	0.11	0.21	0.31	0.43
$v_{\mathrm{opt},\phi}$	0	8.85×10^{-12}	2.56×10^{-12}	1.50×10^{-12}	6.39×10^{-13}	7.42×10^{-12}

Figure 8: Values obtained for $v_{\mathrm{id},\phi}$ and $v_{\mathrm{opt},\phi}$ for different values of ϕ. Once $\phi \neq 0°$, $v_{\mathrm{id},\phi}$ is non-zero whereas $v_{\mathrm{opt},\phi}$ remains close to 0 for each value of ϕ considered.

We note that the time required to compute each distance is not as straightforward to plot as in the previous examples since the dimensions vary, leading to drastically different runtimes.

For the three seeds considered, the pairwise sliced IGW distances took a total of $10453, 34423$, and 34696 seconds to compute, respectively, in the homogeneous setting and $84504, 71983$, and 62987 seconds in the heterogeneous. The corresponding Gaussian IGW distances took a total of $6, 39$, and 36 seconds in the homogeneous case versus $41, 58$, and 45 seconds in the heterogeneous case. In addition to these runs, the embeddings took 1344 seconds to generate. In sum, this experiment required a total of 300615 seconds of compute time. As in the validation of the sample complexity, we run 10 Python instances in parallel to compute the pairwise distances in order to reduce the real world time spent on this experiment.

H Additional Experiments

This section compiles some additional experiments which are not part of the main suite described in the text. Experiments in this section were run on a desktop computer with 64 GB of ram and a Ryzen 9 PRO 7945 processor. To speed up computation, we use multiprocessing to perform multiple experimental runs concurrently. In this section compute time denotes the total compute time (i.e., the sum of runtimes over all threads) and not the wallclock time.

H.1 Necessity of Optimization over the Stiefel Manifold

As noted after Definition 4.1, the minimization over the Stiefel manifold is introduced to make the sliced IGW distance invariant to orthogonal transformations. To supplement the example of comparing μ, the uniform distribution over the rectangle $[-1, 1] \times [-2, 2]$ and ν_ϕ, which is a rotation of μ by an angle $\phi \in \{0°, 15°, 30°, 45°, 60°, 90°\}$, we compile the values of the squared average-sliced IGW distance for this example with and without the minimization over the Stiefel manifold in Fig. 8. Notably, if the optimization over the Stiefel manifold is removed, the distance between the two distributions is non-zero for each $\phi \neq 0°$.

Explicitly, we sample 256 points uniformly at random from $[-1, 1] \times [-2, 2]$ which we endow with equal weights to obtain an empirical version, $\hat{\mu}$, of μ, then we rotate $\hat{\mu}$ by the angle ϕ to obtain $\hat{\nu}_\phi$, a proxy for ν_ϕ. Next, we draw 200 i.i.d. samples, $(\Theta_i)_{i=1}^{200}$, from the uniform distribution on the circle and compare the values of

$$v_{\mathrm{id},\phi} \coloneqq \frac{1}{200} \sum_{i=1}^{200} \mathsf{IGW}((\Theta_i^\intercal)_\sharp \hat{\mu}, (\Theta_i^\intercal)_\sharp \hat{\nu}_\phi)^2,$$

$$v_{\mathrm{opt},\phi} \coloneqq \inf_{\mathbf{\Delta} \in \mathrm{St}(2,2)} \frac{1}{200} \sum_{i=1}^{200} \mathsf{IGW}((\Theta_i^\intercal \mathbf{\Delta})_\sharp \hat{\mu}, (\Theta_i^\intercal)_\sharp \hat{\nu}_\phi)^2.$$

We approximate the value of $v_{\mathrm{opt},\phi}$ by optimizing the objective using the Riemannian subgradient method, Algorithm 3, initialized at the identity with a maximum of 200 iterations. The total compute time for this simple experiment is 0.29 seconds.

H.2 Comparison Between Methods

This section compares the performance of different methods for aligning distributions. Namely, we consider the IGW distance (without slicing) solved using the conditional gradient method as implemented in [98], the sliced quadratic GW distance [33], and Wasserstein–Procrustes following the implementation from [99].

Fraction of runs with $< 10^{-6}$

n	Sliced IGW	IGW	Wasserstein–Procrustes	Sliced Quadratic GW
10	5/5	3/5	4/5	5/5
50	5/5	2/5	5/5	5/5
100	5/5	2/5	4/5	5/5
500	5/5	2/5	5/5	5/5
1000	5/5	2/5	5/5	5/5
2500	5/5	1/5	5/5	5/5
5000	5/5	0/5	4/5	5/5

Figure 9: Fraction of the 5 total runs for different choices of n that terminate with value less than 1×10^{-6} for each metric. The true value for each metric is 0.

We generate the datasets by first drawing N samples from the bivariate Gaussian $\mathcal{N}\left(0, \begin{pmatrix} 0.36 & 0 \\ 0 & 1.96 \end{pmatrix}\right)$. The second dataset is obtained by rotating the first one by 60° and permuting its labels according to a random permutation so that no solver has an advantage based on initialization. Consequently, the distance between these datasets is 0 under each of these metrics. We apply each of the methods stated above to compare datasets of size $n \in \{10, 50, 100, 500, 1000, 2500, 5000\}$. We compile the number of successful runs (defined by obtaining a metric value smaller than 10^{-6}) over 5 repetitions in Fig. 9 and the runtime as a function of the dataset size in Fig. 10. Notably, all of the methods scale similarly except for the Wasserstein–Procrustes solver whose runtime increases faster as n increases. While sliced IGW and sliced quadratic GW succeed for all runs, the IGW and Wasserstein–Procrustes solvers fail for certain runs; this could be mitigated by trying additional initializations for instance. The total compute time for this experiment is 3362.18 seconds.

For the sliced IGW distance, we again use the Riemannian method as before, but decrease the gradient tolerance to 10^{-8} to obtain more accurate results. Additionally, we use $m = 200$ MC slices.

For the IGW distance, we modify the method slightly from the original implementation in [98] to leverage the particular structure of matrices of inner products which enables accelerating matrix-vector and matrix-matrix products. (see [42]). We initialize this method at the independent coupling $\mu \otimes \nu$ which is a canonical choice if no structure is assumed on the problem. We set the maximum number of iterations to 250 and both stopping thresholds to 10^{-10}. Moreover, the line search procedure from that method can be solved directly in this setting.

Wasserstein–Procrustes is solved via Algorithm 1 in [99] with the Fiedler-W initialization as discussed in the same work.

For the sliced quadratic Gromov-Wasserstein distance, we follow the implementation supporting their paper which uses Riemannian Adam to perform the optimization over the Stiefel manifold. We modify the default implementation slightly by setting the learning rate to 0.03 and the maximum number of iterations to 750. Again, we use $m = 200$ MC slices and, as with our method, initialize at the identity.

H.3 Monte-Carlo Error and Convergence Rate for Heavy-Tailed Distributions

As noted in Corollary 4.7, if $\mu \in \mathcal{P}_q(\mathbb{R}^{d_x})$ and $\nu \in \mathcal{P}_q(\mathbb{R}^{d_y})$ for $q \in (2, 4)$, the empirical sliced-IGW distance based on n samples converges to its population-level counterpart at the rate of $n^{-(q-2)/(2q)}$ in expectation. We verify this scaling experimentally in Fig. 11 for a distribution which admits finite moments for each $q < 2.9$ so that the expected rate of convergence is of order $n^{-\alpha}$ for any $\alpha < \frac{0.9}{5.8} \approx 0.1552$.

Indeed, we set $\mu \in \mathcal{P}(\mathbb{R}^5)$ to be the law of

$$\operatorname{diag}(0.65, 0.825, 1, 1.175, 1.35)\frac{G}{\sqrt{V/2.9}} + (B - 0.25)\begin{pmatrix}-0.7 & -0.25 & 0.2 & 0.65 & 1.1\end{pmatrix}^\intercal$$

where $G \sim \mathcal{N}(0, \mathrm{id})$, $V \sim \chi^2_{2.9}$ (the chi-squared distribution with 2.9 degrees of freedom), and $B \sim \text{Bernoulli}(0.25)$ which are sampled independently. By definition, each component of $\frac{G}{\sqrt{V/2.9}}$ has a Student-t distribution with 2.9 degrees of freedom. $\nu \in \mathcal{P}(\mathbb{R}^{10})$ is obtained by setting $\nu = Q_\sharp \mu$

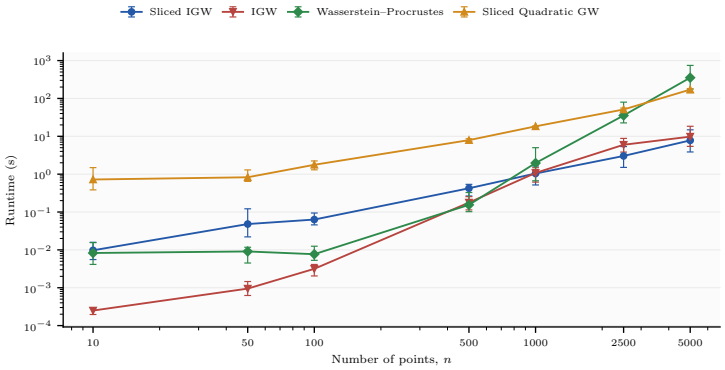


Figure 10: Runtime of each method as a function of the number of points in each dataset. Points indicate means and error bars indicate the maximum and minimum over 5 runs for each method and each choice of n.

where Q is a fixed randomly generated element of $\mathrm{St}(5, 10)$, explicitly,

$$Q = \begin{pmatrix} -0.4280 & 0.6578 & -0.1029 & -0.1446 & 0.0643 \\ 0.7056 & 0.2504 & -0.2616 & 0.4067 & -0.0543 \\ 0.3975 & -0.2287 & 0.2185 & -0.5363 & -0.0399 \\ 0.2337 & 0.3051 & -0.2751 & -0.4556 & -0.3165 \\ 0.0941 & -0.0803 & 0.4057 & 0.0357 & -0.2736 \\ 0.1304 & 0.0894 & 0.1905 & 0.2531 & 0.1298 \\ -0.0214 & 0.2126 & -0.0891 & 0.0292 & -0.6878 \\ -0.2728 & -0.4122 & -0.0715 & 0.1381 & -0.5437 \\ -0.0672 & -0.2904 & -0.6753 & 0.1865 & 0.0606 \\ -0.0332 & 0.2128 & 0.3571 & 0.4454 & -0.1649 \end{pmatrix},$$

where we have again truncated to 4 digits after the decimal.

Therefore, by construction, $\mu \in \mathcal{P}_q(\mathbb{R}^5)$, $\nu \in \mathcal{P}_q(\mathbb{R}^{10})$, for each $q < 2.9$ and $\underline{\mathrm{IGW}}(\mu, \nu) = 0$. We compute an estimate, $\widehat{\underline{\mathrm{IGW}}}_{m,n}$, of $\underline{\mathrm{IGW}}(\mu, \nu)$ by following the same procedure described in Section 5. The total compute time for this experiment is 785102.51 seconds.

H.4 Ablation Study for Heterogeneous Clustering Experiment

This section serves to empirically verify that the results of Fig. 3 are not an artifact of the experimental design and that the advantage that sliced IGW exhibits over Gaussian IGW persists across changes to the various experimental parameters. Notably, we report in Fig. 12 the effects of changing 1) the composition of primary/secondary styles, 2) the number of documents per user, 3) the number of users per style, 4) the fraction of users within each style embedded using MPNet vs MiniLM, and 5) the

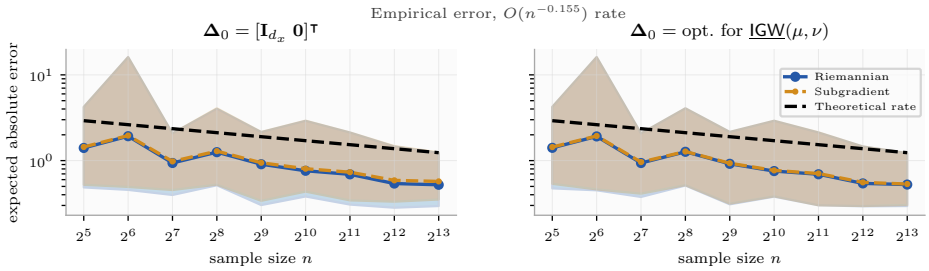


Figure 11: **Left:** Validation of the empirical convergence rate from Corollary 4.7. For a fixed number of Monte Carlo slices $m = 3000$, the estimated value of $\mathbb{E}[|\widehat{\underline{\mathrm{IGW}}}_{m,n}|]$ is seen to scale at least as $C_1 + C_2 n^{-0.155}$. In this plot the shaded area corresponds to the maximum and minimum values from the 25 realizations.

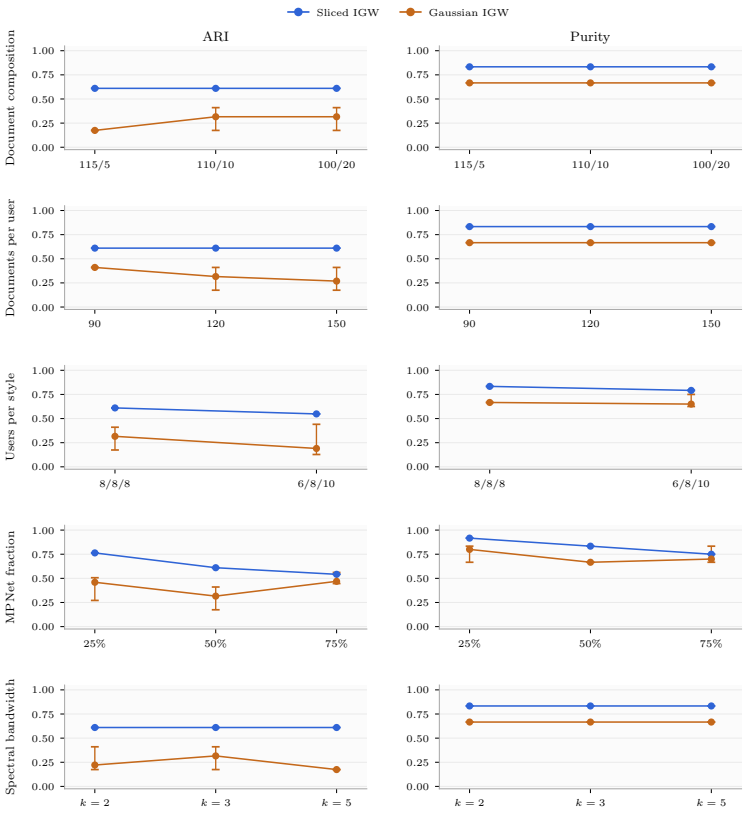


Figure 12: Results of the ablation study with each panel as described above. Points represent the mean over 5 seeds and the error bars illustrate the minimum and maximum over the 5 seeds.

bandwidth used in the spectral clustering algorithm. When we vary each of these parameters, we keep the other ones fixed to those of the original experiment as described in Section 5. The sliced IGW is optimized using the same Riemannian optimization method as the original clustering experiment. The ablation study indicates that sliced IGW outperforms Gaussian IGW over the different choices of parameters for the range of modifications considered. The total runtime for this experiment is 2994719.72 seconds.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Each of the claims made in the abstract and introduction are contained within the main text. Namely, we 1) furnish a formula for the univariate inner product Gromov-Wasserstein distance, 2) analyze a slicing framework using this formula encompassing topological and metric properties, quality of empirical approximation, and analysis of a principled optimization algorithm, and 3) validate the derived theory empirically and provide applications to comparing LLM embeddings and heterogeneous clustering.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper clearly addresses the limitations/assumptions required for the derived theory which is the primary contribution of this contribution. We have honestly stated the limitations with the asymptotic convergence of the algorithm and specified its per iteration complexity in the case of equally weighted measures on the same (finite) number of points under appropriate conditions in the appendix.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best

judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: All of the theoretical results are clearly stated with the full set of assumptions specified. Complete proofs of all of the results in the text are compiled in the appendix. Short summaries of the proofs are included in the main text to aid readability.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The methodology used to obtain the experimental results is clearly explained; additional details are contained in the appendix to aid reproducibility. The supplementary material also contains a zip file with all of the code/resources used to generate the experimental results.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.

(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The code used to generate all of the experimental results is included in the supplementary material as well as all of the resources used. The appendix also explains how the models/datasets were accessed.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The appendix specifies all of the different hyperparameters used to obtain the various results. The code is also available in the supplemental material, where all of these parameters are hardcoded or clearly stated.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: The results we present are not expected to abide by a normal distribution and hence error bars based on standard deviation are difficult to interpret. We have thus presented average values along with maximum and minimum values among repetitions in all of the plots where relevant. For experiments with a low number of repetitions, all values were stated explicitly in place of a summary statistic (e.g., the mean).

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The appendix includes details on compute resources including the specifications of the machines used to generate the experimental results as well as the compute time for each run and the total run time for each experiment.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have reviewed the Code of Ethics and confirm that the research abides, in every respect, with this code.

   Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: The paper proposes a contribution of a mostly theoretical nature. In effect, we propose a distance between probability measures which, on its own, has no clear social impact. The applications we develop, comparison of LLM representations and heterogeneous clustering also do not appear to introduce any additional positive or negative impacts.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release data or models with a high risk of misuse.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We have explicitly cited the creators of the assets used within the paper and clarified the terms of use and licenses associated with each of these assets in the appendix to the best of our abilities.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: This paper does not release new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.

- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs are not an important, original, or non-standard component of the core methods in this research.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.