COMPUTATIONAL AND STATISTICAL PROPERTIES OF OPTIMAL TRANSPORT-BASED DISTANCES

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Gabriel Rioux

August 2025

COMPUTATIONAL AND STATISTICAL PROPERTIES OF OPTIMAL TRANSPORT-BASED DISTANCES

Gabriel Rioux, Ph.D.

Cornell University 2025

Over the past two decades, the statistical and computational properties of optimal transport (OT) have been systematically studied, driven in part by OT's broad applicability across data science, statistics, economics, physics, and other fields. This body of work has identified the curse of dimensionality inherent in statistical estimation with OT distances and has led the development of computationally and statistically efficient regularizations of OT. The first contribution of this thesis is a general framework for deriving limit distributions and other asymptotic properties of empirical regularized OT distances.

While OT distances enable a natural comparison between distributions on a common space, comparing datasets of different types – such as text and images – requires specifying an ad hoc cost function, which may fail to capture a meaningful correspondence between data points. To address this issue, Gromov-Wasserstein (GW) distances have been proposed, enabling a comparison of metric measure spaces based on their intrinsic structure. As a result, GW distances have found widespread use in applications involving heterogeneous data. The second contribution of this thesis is to establish the first limit laws for empirical GW distances and to derive consistent resampling schemes.

Given the broad applicability of GW distances, their computation is of a particular interest. However, existing algorithms for computing regularized GW distances lack comprehensive convergence rate guarantees. To address this gap, the final contribution of this thesis is to introduce the first algorithms for computing regularized GW distances subject to formal non-asymptotic convergence rate guarantees.

BIOGRAPHICAL SKETCH

Gabriel Rioux is a graduate student at the Center for Applied Mathematics at Cornell University. Before joining Cornell, he earned a Bachelor's degree in mathematics and physics and a Master's degree in mathematics and statistics from McGill University in his hometown of Montréal. His principal research interests lie in mathematical statistics, applied probability, and optimization, with a particular focus on the interplay between these areas. His graduate work has centered on establishing the statistical and computational properties of regularized optimal transport distances and Gromov-Wasserstein distances as well as exploring applications of optimal transport in statistics and data science.

To my family and friends, for their endless support.

ACKNOWLEDGEMENTS

I am deeply indebted to my advisor, Prof. Ziv Goldfeld, and to Prof. Kengo Kato for their support and invaluable advice throughout the degree, and for being excellent collaborators. I am also sincerely grateful to Prof. Yunan Yang for graciously rounding out my committee.

I am beholden to Prof. Rustum Choksi and Prof. Tim Hoheisel for their continued encouragement and for giving me my first taste of research, effectively setting me on this path. To my collaborators, thank you for for all you have taught me and for the ideas we have shared. To my friends, any thanks that I can express here would be a huge understatement; just know that none of this would have been possible without you.

My graduate work was supported, in part, by the National Sciences and Engineering Research Council of Canada's Postgraduate Fellowship PGSD-567921-2022.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1
INTRODUCTION

This thesis comprises three papers: two have been published in journals in applied probability and statistics, and the third is currently under review. This first chapter serves as a brief primer to optimal transport theory and the related Gromov-Wasserstein distance. The second chapter develops a generic framework for deriving limit distributions for optimal transport-based distances and applies it to three different regularizations of optimal transport. The third chapter applies this framework to derive the first limit distribution results for the Gromov-Wasserstein distance and proposes an application to the problem of testing for graph distribution isomorphisms. The fourth chapter regards computation of entropically regularized Gromov-Wasserstein distances and introduces the first algorithms for this problem with non-asymptotic convergence rate guarantees.

1.1 Optimal Transport and the Gromov-Wasserstein Distance

Optimal transport (OT) has a storied history, originating nearly 250 years ago with the pioneering work of Gaspard Monge [146]. At its core is the purely operational question of how to transport particles (e.g. grains of sand) into a desired configuration (e.g. a sand castle) in a way that the average displacement of the particles be minimized. The modern mathematical formulation of Monge's original problem is as follows. Given probability measures μ,ν on possibly distinct Polish spaces $\mathcal{X},\mathcal{Y}$ and a cost function $c:\mathcal{X}\times\mathcal{Y}\to\mathbb{R}$ which quantifies the cost of transporting a unit mass from a point $x\in\mathcal{X}$ to its target $y\in\mathcal{Y}$, we wish to find a measurable map $T:\mathcal{X}\to\mathcal{Y}$ which transports μ onto ν, in the sense that $\mu\left(T^{-1}(A)\right)=\nu(A)$ for every measurable set $A\subset\mathcal{Y}$ (this relation is denoted by

$T_\sharp\mu = \nu$), and solves the following problem

$$\inf_{\substack{T \text{ measurable} \\ T_\sharp\mu=\nu}} \int_{\mathcal{X}} c\,(x, T(x))\, d\mu(x).$$

A map solving this problem is known as an OT map and assigns to each point $x \in \mathcal{X}$ its optimal target $y \in \mathcal{Y}$ under the cost c.

Regrettably, Monge's problem is quite brittle, as the existence of a map T satisfying $T_\sharp\mu = \nu$ is not guaranteed even in anodyne settings. For instance, if μ is supported on one point whereas ν is supported on two points, no such map can exist. It was not until the work of Leonid Kantorovich [120] that a satisfactory resolution of this problem was proposed. Rather than imposing that the transport be described by a map, Kantorovich's problem allows mass splitting by working with couplings – probability measures on the product space $\mathcal{X} \times \mathcal{Y}$ which admit μ and ν as marginals on $\mathcal{X}$ and $\mathcal{Y}$ respectively. Namely, we seek $\pi \in \Pi(\mu, \nu)$, the set of all couplings of μ and ν, which minimizes the expected cost of transportation, yielding the problem

$$\inf_{\pi\in\Pi(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{Y}} c(x, y) d\pi(x, y).$$

Crucially, the existence of couplings solving this problem, called OT plans, is guaranteed under mild conditions. A connection between the Monge and Kantorovich problems is afforded in the case that an OT plan is concentrated on the graph of a map, T, in which T is an OT map. When μ and ν are supported on a finite number of points, the Kantorovich problem reduces to a linear program for which a variety of efficient (polynomial time) algorithms are available.

Perhaps the most important instance of the OT problem arises when μ and ν are probability measures on a common Polish space $(\mathcal{X}, \mathsf{d})$, and the cost function is taken to be the p-th power of the metric for some $p \in [1, \infty)$. In this setting, the quantity

$$\mathsf{W}_p(\mu, \nu) := \left(\inf_{\pi\in\Pi(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{X}} \mathsf{d}(x, y)^p d\pi(x, y)\right)^{1/p},$$

is called the p-Wasserstein distance between μ and ν. This moniker is well-deserved, as W_p is a metric on the space of all probability measures on $\mathcal{X}$ which have finite p-th moment. Furthermore, it metrizes the weak convergence of probability measures together with the convergence of p-th moments. Importantly, Wasserstein distances endow the space of distributions with a rich geometry, enabling practitioners to reason about complicated transformations of distributions in terms of concrete geometric operations.

Given that the Wasserstein distance vanishes if and only if the input measures are identical, it is natural to consider its application to statistical problems such as testing for equality of distributions based on samples. Concretely, suppose that we observe independent and identically distributed samples $X_1, \ldots, X_n$ from an unknown probability distribution μ on $\mathbb{R}^d$, which we assume to be compactly supported for simplicity. We aim to test whether μ coincides with a reference distribution ν by computing the statistic $\mathsf{W}_2(\hat{\mu}_n, \nu)$, where $p = 2$ is chosen for simplicity and $\hat{\mu}_n := \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ is the empirical measure of the samples, with δ_x denoting the Dirac measure at $x \in \mathbb{R}^d$. The null hypothesis $\mathsf{W}_2(\mu, \nu) = 0$ is then rejected if $\mathsf{W}_2(\hat{\mu}_n, \nu)$ exceeds a set critical value.

The standard approach to computing critical values for this testing problem requires deriving a central limit theorem under the null and estimating the quantiles of the limiting distribution. However, once $d \geq 4$, the quantity $\mathbb{E}[\mathsf{W}_2(\hat{\mu}_n, \mu)^2]$ is generally of order $O\left(n^{-2/d}\right)$, up to logarithmic factors when $d = 4$, (cf. e.g. [83]). This scaling suggests that a central limit theorem with the conventional $\sqrt{n}$ scaling should not hold in general for $d \geq 4$, see Section 5.4 in [117] for details. The poor scaling of the mean rate of convergence with the dimension is often referred to as the *curse of dimensionality* and is an important consideration when applying OT-based methods to sampled data. To address this issue, computationally and statistically efficient regularizations of OT have seen much interest in the past decade. A complete introduction to regularized OT is included in Chapter 2 in which a platform for valid statistical inference with regularized distances is developed.

While OT is particularly well-suited for comparing distributions supported on the same space, modern data science pipelines often involve comparing datasets that differ in modality or semantic meaning. Applying OT in such settings requires defining a cost function on the joint space which accurately captures the similarities and dissimilarities of heterogeneous data – a task which is often nontrivial. To account for this, practitioners have recently turned to the Gromov-Wasserstein (GW) distance which enables comparing arbitrary metric measure spaces based solely on their intrinsic structure.

Given two complete and separable metric measure spaces $(\mathcal{X}, \mathsf{d}_{\mathcal{X}}, \mu)$ and $(\mathcal{Y}, \mathsf{d}_{\mathcal{Y}}, \nu)$, the (p, q)-GW distance, for $p, q \in [1, \infty)$, between them is defined as

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

and is finite provided that μ and ν have finite pq-size, that is,

$$\int_{\mathcal{X}} \int_{\mathcal{X}} \mathsf{d}_{\mathcal{X}}(x, x')^{pq} d\mu(x) d\mu(x') < \infty \text{ and } \int_{\mathcal{Y}} \int_{\mathcal{Y}} \mathsf{d}_{\mathcal{Y}}(y, y')^{pq} d\nu(y) d\nu(y') < \infty.$$

It is straightforward to verify that $\mathsf{GW}_{p,q}(\mu, \nu) = 0$ if and only if there exists a measurable map $T : \mathcal{X} \to \mathcal{Y}$ for which $T_\sharp \mu = \nu$ and $\mathsf{d}_{\mathcal{X}}(x, x') = \mathsf{d}_{\mathcal{Y}}(T(x), T(x'))$ for μ-almost every $x \in \mathcal{X}$. In this case, $(\mathcal{X}, \mathsf{d}_{\mathcal{X}}, \mu)$ and $(\mathcal{Y}, \mathsf{d}_{\mathcal{Y}}, \nu)$ are isomorphic metric measure spaces. Remarkably, $\mathsf{GW}_{p,q}$ defines a metric on the collection of all isomorphism classes of metric measure spaces with finite pq-size.

In contrast to the standard Wasserstein distance, the statistical properties of the GW distance remained widely unknown until the recent work of [223]. In that paper, it was shown that if μ and ν are compactly supported probability measures on $\mathbb{R}^{d_x}$ and $\mathbb{R}^{d_y}$ respectively with $d_x, d_y \geq 4$, $\mathbb{E}\left[\left|\mathsf{GW}_{2,2}(\hat{\mu}_n, \hat{\nu}_n)^2 - \mathsf{GW}_{2,2}(\mu, \nu)^2\right|\right] = O(n^{-2/\min\{d_x, d_y\}})$, up to logarithmic factors when $\min\{d_x, d_y\} = 4$. This rate matches the curse of dimensionality rate of W_2^2 when $d_x = d_y$, but adapts to the minimum of the two dimensions – a property which also holds for W_2 (see [21]) – suggesting that the Wasserstein and Gromov-Wasserstein distances share similar statistical properties. Chapter 3 strengthens this heuristic by deriving the first limit laws for GW distances in many cases of interest using the framework developed in Chapter 2.

While the Wasserstein and Gromov-Wasserstein distances behave similarly at the statistical level, their computational properties are quite dissimilar. Indeed, OT between two finitely supported distributions is simply a linear program. By contrast, computing the GW distance entails solving a quadratic program which is generally nonconvex – a class of problems known for their intractability. Although local solvers and semidefinite programming relaxations for this problem have been proposed, exact resolution of the GW problem remains challenging in most cases. Alternative approaches seek to estimate a regularized Gromov-Wasserstein distance via local optimization techniques. These methods are appealing due to their inexpensive per-iteration cost, but alas there are no currently guarantees on the number of iterations required for convergence. Chapter 4 introduces the first algorithm for approximating a regularized Gromov-Wasserstein distance which is subject to formal convergence rate guarantees.

1.2 Contributions

The first contribution of this work, provided in Chapter 2, is the development of a general framework for establishing limit theorems, bootstrap consistency, and semiparametric efficiency for OT-based distances. This methodology is applied to three distinct regularization schemes for OT and resolves a number of open questions concerning the statistical properties of sliced Wasserstein distances. The approach is based on the delta method, which has proven effective in deriving limit theorems for OT distances [187, 190, 117]. Following its introduction in [106], this methodology was applied in [105, 166] to regularized versions of certain quantities of interest in OT, including optimal maps, dual potentials, and an index of stochastic dominance.

Beyond standard OT, this approach also enabled deriving limit laws for the empirical Gromov-Wasserstein distance in three settings of interest, namely 1) for finitely supported distributions, 2) in the semidiscrete setting, and 3) for compactly supported distributions albeit with an entropically regularized distance. Although the overall technique is similar to that used for regularized OT, each case requires a substantially different approach for proving that the required conditions hold and the analysis differs substantially from OT due to the quadratic nature of the underlying problem. For distributions with finite support, a consistent scheme for sampling from the null distribution is also proposed. Notably, this scheme requires solving random linear programs and can thus be implemented efficiently whereas methods based on bootstrapping would require solving quadratic programs which are generally intractable. Leveraging this, the empirical Gromov-Wasserstein distance is proposed as a test statistic for assesing whether two collections of random graphs are generated from the same distribution under the inhomogeneous Erdős-Rényi model, up to a relabeling of the nodes. This constitutes the second contribution of this work, detailed in Chapter 3.

While central limit theorems are now available for the empirical Gromov-Wasserstein distance in various settings, accurately computing the test statistic, which is essential for valid statistical inference, remains onerous. Chapter 4 contains the final contribution of this thesis, the first algorithms for approximating the entropically regularized Gromov-Wasserstein distance with known non-asymptotic convergence rates. While these algorithms are generally only guaranteed to converge to local solutions of the regularized Gromov-Wasserstein problem, they serve as a first step towards principled and efficient estimation of Gromov-Wasserstein distances.

CHAPTER 2

STATISTICAL INFERENCE WITH REGULARIZED OPTIMAL TRANSPORT

This chapter is based on the paper "Statistical Inference with Regularized Optimal Transport" [106], published in *Information and Inference: A journal of the IMA*. This is joint work with Ziv Goldfeld, Kengo Kato, and Ritwik Sadhu.

2.1 Introduction

Optimal transport (OT) theory [208, 175] provides a versatile framework for comparing probability distributions. Introduced by Monge [146] and later formulated by Kantorovich [120], the OT problem between two Borel probability measures μ, ν on $\mathbb{R}^d$ is defined by

$$\mathsf{T}_c(\mu,\nu) := \inf_{\pi\in\Pi(\mu,\nu)} \int_{\mathbb{R}^d\times\mathbb{R}^d} c(x,y)d\pi(x,y), \tag{2.1}$$

where $\Pi(\mu,\nu)$ is the set of couplings between μ and ν. The special case of the p-Wasserstein distance for $p \in [1,\infty)$ is given by $\mathsf{W}_p(\mu,\nu) := (\mathsf{T}_{\|\cdot\|^p}(\mu,\nu))^{1/p}$. Thanks to an array of favorable properties, including the Wasserstein metric structure, a convenient duality theory, robustness to support mismatch, and the rich geometry induced on the space of probability measures, OT and the Wasserstein distance have seen a surge of applications in statistics, machine learning, and applied mathematics. These include generative modeling [10, 111, 193, 16, 45], robust/adversarial machine learning (ML) [20, 213], domain adaptation [185, 52], image recognition [170, 174, 132], vector quantile regression [36, 46, 91, 113], Bayesian estimation [15], and causal inference [194]. Unfortunately, OT distances are generally hard to compute and suffer from the curse of dimensionality in empirical estimation, whereby the number of samples needed for reliable estimation grows exponentially with dimension.

These deficits have motivated the introduction of regularized OT methods that aim to alleviate the said computational and statistical bottlenecks. Three prominent regularizations are: (1) slicing via lower-dimensional projections [163, 28, 148, 12, 147]; (2) smoothing via convolution with a chosen kernel [102, 100, 103, 101, 152, 172, 221, 40, 114, 105, 22]; and (3) convexification via entropic penalty [179, 130, 53, 4, 87, 142, 65]. These techniques preserve many properties of classic OT but avoid the curse of dimensionality, which enables a scalable statistical theory. As reviewed below[1], much effort was devoted to exploring dimension-free empirical convergence rates and limit distributions, bootstrapping, and other statistical aspects of regularized OT, although several notable gaps in the literature remain. Furthermore, proof techniques for such results are typically on a case-by-case basis and do not follow a unified approach, despite evident similarities between the three regularization methods as complexity reduction techniques of the classic OT framework.

The present paper develops a unified framework for deriving limit distributions, semiparametric efficiency bounds, and bootstrap consistency for a broad class of functionals that, in particular, encompasses the empirical regularized OT distances mentioned above (Section 2.3). As example applications of the general framework, we explore a comprehensive treatment of the following problems:

- **Average- and max-sliced W_p (Section 2.4):** Our limit distribution theory closes existing gaps in the literature (e.g., a limit distribution result for sliced W_1 was assumed in [147] but left unproven), with the efficiency and bootstrap consistency results providing additional constituents for valid statistical inference.
- **Smooth W_p with compactly supported kernels (Section 2.5):** Gaussian-smoothed OT was previously shown to preserve the classic Wasserstein structure while alleviating the curse of dimensionality. Motivated by computational considerations,

[1]We postpone the literature review on each regularization method to its respective section.

herein we study smoothing with compactly supported kernels. We explore the metric, topological, and statistical aspects previously derived under Gaussian smoothing.

- **Entropic OT (Section 2.6):** A central limit theorem (CLT) for empirical entropic OT (EOT) was derived [142, 65] for independent data via a markedly different proof technique than proposed herein. Revisiting this problem using our general machinery, we rederive this CLT allowing for dependent data, and also obtain new results on semiparametric efficiency and bootstrap consistency.

The unified limit distribution framework, stated in Proposition 1, relies on the extended functional delta method for Hadamard directionally differentiable functionals [181, 169]. To match the delta method with the regularized OT setup, we focus on a functional on a space of probability measures that is (a) locally Lipschitz with respect to (w.r.t.) the sup-norm for a Donsker function class and (b) Gâteaux directionally differentiable at the population distribution. To apply this framework, we seek to: (i) set up the regularized distance as a locally Lipschitz functional δ w.r.t. $\|\cdot\|_{\infty,\mathcal{F}} = \sup_{f\in\mathcal{F}}|\cdot|$; (ii) show $\mathcal{F}$ to be Donsker to obtain convergence of the empirical process in $\ell^\infty(\mathcal{F})$; (iii) characterize the Gâteaux directional derivative of δ at μ. For each regularized distance (sliced, smooth, and entropic), we identify the appropriate function class $\mathcal{F}$ and establish the desired Lipschitz continuity and differentiability, relying on OT duality theory. Regularization enforces the dual potentials to possess smoothness or low-dimensionality properties, which are leveraged to show that $\mathcal{F}$ is Donsker. Of note is that our framework does not require independent and identically distributed (i.i.d.) data and can be applied for any estimate (not only the empirical distribution) of the population distribution, so long as the uniform limit theorem mentioned in (ii) holds true.

As the general framework stems from the extended functional delta method, the limiting variable of the (scaled and centered) empirical regularized distance is given by the directional derivative of δ at the population distribution. Linearity of the derivative implies that the limit variable is centered Gaussian. In this case, it is natural to ask whether the empirical distance attains the semiparametric efficiency lower bound (cf. [200, Chapter 25]). Semiparametric efficiency bounds serve as analogs of Cramér-Rao lower bounds in semiparametric estimation and account for the fundamental difficulty of estimating functionals of interest. We show that the asymptotic variance of the empirical distance indeed agrees with the semiparametric efficiency bound, relative to a certain tangent space. Still, even when the limiting variable is Gaussian, direct analytic estimation of the asymptotic variance may be nontrivial. To account for that, we explore bootstrap consistency for empirical regularized OT distances. Altogether, the limit distribution theory, semiparametric efficiency, and bootstrap consistency provide a comprehensive statistical account of the considered regularized OT distances.

A unifying approach of a similar flavor to ours, but for classic OT distances, was proposed in [117]. Focusing solely on the supremum functional, they used the extended functional delta method to derive limit distributions for classic W_p, with $p \geq 2$, for compactly supported distributions under the alternative in dimensions $d \leq 3$. In comparison, our approach is more general and can treat any functional that adheres to the aforementioned local Lipschitz continuity and differentiability. This is crucial for analyzing regularized OT distances as some instances do not amount to a supremum functional. For instance, average-sliced Wasserstein distances correspond to mixed L^1-L^∞ functionals, which are not accounted for by the setup from [117]. The functional delta method was also used in [187, 190] to derive limit distributions for OT between discrete population distributions by parametrizing them using simplex vectors. This result was extended to semi-discrete OT in [62] by exploiting the fact that complexity of the optimal potentials class is reduced when one of the measures is supported on a discrete set. Another recent application can be found in [105], where this approach was leveraged for Gaussian-smoothed W_p by embedding the domain of the Wasserstein distance into a certain dual Sobolev space.

The paper is organized as follows. Section 2.2 presents notation used throughout the paper and background on Wasserstein distances and the extended functional delta method. Section 2.3 presents a unified framework for deriving limit distributions, bootstrap consistency, and semiparametric efficiency bounds for regularized OT distances. The tools developed therein will be applied to sliced Wasserstein distances in Section 2.4, smooth Wasserstein distances with compactly supported kernels in Section 2.5, and EOT in Section 2.6. Section 2.7 leaves some concluding remarks. Proofs for the results in Sections 2.2–2.6 are found in Sections A.1–A.5 of the supplementary material.

2.2 Background and Preliminaries

This section collects notation used throughout the paper and sets up necessary background on Wasserstein distances and the extended functional delta method.

2.2.1 Notation

Let $\|\cdot\|$ denote the Euclidean norm and $B(x,r)$ be the open ball with center $x \in \mathbb{R}^d$ and radius $r > 0$. For a subset A of a topological space S, let $\overline{A}^S$ denote the closure of A; if the space S is clear from the context, then we simply write $\overline{A}$ for the closure. The space of Borel probability measures on S is denoted by $\mathcal{P}(S)$. When S is a normed space with norm $\|\cdot\|_S$, we denote $\mathcal{P}_p(S) := \{\mu \in \mathcal{P}(S) : \int \|x\|_S^p d\mu(x) < \infty\}$ for $1 \le p < \infty$. The (topological) support of $\mu \in \mathcal{P}(S)$ is denoted as $\mathrm{spt}(\mu)$. For any finite signed Borel measure γ on S, we identify γ with the linear functional $f \mapsto \gamma(f) = \int f d\gamma$. For $\mu \in \mathcal{P}(S)$

and a μ-integrable function h on S, $h\mu$ denotes the signed measure $hd\mu$. Let $\stackrel{w}{\to}$, $\stackrel{d}{\to}$, and $\stackrel{\mathbb{P}}{\to}$ denote weak convergence of probability measures, convergence in distribution of random variables, and convergence in probability, respectively. When necessary, convergence in distribution is understood in the sense of Hoffmann-Jørgensen (cf. Chapter 1 in [199]). For any nonempty set S, let $\ell^\infty(S)$ be the Banach space of bounded real functions on S equipped with the sup-norm $\|\cdot\|_{\infty,S} = \sup_{x\in S}|\cdot|$. For any measure space $(S,\mathcal{S},\mu)$ and $1 \le p < \infty$, let $L^p(\mu) = L^p(S,\mathcal{S},\mu)$ denote the Banach space of measurable functions $f : S \to \mathbb{R}$ with $\|f\|_{L^p(\mu)} = (\int |f|^p d\mu)^{1/p} < \infty$. If μ is σ-finite and $\mathcal{S}$ is countably generated, then the space is separable. For two numbers a and b, we use the notation $a \wedge b = \min\{a,b\}$ and $a \vee b = \max\{a,b\}$.

2.2.2 Wasserstein distances

The Wasserstein distance is a specific instance of the OT problem from (2.1), defined as follows.

Definition 1 (Wasserstein distance)**.** Let $1 \le p < \infty$. The p-th Wasserstein distance between $\mu, \nu \in \mathcal{P}_p(\mathbb{R}^d)$ is defined as

$$\mathsf{W}_p(\mu,\nu) := \inf_{\pi\in\Pi(\mu,\nu)} \left[\int_{\mathbb{R}^d\times\mathbb{R}^d} \|x-y\|^p \, d\pi(x,y)\right]^{1/p}, \tag{2.2}$$

where $\Pi(\mu,\nu)$ is the set of couplings of μ and ν.

The p-Wasserstein distance is a metric on $\mathcal{P}_p(\mathbb{R}^d)$ and metrizes weak convergence plus convergence of pth moments, i.e., $\mathsf{W}_p(\mu_n,\mu) \to 0$ if and only if $\mu_n \stackrel{w}{\to} \mu$ and $\int \|x\|^p d\mu_n(x) \to \int \|x\|^p d\mu(x)$. Wasserstein distances admit the following dual form (cf. [208, Theorem 5.9]):

$$\mathsf{W}_p^p(\mu,\nu) = \sup_{\varphi\in L^1(\mu)} \left[\int_{\mathbb{R}^d} \varphi d\mu + \int_{\mathbb{R}^d} \varphi^c d\nu\right], \tag{2.3}$$

where $\varphi^c(y) = \inf_{x\in\mathbb{R}^d}[\|x-y\|^p - \varphi(x)]$ is the c-transform of φ (for the cost $c(x,y) = \|x-y\|^p$). A function $f : \mathbb{R}^d \to [-\infty,\infty)$ is called *c-concave* if $f = g^c$ for some function $g : \mathbb{R}^d \to [-\infty,\infty)$. There is at least one c-concave $\varphi \in L^1(\mu)$ that attains the supremum in (2.3), and we call this φ an *OT potential* from μ to ν for W_p. Further, when $1 < p < \infty$ and μ is supported on a connected set with negligible boundary and has a (Lebesgue) density, then the OT potential from μ to ν is unique on int(spt(μ)) up to additive constants [61, Corollary 2.7]. Various smoothness properties of the potentials can be established under appropriate regularity conditions on the cost and μ,ν—a fact that we shall leverage in our derivations.

Remark 1 (Literature review on W_p limit distribution theory)**.** Distributional limits of $\sqrt{n}(\mathsf{W}_p^p(\hat{\mu}_n,\nu) - \mathsf{W}_p^p(\mu,\nu))$ and its two-sample analogue for discrete μ,ν under both the null $\mu = \nu$ and the alternative $\mu \neq \nu$ were derived in [187, 190]. Similar results for general distributions are known only in the one-dimensional case. Specifically, for $p = 1, 2$, [59, 60] leverage the representations of W_p in $d = 1$ as the L^p norm between distribution functions ($p = 1$) and quantile functions ($p = 2$) to derive distributional limits under the null. Limit distributions in $d = 1$ for $p \geq 2$ under the alternative ($\mu \neq \nu$) were derived in [63]. In arbitrary dimension, [64] establish asymptotic normality of $\sqrt{n}(\mathsf{W}_2^2(\hat{\mu}_n,\nu) - \mathbb{E}[\mathsf{W}_2^2(\hat{\mu}_n,\nu)])$ under the alternative $\mu \neq \nu$ by deriving an asymptotic linear representation using the Efron-Stein inequality. This was extended to general transportation costs satisfying certain regularity conditions in [61]. The main limitation of these results is the centering around the expected empirical distance (and not the population one), which does not enable performing inference for W_p. This gap was addressed in [137], where a CLT for $\sqrt{n}(\mathsf{W}_2^2(\tilde{\mu}_n,\nu) - \mathsf{W}_2^2(\mu,\nu))$ was established, but for a wavelet-based estimator $\tilde{\mu}_n$ of μ (as opposed to the empirical distribution), while assuming several technical conditions on the Lebesgue densities of μ,ν. As mentioned in the introduction, [117] leverage the extended functional delta method for the supremum functional to obtain limit distributions for W_p, with $p \geq 2$, for compactly supported distributions under the alternative in dimensions $d \leq 3$.

2.2.3 Extended functional delta method

Our unified framework for deriving limit distributions of empirical regularized OT distances relies on the extended functional delta method, which we set up next. Let $\mathfrak{D}, \mathfrak{E}$ be normed spaces and $\phi : \Theta \subset \mathfrak{D} \to \mathfrak{E}$ be a map. Following [181, 169], we say that ϕ is *Hadamard directionally differentiable* at $\theta \in \Theta$ if there exists a map $\phi'_\theta : \mathcal{T}_\Theta(\theta) \to \mathfrak{E}$ such that

$$\lim_{n\to\infty} \frac{\phi(\theta + t_n h_n) - \phi(\theta)}{t_n} = \phi'_\theta(h) \tag{2.4}$$

for any $h \in \mathcal{T}_\Theta(\theta)$, $t_n \downarrow 0$, and $h_n \to h$ in $\mathfrak{D}$ such that $\theta + t_n h_n \in \Theta$. Here $\mathcal{T}_\Theta(\theta)$ is the *tangent cone* to Θ at θ defined as

$$\mathcal{T}_\Theta(\theta) := \left\{ h \in \mathfrak{D} : \ h = \lim_{n\to\infty} \frac{\theta_n - \theta}{t_n} \text{ for some } \theta_n \to \theta \text{ in } \Theta \text{ and } t_n \downarrow 0 \right\}.$$

The tangent cone $\mathcal{T}_\Theta(\theta)$ is closed, and if Θ is convex, then $\mathcal{T}_\Theta(\theta)$ coincides with the closure in $\mathfrak{D}$ of $\{(\vartheta - \theta)/t : \ \vartheta \in \Theta, t > 0\}$. The derivative ϕ'_θ is positively homogeneous and continuous but need not be linear.

Lemma 1 (Extended functional delta method [182, 74, 169, 79])**.** Let $\mathfrak{D}, \mathfrak{E}$ be normed spaces and $\phi : \Theta \subset \mathfrak{D} \to \mathfrak{E}$ be a map that is Hadamard directionally differentiable at $\theta \in \Theta$ with derivative $\phi'_\theta : \mathcal{T}_\Theta(\theta) \to \mathfrak{E}$. Let $T_n : \Omega \to \Theta$ be maps such that $r_n(T_n - \theta) \stackrel{d}{\to} T$ for some $r_n \to \infty$ and Borel measurable map $T : \Omega \to \mathfrak{D}$ with values in $\mathcal{T}_\Theta(\theta)$. Then, $r_n(\phi(T_n) - \phi(\theta)) \stackrel{d}{\to} \phi'_\theta(T)$. Further, if Θ is convex, then we have $r_n(\phi(T_n) - \phi(\theta)) - \phi'_\theta(r_n(T_n - \theta)) \to 0$ in outer probability.

Lemma 1 is at the core of our framework for deriving limit distributions. It is termed the "extended" functional delta method as it extends the (classical) functional delta method for Hadamard differentiable maps to directionally differentiable ones.

While Hadamard directional differentiability is sufficient to derive limit distributions, bootstrap consistency often requires (full) Hadamard differentiability. Recall that the map ϕ is *Hadamard differentiable* at θ *tangentially to* a vector subspace $\mathfrak{D}_0 \subset \mathfrak{D}$ if there exists a continuous *linear* map $\phi'_\theta : \mathfrak{D}_0 \to \mathfrak{E}$ satisfying (2.4) for any $h \in \mathfrak{D}_0$, $t_n \to 0$ ($t_n \neq 0$), and $h_n \to h$ in $\mathfrak{D}$ such that $\theta + t_n h_n \in \Theta$. The differences from Hadamard directional differentiability is that the derivative ϕ'_θ must be linear and thus the domain must be a vector subspace of $\mathfrak{D}$, and the sequence $t_n \to 0$ must be a generic (nonzero) sequence converging to zero. The next lemma is useful for verifying Hadamard differentiability from the directional one.

Lemma 2. Let $\phi : \Theta \subset \mathfrak{D} \to \mathfrak{E}$ be Hadamard directionally differentiable at $\theta \in \Theta$ with derivative $\phi'_\theta : \mathcal{T}_\Theta(\theta) \to \mathfrak{E}$. If $\mathcal{T}_\Theta(\theta)$ contains a subspace $\mathfrak{D}_0$ on which ϕ'_θ is linear, then ϕ is Hadamard differentiable at θ tangentially to $\mathfrak{D}_0$.

2.3 Unified Framework for Statistical Inference

This section develops a general framework for deriving limit distributions, bootstrap consistency, and semiparametric efficiency bounds for regularized OT distances. We first treat the former two aspects together, and then move on to discuss efficiency. Throughout this section, μ_n designates an arbitrary random probability measure and not necessarily the empirical measure (unless explicitly stated otherwise).

2.3.1 Limit distributions and bootstrap consistency

The following result is an adaptation of the extended functional delta method from Lemma 1 to the space of probability measures, which enables directly applying it to empirical regularized OT.

Proposition 1 (Limit distributions)**.** Consider the setting:

> **(Setting $\circledast$)** Let $\mathcal{F}$ be a class of Borel measurable functions on a topological space S with a finite envelope F. For a given $\mu \in \mathcal{P}(S)$, let δ be a map from $\mathcal{P}_0 \subset \mathcal{P}(S)$ into a Banach space $(\mathfrak{E}, \|\cdot\|_{\mathfrak{E}})$, where $\mathcal{P}_0$ is a convex subset such that $\mu \in \mathcal{P}_0$ and $\int F d\nu < \infty$ for all $\nu \in \mathcal{P}_0$.

Further suppose that

(a) $\mu_n : \Omega \to \mathcal{P}_0$ are random probability measures with values in $\mathcal{P}_0$ for all $n \in \mathbb{N}$, such that there exists a tight random variable G_μ in $\ell^\infty(\mathcal{F})$ with $\sqrt{n}(\mu_n - \mu) \stackrel{d}{\to} G_\mu$ in $\ell^\infty(\mathcal{F})$;

(b) δ is locally Lipschitz continuous at μ with respect to $\|\cdot\|_{\infty,\mathcal{F}}$, in the sense that there exist constants $\epsilon > 0$ and $C < \infty$ such that

$$\|\nu - \mu\|_{\infty,\mathcal{F}} \vee \|\nu' - \mu\|_{\infty,\mathcal{F}} < \epsilon \quad \Longrightarrow \quad \|\delta(\nu) - \delta(\nu')\|_{\mathfrak{E}} \leq C\|\nu - \nu'\|_{\infty,\mathcal{F}};$$

(c) For every $\nu \in \mathcal{P}_0$, the mapping $t \mapsto \delta(\mu + t(\nu - \mu))$ is right differentiable at $t = 0$, and denote its right derivative by

$$\delta'_\mu(\nu - \mu) = \lim_{t \downarrow 0} \frac{\delta(\mu + t(\nu - \mu)) - \delta(\mu)}{t}. \tag{2.5}$$

Then (i) δ'_μ uniquely extends to a continuous, positively homogeneous map on the tangent cone of $\mathcal{P}_0$ at μ:

$$\mathcal{T}_{\mathcal{P}_0}(\mu) := \overline{\{t(\nu - \mu) : \nu \in \mathcal{P}_0, t > 0\}}^{\ell^\infty(\mathcal{F})};$$

(ii) $G_\mu \in T_{\mathcal{P}_0}(\mu)$ almost surely (a.s.); and (iii) $\sqrt{n}(\delta(\mu_n)-\delta(\mu))-\delta'_\mu(\sqrt{n}(\mu_n-\mu)) \to 0$ holds in outer probability. Consequently, we have the following convergence in distribution $\sqrt{n}(\delta(\mu_n) - \delta(\mu)) \xrightarrow{d} \delta'_\mu(G_\mu)$.

The proof first identifies δ as a map defined on a subset of $\ell^\infty(\mathcal{F})$. Formally, let $\tau : \mathcal{P}_0 \ni \nu \mapsto (f \mapsto \nu(f)) \in \ell^\infty(\mathcal{F})$, and we identify δ with $\bar{\delta} : \tau\mathcal{P}_{0,\epsilon} \to \mathfrak{E}$ defined by $\bar{\delta}(\tau\nu) = \delta(\nu)$, where $\mathcal{P}_{0,\epsilon} = \{\nu \in \mathcal{P}_0 : \|\nu - \mu\|_{\infty,\mathcal{F}} < \epsilon\}$. The local Lipschitz condition (b) guarantees that the map $\bar{\delta}$ is well-defined (indeed, without the local Lipschitz condition, $\bar{\delta}$ may not be well-defined as τ may fail to be one-to-one). With this identification, we apply the extended functional delta method, Lemma 1, by establishing Hadamard directional differentiability of δ at μ. The latter essentially follows by local Lipschitz continuity (condition (b)) and Gâteaux directional differentiability (condition (c)). Since the derivative δ'_μ is a priori defined only on $\mathcal{P}_{0,\epsilon} - \mu$, we need to extend the derivative to the tangent cone $\mathcal{T}_{\mathcal{P}_0}(\mu)$, for which we need completeness of the space $\mathfrak{E}$; see the proof in Section A.1 of the supplementary material for details.

For i.i.d. data $X_1, \ldots, X_n \sim \mu$ and $\mu_n = \hat{\mu}_n$ as the empirical measure, to apply Proposition 1 we will: (i) find a μ-Donsker function class $\mathcal{F}$ such that the functional δ is locally Lipschitz w.r.t. $\|\cdot\|_{\infty,\mathcal{F}}$ at μ; and (ii) find the Gâteaux directional derivative (2.5). In our applications to regularized OT, such a function class $\mathcal{F}$ will be chosen to contain dual potentials corresponding to a proper class of distributions. Regularization enforces dual potentials to possess certain smoothness or low-dimensionality properties, guaranteeing that $\mathcal{F}$ is indeed μ-Donsker. The dual OT formulation also plays a crucial role in finding the Gâteaux directional derivative (2.5).

Remark 2 (On Proposition 1)**.** We now clarify certain aspects of Proposition 1.

(Relaxed condition): When $\delta(\mu_n)$ is well-defined, the condition that μ_n takes values in

$\mathcal{P}_0$ can be relaxed to $\mu_n \in \mathcal{P}_0$ with inner probability approaching one.

(Data generating process): Proposition 1 does not impose any dependence conditions on the data. In particular, it can be applied to dependent data as long as one can verify the uniform limit theorem in Condition (a). See, e.g., [131, 7, 9, 71, 11, 154, 56] on uniform CLTs for dependent data.

(Convexity of $\mathcal{P}_0$): The assumption that $\mathcal{P}_0$ is convex can be replaced with the condition that $\mathcal{P}_0$ is a convex subset of $\ell^\infty(\mathcal{F})$. Namely, using the mapping $\tau : \mathcal{P}_0 \ni \nu \mapsto (f \mapsto \nu(f)) \in \ell^\infty(\mathcal{F})$, we only need that $\tau\mathcal{P}_0 = \{\tau\nu : \nu \in \mathcal{P}_0\} \subset \ell^\infty(\mathcal{F})$ is convex. Condition (c) then should read that $t \mapsto \bar{\delta}((1-t)\tau\mu + t\tau\nu)$ is differentiable from the right at $t = 0$ with derivative $\delta'_\mu(\mu - \nu) = \lim_{t\downarrow 0} t^{-1}\{\bar{\delta}((1-t)\tau\mu + t\tau\nu) - \bar{\delta}(\tau\mu)\}$, where $\bar{\delta}(\tau\nu) = \delta(\nu)$. This modification is needed to cover the two-sample setting; see, e.g., the proof of Theorem 1 Part (ii).

Bootstrap consistency

In applications of Proposition 1, the obtained limit distribution is often non-pivotal in the sense that it depends on the population distribution μ, which is unknown in practice. To circumvent the difficulty of estimating the distribution of $\delta'_\mu(G_\mu)$ directly, one may apply the bootstrap. When $\mathcal{F}$ is μ-Donsker and $\mu_n = \hat{\mu}_n$ is the empirical distribution of i.i.d. data from μ, then the bootstrap (applied to the functional δ) is consistent for estimating the distribution of $\delta'_\mu(G_\mu)$ provided that the map $\nu \mapsto \delta(\nu)$ is Hadamard differentiable w.r.t. $\|\cdot\|_{\infty,\mathcal{F}}$ at $\nu = \mu$ tangentially to a subspace of $\ell^\infty(\mathcal{F})$ that contains the support of G_μ; cf. Theorem 23.9 in [200] or Theorem 3.9.11 in [201]. The following corollary is useful for invoking such theorems under the setting of Proposition 1.

Corollary 1 (Bootstrap consistency via Hadamard differentiability)**.** Consider the setting of Proposition 1. If, in addition, G_μ is a mean-zero Gaussian variable in $\ell^\infty(\mathcal{F})$, then $\mathrm{spt}(G_\mu)$ is a vector subspace of $\ell^\infty(\mathcal{F})$. If further δ'_μ is linear on $\mathrm{spt}(G_\mu)$, then $\nu \mapsto \delta(\nu)$ is Hadamard differentiable w.r.t. $\|\cdot\|_{\infty,\mathcal{F}}$ at $\nu = \mu$ tangentially to $\mathrm{spt}(G_\mu)$.

In general, when the functional is Hadamard directionally differentiable with a nonlinear derivative, the bootstrap fails to be consistent; cf. [74, 79]. An alternative way to estimate the limit distribution in such cases is to use subsampling or the "m-out-of-n" bootstrap [161, 74]; see Lemma 3 for the max-slicing case.

2.3.2 Semiparametric efficiency

In Proposition 1, if δ'_μ is linear and G_μ is mean-zero Gaussian, then the limit distribution $\delta'_\mu(G_\mu)$ is mean-zero Gaussian as well. In such cases, it is natural to ask if the plug-in estimator $\delta(\mu_n)$ is asymptotically efficient in the sense of [200, p. 367], relative to a certain tangent space. Informally, the semiparametric efficiency bound at μ is computed as the largest Cramér-Rao lower bound among one-dimensional submodels passing through μ.

Formally, consider estimating a functional $\kappa : \mathcal{P} \subset \mathcal{P}(S) \to \mathbb{R}$ at $\mu \in \mathcal{P}$ from i.i.d. data $X_1, \dots, X_n \sim \mu$. We consider submodels $\{\mu_t : 0 \le t < \epsilon'\}$ with $\mu_0 = \mu$ such that, for some measurable score function $h : S \to \mathbb{R}$, we have

$$\int \left[\frac{d\mu_t^{1/2} - d\mu^{1/2}}{t} - \frac{1}{2} h d\mu^{1/2} \right]^2 \to 0,$$

where $d\mu_t$ and $d\mu$ are Radon-Nikodym densities w.r.t. a common dominating measure and the integration is taken w.r.t. the dominating measure. Score functions are square integrable w.r.t. μ and μ-mean zero. A *tangent set* $\dot{\mathcal{P}}_\mu \subset L^2(\mu)$ of the model $\mathcal{P}$ at μ is the set of score functions corresponding to a collection of such submodels. If $\dot{\mathcal{P}}_\mu$ is a vector subspace of $L^2(\mu)$, then it is called a *tangent space*. Relative to a given tangent set $\dot{\mathcal{P}}_\mu$, the functional $\kappa : \mathcal{P} \to \mathbb{R}$ is called *differentiable* at μ if there exists a continuous linear functional $\dot{\kappa}_\mu : L^2(\mu) \to \mathbb{R}$ such that, for every $h \in \dot{\mathcal{P}}_\mu$ and a submodel $t \mapsto \mu_t$ with score function h,

$$\frac{\kappa(\mu_t) - \kappa(\mu)}{t} \to \dot{\kappa}_\mu h, \quad t \downarrow 0.$$

The semiparametric efficiency bound for estimating κ at μ, relative to $\dot{\mathcal{P}}_\mu$, is defined as

$$\sigma^2_{\kappa,\mu} = \sup_{h \in \mathrm{lin}(\dot{\mathcal{P}}_\mu)} \frac{(\dot{\kappa}_\mu h)^2}{\|h\|^2_{L^2(\mu)}},$$

where $\mathrm{lin}(\dot{\mathcal{P}}_\mu)$ is the linear span of $\dot{\mathcal{P}}_\mu$. In particular, the $N(0, \sigma^2_{\kappa,\mu})$ distribution serves as the "optimal" limit distribution for estimating κ at μ in the sense of the Hájek-Le Cam convolution theorem and also in the local asymptotic minimax sense; see Chapter 25 in [200] for details.

The next proposition concerns the computation of the semiparametric efficiency bound.

Proposition 2 (Semiparametric efficiency)**.** For Setting ⊛ from Proposition 1 with $\mathfrak{E} = \mathbb{R}$, consider estimating $\delta : \mathcal{P}_0 \to \mathbb{R}$ at μ from i.i.d. data $X_1, \dots, X_n \sim \mu$. Set

$$\dot{\mathcal{P}}_{0,\mu} = \{h : \ h : S \to \mathbb{R} \text{ is bounded and measurable with } \mu\text{-mean zero}\}.$$

Suppose that (a) the function class $\mathcal{F}$ is μ-pre-Gaussian, i.e., there exists a tight mean-zero Gaussian process $G_\mu = (G_\mu(f))_{f \in \mathcal{F}}$ in $\ell^\infty(\mathcal{F})$ with covariance function $\mathrm{Cov}(G_\mu(f), G_\mu(g)) = \mathrm{Cov}_\mu(f, g)$; (b) for every $h \in \dot{\mathcal{P}}_{0,\mu}$, $(1 + th)\mu \in \mathcal{P}_0$ for sufficiently small $t > 0$; and (c) there exists a continuous linear functional $\delta'_\mu : \ell^\infty(\mathcal{F}) \to \mathbb{R}$ such that (2.5) holds for every $\nu \in \mathcal{P}_0$ of the form $\nu = (1 + h)\mu$ for some $h \in \dot{\mathcal{P}}_{0,\mu}$. Then, the semiparametric efficiency bound for estimating δ at μ relative to the tangent space $\dot{\mathcal{P}}_{0,\mu}$ agrees with $\mathrm{Var}(\delta'_\mu(G_\mu))$.

Proposition 2 can be thought of as a variant of Theorem 3.1 in [198], which asserts that a Hadamard differentiable functional (tangentially to a sufficiently large subspace) of an asymptotically efficient estimator is again asymptotically efficient; see Remark 25 in the supplementary material for more details. In Section A.1 of the supplementary material, we provide a direct and self-contained proof of Proposition 2. We note that Proposition 2 covers a slightly more general situation than [198, Theorem 3.1] since it only requires Gâteaux differentiability of the map δ, and choosing a pre-Gaussian function class $\mathcal{F}$ in such a way that the derivative δ'_μ extends to a continuous linear functional on $\ell^\infty(\mathcal{F})$. In particular, the efficiency bound computation in Proposition 2 is applicable even when Proposition 1 is difficult to apply. For instance, when the Gâteaux derivative δ'_μ in (2.5) is a point evaluation, $\delta'_\mu(\nu - \mu) = (\nu - \mu)(f^\star)$ for some function $f^\star \in L^2(\mu)$, we can choose $\mathcal{F} = \{f^\star\}$ (singleton) and apply Proposition 2 to conclude that $\mathrm{Var}_\mu(f^\star)$ agrees with the semiparametric efficiency bound, relative to $\dot{\mathcal{P}}_{0,\mu}$ (note that the function class $\mathcal{F}$ in Proposition 2 need not be the same as the one in Proposition 1).

The following corollary covers the two-sample case. Define $\dot{\mathcal{P}}_{0,\nu}$ analogously to $\dot{\mathcal{P}}_{0,\mu}$ and set $\dot{\mathcal{P}}_{0,\mu} \oplus \dot{\mathcal{P}}_{0,\nu} = \{h_1 \oplus h_2 : h_1 \in \dot{\mathcal{P}}_{0,\mu}, h_2 \in \dot{\mathcal{P}}_{0,\nu}\}$.

Corollary 2 (Semiparametric efficiency in two-sample setting)**.** Let $\mathcal{F}$ be a class of Borel measurable functions on a topological space S with finite envelope F, and for given $\mu, \nu \in \mathcal{P}(S)$, let $\mathcal{P}_{0,\mu}, \mathcal{P}_{0,\nu}$ be subsets of $\mathcal{P}(S)$ containing μ, ν, respectively, such that $\int F d\rho < \infty$ for all $\rho \in \mathcal{P}_{0,\mu} \cup \mathcal{P}_{0,\nu}$. Let $\mathcal{P}_{0,\mu} \otimes \mathcal{P}_{0,\nu} = \{\rho_1 \otimes \rho_2 : \rho_1 \in \mathcal{P}_{0,\mu}, \rho_2 \in \mathcal{P}_{0,\nu}\}$. Consider estimating $\delta : \mathcal{P}_{0,\mu} \otimes \mathcal{P}_{0,\nu} \to \mathbb{R}$ at $\mu \otimes \nu$ from i.i.d. data $(X_1, Y_1), \dots, (X_n, Y_n) \sim \mu \otimes \nu$. Suppose that (a) the function class $\mathcal{F}$ is pre-Gaussian w.r.t. μ and ν; (b) for every $h_1 \oplus h_2 \in \dot{\mathcal{P}}_{0,\mu} \oplus \dot{\mathcal{P}}_{0,\nu}$, $((1 + th_1)\mu) \otimes ((1 + th_2)\nu) \in \mathcal{P}_{0,\mu} \otimes \mathcal{P}_{0,\nu}$ for sufficiently small $t > 0$; (c) there exist continuous linear functionals $\delta'_\mu : \ell^\infty(\mathcal{F}) \to \mathbb{R}$ and $\delta'_\nu : \ell^\infty(\mathcal{F}) \to \mathbb{R}$ such that $t^{-1}\{\delta(((1 + th_1)\mu) \otimes ((1 + th_2)\nu)) - \delta(\mu \otimes \nu)\} \to \delta'_\mu(h_1\mu) + \delta'_\nu(h_2\nu)$ as $t \downarrow 0$ for every $h_1 \oplus h_2 \in \dot{\mathcal{P}}_{0,\mu} \oplus \dot{\mathcal{P}}_{0,\nu}$. Then, the semiparametric efficiency bound for estimating δ at $\mu \otimes \nu$

relative to the tangent space $\dot{\mathcal{P}}_{0,\mu} \oplus \dot{\mathcal{P}}_{0,\nu}$ agrees with $\mathrm{Var}(\delta'_\mu(G_\mu)) + \mathrm{Var}(\delta'_\mu(G_\nu))$, where G_μ and G_ν are tight μ- and ν-Brownian bridges in $\ell^\infty(\mathcal{F})$, respectively.

Remark 3 (Efficiency of wavelet-based estimator of W_2)**.** Theorem 18 in [137] establishes a CLT for a wavelet-based estimator $\mathsf{W}_2(\tilde{\mu}_n, \nu)$ for $\mathsf{W}_2(\mu, \nu)$ in the one-sample case under high-level assumptions that include global regularity of the OT potential φ. Their result reads as $\sqrt{n}(\mathsf{W}_2^2(\tilde{\mu}_n, \nu) - \mathsf{W}_2^2(\mu, \nu)) \xrightarrow{d} N(0, \mathrm{Var}_\mu(\varphi))$. Their proof first establishes a CLT for the expectation centering (similarly to [64]) and then shows that the bias is negligible. To obtain the same result via Proposition 1, one would have to assume a *uniform* bound on the Hölder norm of OT potentials corresponding to a local neighborhood of μ. Unfortunately, such uniform bounds on global regularity of OT potentials are currently unavailable, except for a few limited cases (cf. the discussion after Theorem 3 in [137]). For instance, when the marginals are defined on the flat torus and admit sufficiently smooth densities that are bounded away from 0 and ∞, Theorem 5 in [137] provides such a bound, rendering Proposition 1 applicable. Moreover, in the setting of Theorem 18 in [137], it is readily verified that $\rho \mapsto \mathsf{W}_2(\rho, \nu)$ is Gâteaux differentiable at μ with derivative $(\rho - \mu)(\varphi)$ (cf. the proof of Lemma 45 in the supplementary material), so $\mathrm{Var}_\mu(\varphi)$ indeed coincides with the semiparametric efficiency bound.

2.4 Sliced Wasserstein distances

This section studies statistical aspects of sliced Wasserstein distances, deriving limit distributions, bootstrap consistency, and semiparametric efficiency.

2.4.1 Background

Average- and max-sliced Wasserstein distances are defined next.

Definition 2 (Sliced Wasserstein distances)**.** Let $1 \leq p < \infty$. The average-sliced and max-sliced p-Wasserstein distances between $\mu, \nu \in \mathcal{P}_p(\mathbb{R}^d)$ are defined, respectively, as

$$\underline{\mathsf{W}}_p(\mu,\nu) := \left[\int_{\mathbb{S}^{d-1}} \mathsf{W}_p^p(\mathfrak{p}^\theta_\sharp\mu, \mathfrak{p}^\theta_\sharp\nu) d\sigma(\theta)\right]^{1/p} \quad \text{and} \quad \overline{\mathsf{W}}_p(\mu,\nu) := \max_{\theta\in\mathbb{S}^{d-1}} \mathsf{W}_p(\mathfrak{p}^\theta_\sharp\mu, \mathfrak{p}^\theta_\sharp\nu),$$

where $\mathfrak{p}^\theta : \mathbb{R}^d \to \mathbb{R}$ is the projection map $x \mapsto \theta^\intercal x$, σ is the uniform distribution on the unit sphere $\mathbb{S}^{d-1} := \{x \in \mathbb{R}^d : \|x\| = 1\}$, and $\mathfrak{p}^\theta_\sharp\mu := \mu \circ (\mathfrak{p}^\theta)^{-1}$ is the pushforward of μ under $\mathfrak{p}^\theta$.

The sliced distances $\underline{\mathsf{W}}_p$ and $\overline{\mathsf{W}}_p$ are metrics on $\mathcal{P}_p(\mathbb{R}^d)$ and, in fact, induce the same topology as W_p [12]. Sliced Wasserstein distances are efficiently computable using the closed-form expression for W_p between distributions on $\mathbb{R}$ using quantile functions. For $\mu \in \mathcal{P}(\mathbb{R}^d)$ and $\theta \in \mathbb{S}^{d-1}$, denote by $F_\mu(\cdot\,;\theta)$ and $F_\mu^{-1}(\cdot\,;\theta)$ the distribution and quantile functions of $\mathfrak{p}^\theta_\sharp\mu$, respectively, i.e.,

$$F_\mu(t;\theta) = \mu(\{x \in \mathbb{R}^d : \theta^\intercal x \leq t\}) \quad \text{and} \quad F_\mu^{-1}(\tau;\theta) = \inf\{t \in \mathbb{R} : F_\mu(t;\theta) \geq \tau\}.$$

Then, $\mathsf{W}_p(\mathfrak{p}^\theta_\sharp\mu, \mathfrak{p}^\theta_\sharp\nu)$ equals the L^p-norm between the corresponding quantile functions,

$$\mathsf{W}_p^p(\mathfrak{p}^\theta_\sharp\mu, \mathfrak{p}^\theta_\sharp\nu) = \int_0^1 \left|F_\mu^{-1}(\tau;\theta) - F_\nu^{-1}(\tau;\theta)\right|^p d\tau,$$

which further simplifies for $p = 1$ to the L^1 distance between the corresponding distribution functions. Also, sliced Wasserstein distances between projected empirical distributions is readily computed using order statistics. Let $\hat{\mu}_n := n^{-1}\sum_{i=1}\delta_{X_i}$ and $\hat{\nu}_n := n^{-1}\sum_{i=1}\delta_{Y_i}$ be the empirical distributions of $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$. For each $\theta \in \mathbb{S}^{d-1}$, let $X_i(\theta) = \theta^\intercal X_i$, and let $X_{(1)}(\theta) \leq \cdots \leq X_{(n)}(\theta)$ be the order statistics. Define $Y_{(1)}(\theta) \leq \cdots \leq Y_{(n)}(\theta)$ analogously. By Lemma 4.2 in [25], we have $\mathsf{W}_p^p(\mathfrak{p}^\theta_\sharp\hat{\mu}_n, \mathfrak{p}^\theta_\sharp\hat{\nu}_n) = \frac{1}{n}\sum_{i=1}^n \left|X_{(i)}(\theta) - Y_{(i)}(\theta)\right|^p$. The sliced distances $\underline{\mathsf{W}}_p$ and $\overline{\mathsf{W}}_p$ can be computed by integrating or maximizing the above over $\theta \in \mathbb{S}^{d-1}$.

Literature review

Sliced Wasserstein distances have been applied to various statistical inference and machine learning tasks, including barycenter computation [163], generative modeling [69, 68, 148, 147], and autoencoders [126]. The statistical literature on sliced distances mostly focused on expected value analysis. Specifically, [153] show that if μ satisfies a $T_q(\sigma^2)$ inequality with $q \in [1, 2]$, then $\mathbb{E}[\overline{\mathsf{W}}_p(\hat{\mu}_n, \mu)] \lesssim \sigma(n^{-1/(2p)} + n^{(1/q-1/p)_+}\sqrt{(d\log n)/n})$ up to a constant that depends only on p. Further results on empirical convergence rates can be found in [134], where both $\underline{\mathsf{W}}_p$ and $\overline{\mathsf{W}}_p$ were treated, while replacing the transport inequality assumption of [153] with exponential moment bounds (via Bernstein's tail conditions) or Poincaré type inequalities. A limit distribution result for one-sample sliced W_1 was mentioned in [147] but was left as an unproven assumption. Extensions to sliced W_p and two-sample results, all of which are crucial for principled statistical inference, are currently open. Consistency of the bootstrap and efficiency bounds are also unaccounted for by the existing literature.[2]

2.4.2 Statistical analysis

We move on to the statistical aspects of sliced W_p, closing the aforementioned gaps. The $p > 1$ case is treated under the general framework of Section 2.3 for compactly supported distributions. For $p = 1$, we present a separate derivation that leverages its simplified form to obtain the results under mild moment assumptions.

[2]After the first version of the present paper was posted on the arXiv, we became aware that the latest update of [138] (arXiv update: April 4, 2022) contains limit distribution and bootstrap results for $\underline{\mathsf{W}}_p$ with $p > 1$ under the alternative. Our work is independent of [138] and our approach is distinct; see Remark 5. After our paper was posted, [218] proved similar results to ours for $\underline{\mathsf{W}}_1$ and $\overline{\mathsf{W}}_1$ under a similar set of assumptions to us (see Remark 8 for details) and [214] derived limit distributions for $p > 1$ uniformly in the slicing direction under the assumption of compact support and uniqueness of optimal potentials, but did not cover the max-slicing case in full generality. Neither of these works addressed asymptotic efficiency.

Order $p > 1$

The next theorem characterizes limit distributions for average-sliced p-Wasserstein distances under both the one- and two-sample settings. It also states asymptotic efficiency of the empirical plug-in estimator, and consistency of the bootstrap. The latter facilitates statistical inference by providing a tractable estimate of the limiting distribution, and is set up as follows. Given the data $X_1, \dots, X_n$, let $X_1^B, \dots, X_n^B$ be an independent sample from $\hat{\mu}_n$, and set $\hat{\mu}_n^B := n^{-1} \sum_{i=1}^n \delta_{X_i^B}$ as the bootstrap empirical distribution. Define $\hat{\nu}_n^B$ analogously and let $\mathbb{P}^B$ denote the conditional probability given the data.

Theorem 1 (Limit distribution, efficiency, and bootstrap consistency for $\underline{\mathsf{W}}_p^p$)**.** Let $1 < p < \infty$, and suppose that μ, ν are compactly supported, such that μ is absolutely continuous and $\mathrm{spt}(\mu)$ is convex. For every $\theta \in \mathbb{S}^{d-1}$, let φ^θ be an OT potential from $\mathfrak{p}_\sharp^\theta \mu$ to $\mathfrak{p}_\sharp^\theta \nu$ for W_p, which is unique up to additive constants on $\mathrm{int}(\mathrm{spt}(\mathfrak{p}_\sharp^\theta \mu))$. Also, set $\psi^\theta = [\varphi^\theta]^c$ as the c-transform of φ^θ for $c(s,t) = |s-t|^p$. The following hold.

(i) We have

$$\sqrt{n}(\underline{\mathsf{W}}_p^p(\hat{\mu}_n, \nu) - \underline{\mathsf{W}}_p^p(\mu, \nu)) \xrightarrow{d} N\left(0, v_p^2\right),$$

where $v_p^2 = \iint \mathrm{Cov}_\mu(\varphi^\theta \circ \mathfrak{p}^\theta, \varphi^\vartheta \circ \mathfrak{p}^\vartheta) d\sigma(\theta) d\sigma(\vartheta)$, which is well-defined under the current assumption. The asymptotic variance v_p^2 coincides with the semiparametric efficiency bound for estimating $\underline{\mathsf{W}}_p^p(\cdot, \nu)$ at μ. Also, provided that $v_p^2 > 0$, we have

$$\sup_{t \in \mathbb{R}} \left| \mathbb{P}^B\Big(\sqrt{n}(\underline{\mathsf{W}}_p^p(\hat{\mu}_n^B, \nu) - \underline{\mathsf{W}}_p^p(\hat{\mu}_n, \nu)) \le t \Big) - \mathbb{P}(N(0, v_p^2) \le t) \right| \xrightarrow{\mathbb{P}} 0.$$

(ii) If in addition ν is absolutely continuous with convex support, then

$$\sqrt{n}(\underline{\mathsf{W}}_p^p(\hat{\mu}_n, \hat{\nu}_n) - \underline{\mathsf{W}}_p^p(\mu, \nu)) \xrightarrow{d} N\left(0, v_p^2 + w_p^2\right),$$

where v_p^2 is given in (i) and $w_p^2 = \iint \mathrm{Cov}_\nu(\psi^\theta \circ \mathfrak{p}^\theta, \psi^\vartheta \circ \mathfrak{p}^\vartheta) d\sigma(\theta) d\sigma(\vartheta)$. The asymptotic variance $v_p^2 + w_p^2$ coincides with the semiparametric efficiency bound for estimating

$\underline{\mathsf{W}}_p^p(\cdot,\cdot)$ at (μ,ν). Also, provided that $v_p^2 + w_p^2 > 0$, we have

$$\sup_{t\in\mathbb{R}} \left|\mathbb{P}^B\Big(\sqrt{n}(\underline{\mathsf{W}}_p^p(\hat{\mu}_n^B,\hat{\nu}_n^B) - \underline{\mathsf{W}}_p^p(\hat{\mu}_n,\hat{\nu}_n)) \le t\Big) - \mathbb{P}(N(0, v_p^2 + w_p^2) \le t)\right| \xrightarrow{\mathbb{P}} 0.$$

The derivation of the limit distributions in Theorem 1 follows from Proposition 1. We outline the main idea for the one-sample case. The functional of interest is set as the pth power of the average-sliced p-Wasserstein distance. Leveraging compactness of supports, we then show that $\underline{\mathsf{W}}_p^p$ is Lipschitz w.r.t. $\overline{\mathsf{W}}_1$ (cf. Lemma 9 in the supplementary material). From the Kantorovich-Rubinstein duality, $\overline{\mathsf{W}}_1$ can be expressed as $\overline{\mathsf{W}}_1(\mu,\nu) = \|\mu-\nu\|_{\infty,\mathcal{F}}$ with $\mathcal{F} = \{\varphi\circ\mathfrak{p}^\theta : \theta\in\mathbb{S}^{d-1}, \varphi\in\mathsf{Lip}_{1,0}(\mathbb{R})\}$, which is shown to be μ-Donsker ($\mathsf{Lip}_{1,0}(\mathbb{R})$ denotes the class of 1-Lipschitz functions φ on $\mathbb{R}$ with $\varphi(0)=0$). Evaluating the Gâteaux directional derivative of the sliced distance, we have all the conditions needed to invoke Proposition 1, which in turn yields the distributional limits. For $\underline{\mathsf{W}}_p$, the corresponding derivative turns out to be linear (in a suitable sense), so that asymptotic efficiency of the plug-in estimator and the bootstrap consistency follow from Proposition 2 and Corollary 1 combined with Theorem 23.9 in [200].

For the two-sample case, we think of $\underline{\mathsf{W}}_p(\mu,\nu)$ as a functional of the product measure $\mu\otimes\nu$, as the correspondence between (μ,ν) and $\mu\otimes\nu$ is one-to-one. With this identification, the rest of the argument is analogous to the one-sample case. We note that in the two-sample case, the semiparametric efficiency bound is defined relative to the tangent space

$$\{h_1\oplus h_2 : h_1 \text{ and } h_2 \text{ are bounded measurable functions with } \mu(h_1)=\nu(h_2)=0\}.$$

This convention is adopted throughout when discussing semiparametric efficiency bounds in the two-sample case.

The asymptotic variances in Theorem 1 involve potentials between all slices of the marginal distributions, so direct estimation of the asymptotic variances seems highly nontrivial from a computational standpoint. Hence, the bootstrap offers a particularly appealing alternative for estimating the sampling distributions of empirical sliced Wasserstein distances.

Remark 4 (Removing pth power)**.** While Theorem 1 states limit distributions for the pth power of $\underline{\mathsf{W}}_p$, we can readily obtain corresponding results for the average-sliced p-Wasserstein distance itself by invoking the delta method for the map $s \mapsto s^{1/p}$.

Remark 5 (Comparison with [138])**.** Theorem 4 in [138] derives limit distributions and bootstrap consistency for empirical $\underline{\mathsf{W}}_p$ under the alternative, subject to the assumption that the projected densities $\{f_\mu(\cdot;\theta)\}_{\theta\in\mathbb{S}^{d-1}}$ and $\{f_\nu(\cdot;\theta)\}_{\theta\in\mathbb{S}^{d-1}}$ are uniformly integrable with

$$\sup_{\theta\in\mathbb{S}^{d-1}} \operatorname{esssup}_{0<u<1} \frac{1}{f_\mu(F_\nu^{-1}(t;\theta);\theta)} \vee \frac{1}{f_\nu(F_\nu^{-1}(t;\theta);\theta)} < \infty.$$

Here $f_\mu(\cdot;\theta)$ and $F_\mu^{-1}(\cdot;\theta)$ are the (Lebesgue) density and quantile function of $\mathfrak{p}^\theta_\sharp\mu$, and their composition is the so-called I-function; cf. [25, Equation (5.2)]. Verification of this condition for given distributions seems nontrivial. The proof of [138, Theorem 4] exploits the quantile function representation of W_p in $d = 1$ along with a linearization step (of quantile functions). Our limit theorem, on the other hand, assumes that μ has a density with compact and convex support, and employs a markedly different proof via the general framework of Proposition 1.

We next provide one- and two-sample limit distributions for the max-sliced Wasserstein distance. In this case, the Hadamard directional derivative is nonlinear and therefore the limit is non-Gaussian and the nonparametric bootstrap is inconsistent (cf. [74, 79]).

Theorem 2 (Limit distribution for $\overline{\mathsf{W}}_p$)**.** Consider the assumption of Theorem 1.

(i) Setting $\mathfrak{S}_{\mu,\nu} := \{\theta \in \mathbb{S}^{d-1} : \mathsf{W}_p(\mathfrak{p}^\theta_\sharp\mu, \mathfrak{p}^\theta_\sharp\nu) = \overline{\mathsf{W}}_p(\mu,\nu)\}$, we have

$$\sqrt{n}(\overline{\mathsf{W}}_p^p(\hat\mu_n,\nu) - \overline{\mathsf{W}}_p^p(\mu,\nu)) \stackrel{d}{\to} \sup_{\theta\in\mathfrak{S}_{\mu,\nu}} \mathbb{G}_\mu(\theta),$$

where $(\mathbb{G}_\mu(\theta))_{\theta\in\mathbb{S}^{d-1}}$ is a centered Gaussian process with continuous paths and covariance function $\mathrm{Cov}(\mathbb{G}_\mu(\theta), \mathbb{G}_\mu(\vartheta)) = \mathrm{Cov}_\mu(\varphi^\theta \circ \mathfrak{p}^\theta, \varphi^\vartheta \circ \mathfrak{p}^\vartheta)$, which is well-defined.

(ii) If in addition ν is also absolutely continuous with convex support, then

$$\sqrt{n}(\overline{\mathsf{W}}_p^p(\hat{\mu}_n, \hat{\nu}_n) - \overline{\mathsf{W}}_p^p(\mu, \nu)) \stackrel{d}{\to} \sup_{\theta\in\mathfrak{S}_{\mu,\nu}} [\mathbb{G}_\mu(\theta) + \mathbb{G}'_\nu(\theta)], \tag{2.6}$$

where $(\mathbb{G}'_\nu(\theta))_{\theta\in\mathbb{S}^{d-1}}$ is independent of $\mathbb{G}_\mu$ given in (i) and defined analogously.

Observe that, for $\mu, \nu \in \mathcal{P}_p(\mathbb{R}^d)$, the map $\theta \mapsto \mathsf{W}_p(\mathfrak{p}^\theta_\sharp\mu, \mathfrak{p}^\theta_\sharp\nu)$ is continuous, so the set $\mathfrak{S}_{\mu,\nu}$ is nonempty. The proof of Theorem 2 also relies on the general framework of Proposition 1. As in the average case, $\overline{\mathsf{W}}_p$ is Lipschitz w.r.t. $\overline{\mathsf{W}}_1$. This reduces the argument to characterizing the Gâteaux directional derivative, which requires extra work.

The nonlinearity of the Hadamard directional derivative means that the nonparametric bootstrap is inconsistent for $\overline{\mathsf{W}}_p$. Nevertheless, subsampling or the m-out-of-n bootstrap can still consistently estimate the limit law. The next lemma deals with the m-out-of-n bootstrap. Below, we say that the bootstrap quantity S_n^B converges in distribution to a nonrandom law ρ in probability if $\sup_{g\in\mathsf{BL}_1(\mathbb{R})} \left|\mathbb{E}^B[g(S_n^B)] - \mathbb{E}_{S\sim\rho}[g(S)]\right| \to 0$ in probability, where $\mathsf{BL}_1(\mathbb{R})$ denotes the class of (bounded) 1-Lipschitz functions $g : \mathbb{R} \to [-1, 1]$ and $\mathbb{E}^B$ is the conditional expectation given the sample.

Lemma 3. Let $X_1^B, \dots, X_m^B$ and $Y_1^B, \dots, Y_m^B$ be independent samples from $\hat{\mu}_n$ and $\hat{\nu}_n$, respectively, where $m = m_n \to \infty$ with $m = o(n)$. Set $\hat{\mu}_{m,n}^B = m^{-1}\sum_{i=1}^m \delta_{X_i^B}$ and $\hat{\nu}_{m,n}^B = m^{-1}\sum_{i=1}^m \delta_{Y_i^B}$. Under the setting of Theorem 1 and conditionally on the data, the sequence $\sqrt{m}(\overline{\mathsf{W}}_p^p(\hat{\mu}_{m,n}^B, \hat{\nu}_{m,n}^B) - \overline{\mathsf{W}}_p^p(\hat{\mu}_n, \hat{\nu}_n))$ converges in distribution to the limit in (2.6), in probability.

Remark 6 (Bias of plug-in estimator for $\overline{\mathsf{W}}_p$ and correction)**.** In general, the limit distributions for the max-sliced distance in Theorem 2 have positive means, which implies that empirical $\overline{\mathsf{W}}_p$ tends to be upward biased at the order of $n^{-1/2}$. Such an upward bias commonly appears in plug-in estimation of the maximum of a nonparametric function (cf. [47]). One may correct this bias using a precision correction similar to [47]. Namely, define $\widehat{\mathbb{W}}_p(\alpha) := \overline{\mathsf{W}}_p^p(\hat{\mu}_n, \hat{\nu}_n) - k_\alpha/\sqrt{n}$, where k_α is the α-quantile of $\sup_{\theta \in \mathfrak{S}_{\mu,\nu}} [\mathbb{G}_\mu(\theta) + \mathbb{G}'_\nu(\theta)]$, which can be estimated via the subsampling or the m-out-of-n bootstrap. Provided that k_α is a continuity point of the distribution function of the limit variable, this estimator is upward α-quantile unbiased, and in particular, median unbiased when $\alpha = 1/2$, meaning that $\mathbb{P}(\widehat{\mathbb{W}}_p(\alpha) \le \overline{\mathsf{W}}_p^p(\mu, \nu)) = \alpha + o(1)$.

Remark 7 (Extensions)**.** We address two possible extensions of Theorems 1 and 2.

(Null case): As $\varphi^\theta \circ \mathfrak{p}^\theta$ and $\psi^\theta \circ \mathfrak{p}^\theta$ are constant μ- and ν-a.e., respectively, for each $\theta \in \mathbb{S}^{d-1}$, the limit distributions in Theorems 1 and 2 degenerate to zero under the null, i.e., $\mu = \nu$. In $d = 1$, [60] derived a limit distribution for $\mathsf{W}_2(\hat{\mu}_n, \mu)$ using the quantile function representation of W_2, which requires several technical conditions concerning the tail of the Lebesgue density of μ. Their argument hinges on approximating the (general) sample quantile process by the uniform quantile process, and applying results for the latter. This argument does not directly extend to the sliced distances, as the quantile process is indexed by the additional projection parameter $\theta \in \mathbb{S}^{d-1}$. Also, it seems nontrivial to find simple conditions on μ itself under which the projected distributions $\mathfrak{p}^\theta_\# \mu$ satisfy the conditions from [60] for all (or uniformly over) $\theta \in \mathbb{S}^{d-1}$. We leave null limit distributions for $\underline{\mathsf{W}}_p$ and $\overline{\mathsf{W}}_p$ for future research.

(Unbounded support): For compactly supported distributions, dual potentials are Lipschitz continuous whose Lipschitz constants depend only on the order p and the radius of the support. This is a key result when we apply Proposition 1 to sliced Wasserstein distances; see also the discussion after Theorem 1. For distributions with unbounded supports, the recent work of [139] derives local Lipschitz estimates for dual potentials under a high-level anti-concentration assumption (see the discussion around their Lemma 11), but for empirical distributions, the local Lipschitz constants depend on the sample size n, which hinders the application of Proposition 1 (the function class being dependent on n does not bring any difficulty to finding error bounds but would require a very delicate argument for limit theorems). The extension of Theorems 1 and 2 to simple moment conditions would require highly technical arguments and hence is beyond the scope of the present paper.

Order $p = 1$

The analysis for $\underline{\mathsf{W}}_1$ relies on the explicit expression of W_1 between distributions on $\mathbb{R}$ as the L^1 distance between distribution functions, whereby

$$\underline{\mathsf{W}}_1(\mu, \nu) = \int_{\mathbb{S}^{d-1}} \int_{\mathbb{R}} |F_\mu(t;\theta) - F_\nu(t;\theta)|\, dt\, d\sigma(\theta). \tag{2.7}$$

For $\overline{\mathsf{W}}_1$, the Kantorovich-Rubinstein duality yields

$$\overline{\mathsf{W}}_1(\mu, \nu) = \sup_{\theta\in\mathbb{S}^{d-1}} \sup_{\varphi\in\mathsf{Lip}_{1,0}(\mathbb{R})} \left[\int \varphi(\theta^\intercal x) d(\mu - \nu)(x) \right]. \tag{2.8}$$

Here $\mathsf{Lip}_{1,0}(\mathbb{R})$ denotes the class of 1-Lipschitz functions φ on $\mathbb{R}$ with $\varphi(0) = 0$. These explicit expressions enable us to derive a limit distribution theory for $\underline{\mathsf{W}}_1$ and $\overline{\mathsf{W}}_1$ under mild moment conditions, as presented next.

To state the result, set λ as the Lebesgue measure on $\mathbb{R}$ and denote

$$\mathrm{sign}(t) = \mathbb{1}_{(0,\infty)}(t) - \mathbb{1}_{(-\infty,0)}(t).$$

Recall that a stochastic process $(Y(t))_{t\in T}$ indexed by a measurable space T is called measurable if $(t, \omega) \mapsto Y(t, \omega)$ is jointly measurable.

Theorem 3 (Limit distribution for $\underline{\mathsf{W}}_1$ and $\overline{\mathsf{W}}_1$)**.** Let $\epsilon > 0$ be arbitrary.

(i) If $\mu \in \mathcal{P}_{2+\epsilon}(\mathbb{R}^d)$ and $\nu \in \mathcal{P}_1(\mathbb{R}^d)$, then there exists a measurable, centered Gaussian process $\mathsf{G}_\mu = (\mathsf{G}_\mu(t,\theta))_{(t,\theta)\in\mathbb{R}\times\mathbb{S}^{d-1}}$ with paths in $L^1(\lambda \otimes \sigma)$ and covariance function

$$\mathrm{Cov}(\mathsf{G}_\mu(s,\theta), \mathsf{G}_\mu(t,\vartheta)) = \mu(\{x \in \mathbb{R}^d : \theta^\intercal x \leq s, \vartheta^\intercal y \leq t\}) - F_\mu(s;\theta)F_\mu(t;\vartheta), \quad (2.9)$$

such that

$$\sqrt{n}(\underline{\mathsf{W}}_1(\hat{\mu}_n,\nu) - \underline{\mathsf{W}}_1(\mu,\nu)) \xrightarrow{d} \iint [\mathrm{sign}(F_\mu - F_\nu)]\mathsf{G}_\mu d\lambda d\sigma + \iint_{F_\mu = F_\nu} |\mathsf{G}_\mu| d\lambda d\sigma. \quad (2.10)$$

(ii) If $\mu, \nu \in \mathcal{P}_{2+\epsilon}(\mathbb{R}^d)$, then

$$\begin{aligned} &\sqrt{n}(\underline{\mathsf{W}}_1(\hat{\mu}_n,\hat{\nu}_n) - \underline{\mathsf{W}}_1(\mu,\nu)) \\ &\qquad \xrightarrow{d} \iint [\mathrm{sign}(F_\mu - F_\nu)](\mathsf{G}_\mu - \mathsf{G}'_\nu) d\lambda d\sigma + \iint_{F_\mu = F_\nu} |\mathsf{G}_\mu - \mathsf{G}'_\nu| d\lambda d\sigma, \end{aligned}$$

where G'_ν is independent of G_μ given in (i) and defined analogously.

(iii) Assume $\mu \in \mathcal{P}_{4+\epsilon}(\mathbb{R}^d)$ and $\nu \in \mathcal{P}_1(\mathbb{R}^d)$. Consider the function class

$$\mathcal{F} = \left\{\varphi \circ \mathfrak{p}^\theta : \theta \in \mathbb{S}^{d-1}, \varphi \in \mathsf{Lip}_{1,0}(\mathbb{R})\right\}.$$

Then, there exists a tight μ-Brownian bridge process G_μ in $\ell^\infty(\mathcal{F})$ such that

$$\sqrt{n}(\overline{\mathsf{W}}_1(\hat{\mu}_n,\nu) - \overline{\mathsf{W}}_1(\mu,\nu)) \xrightarrow{d} \sup_{f\in M_{\mu,\nu}} G_\mu(f),$$

where $M_{\mu,\nu} = \{f \in \overline{\mathcal{F}}^\mu : \mu(f - \nu(f)) = \overline{\mathsf{W}}_1(\mu,\nu)\}$ and $\overline{\mathcal{F}}^\mu$ is the completion of $\mathcal{F}$ for the standard deviation pseudometric $(f,g) \mapsto \sqrt{\mathrm{Var}_\mu(f-g)}$.

(iv) If $\mu, \nu \in \mathcal{P}_{4+\epsilon}(\mathbb{R}^d)$, then there exists a tight ν-Brownian bridge process G'_ν in $\ell^\infty(\mathcal{F})$ independent of G_μ above such that

$$\sqrt{n}(\overline{\mathsf{W}}_1(\hat{\mu}_n,\hat{\nu}_n) - \overline{\mathsf{W}}_1(\mu,\nu)) \xrightarrow{d} \sup_{f\in M'_{\mu,\nu}} [G_\mu(f) - G'_\nu(f)],$$

where $M'_{\mu,\nu} = \{f \in \overline{\mathcal{F}}^{\mu,\nu} : (\mu - \nu)(f) = \overline{\mathsf{W}}_1(\mu,\nu)\}$ and $\overline{\mathcal{F}}^{\mu,\nu}$ is the completion of $\mathcal{F}$ for the pseudometric $(f,g) \mapsto \sqrt{\mathrm{Var}_\mu(f-g)} + \sqrt{\mathrm{Var}_\nu(f-g)}$.

While Theorem 1 requires distributions to have compact and convex support, Theorem 3 holds under mild moment assumptions. The derivation for $\underline{\mathsf{W}}_1$ uses the CLT in L^1 to deduce convergence of empirical projected distribution functions in $L^1(\lambda\otimes\sigma)$. The limit distribution is then obtained via the functional delta method by casting $\underline{\mathsf{W}}_1$ as the $L^1(\lambda\otimes\sigma)$ norm between distribution functions and characterizing the corresponding Hadamard directional derivative. For $\overline{\mathsf{W}}_1$, we use the Kantorovich-Rubinstein duality in conjunction with the fact that the class of projection 1-Lipschitz functions is Donsker under the said moment condition. In Part (iii), if $\mu = \nu$, then $M_{\mu,\nu} = \overline{\mathcal{F}}^{\mu}$, and since G_μ has uniformly continuous paths w.r.t. the standard deviation pseudometric, the limit variable becomes $\sup_{f\in\mathcal{F}} G_\mu(f)$. Likewise, in Part (iv), the limit variable becomes $\sup_{f\in\mathcal{F}}[G_\mu(f) - G'_\nu(f)]$ when $\mu = \nu$.

For $\underline{\mathsf{W}}_1$, if the second term on the right-hand side of (2.10) is zero, then the asymptotic normality holds. We state this result including its two-sample analogue next.

Corollary 3 (Asymptotic normality for $\underline{\mathsf{W}}_1$)**.** For $\mu \in \mathcal{P}(\mathbb{R}^d)$ and $\theta \in \mathbb{S}^{d-1}$, define $\bar{l}^{\theta}_{\mu} = \sup \operatorname{spt}(\mathfrak{p}^{\theta}_{\sharp}\mu)$ and $\underline{l}^{\theta}_{\mu} = \inf \operatorname{spt}(\mathfrak{p}^{\theta}_{\sharp}\mu)$. The following hold.

(i) Under the assumption of Theorem 3 Part (i), if in addition $F_\mu(t;\theta) \neq F_\nu(t;\theta)$ for $(\lambda\otimes\sigma)$-almost all $(t,\theta) \in \{(s,\vartheta) : s \in [\underline{l}^{\vartheta}_{\mu}, \bar{l}^{\vartheta}_{\mu}], \vartheta \in \mathbb{S}^{d-1}\}$, then

$$\sqrt{n}(\underline{\mathsf{W}}_1(\hat{\mu}_n,\nu) - \underline{\mathsf{W}}_1(\mu,\nu)) \xrightarrow{d} N(0, v_1^2),$$

where v_1^2 is the variance of $\iint [\operatorname{sign}(F_\mu - F_\nu)]\mathsf{G}_\mu d\lambda d\sigma$. The asymptotic variance v_1^2 agrees with the semiparametric efficiency bound for estimating $\underline{\mathsf{W}}_1(\cdot,\nu)$ at μ.

(ii) Under the assumption of Theorem 3 Part (ii), if in addition $F_\mu(t;\theta) \neq F_\nu(t;\theta)$ for $(\lambda\otimes\sigma)$-almost all $(t,\theta) \in \{(s,\vartheta) : s \in [\underline{l}^{\vartheta}_{\mu} \wedge \underline{l}^{\vartheta}_{\nu}, \bar{l}^{\vartheta}_{\mu} \vee \bar{l}^{\vartheta}_{\nu}], \vartheta \in \mathbb{S}^{d-1}\}$, then

$$\sqrt{n}(\underline{\mathsf{W}}_1(\hat{\mu}_n,\hat{\nu}_n) - \underline{\mathsf{W}}_1(\mu,\nu)) \xrightarrow{d} N(0, v_1^2 + w_1^2),$$

where v_1^2 is as above and w_1^2 is the variance of $\iint [\mathrm{sign}(F_\mu - F_\nu)]\mathsf{G}'_\nu d\lambda d\sigma$. The asymptotic variance $v_1^2 + w_1^2$ agrees with the semiparametric efficiency bound for estimating $\underline{\mathsf{W}}_1$ at (μ, ν).

Finally, bootstrap consistency (as in Theorem 1) holds for both cases (i) and (ii).

As an example, the above asymptotic normality holds when the population distributions are both Gaussian.

Example 1. Consider $\mu = N(\xi_1, \Sigma_1)$ and $\nu = N(\xi_2, \Sigma_2)$. Then $\mathfrak{p}^\theta_\sharp \mu = N(\theta^\intercal \xi_1, \theta^\intercal \Sigma_1 \theta)$ and $\mathfrak{p}^\theta_\sharp \mu = N(\theta^\intercal \xi_2, \theta^\intercal \Sigma_2 \theta)$. In this case, as long as $\xi_1 \neq \xi_2$ or $\Sigma_1 \neq \Sigma_2$, we have $\sigma(\{\theta : \theta^\intercal \xi_1 = \theta^\intercal \xi_2 \text{ and } \theta^\intercal \Sigma_1 \theta = \theta^\intercal \Sigma_2 \theta\}) = 0$, which follows from the fact that $\frac{Z}{\|Z\|} \sim \sigma$ for $Z \sim N(0, I_d)$ and Lemma 1 in [158]. Thus, $F_\mu(t;\theta) \neq F_\nu(t;\theta)$ for $(\lambda \otimes \sigma)$-almost all $(t, \theta) \in \mathbb{R} \times \mathbb{S}^{d-1}$ and the conclusion of Corollary 3 applies.

Remark 8 (Comparison with [218])**.** The work [218], which appeared after our paper was posted on the arXiv, derived limit distribution results for $\underline{\mathsf{W}}_1$ and $\overline{\mathsf{W}}_1$ similar to the above. For the average-slicing in the one-sample case, they assume the slightly weaker moment condition $\int \sqrt{\mathbb{P}(\|X\| > t)}dt < \infty$, where $X \sim \mu$. Their proof strategy is essentially the same as ours and the above condition is imposed to verify a CLT for the empirical projected distribution function in $L^1(\lambda \otimes \sigma)$, as stated in the beginning of Step 1 in the proof of our Theorem 3(i). Our assumption, which requires finite $(2 + \epsilon)$-th moment, is meant to provide an elementary moment condition to guarantee the said CLT in $L^1(\lambda \otimes \sigma)$, and it is not difficult to see that the weaker (yet more high-level) condition $\int \sqrt{\mathbb{P}(\|X\| > t)}dt < \infty$ suffices for our proof to go through; see the discussion around equation (A.9) in the supplementary material. For the max-slicing case (with $p = 1$), [214] assume the exact same moment condition, although they do not cover the alternative case.

2.5 Smooth Wasserstein Distance with Compactly Supported Kernels

We study smooth Wasserstein distances with compactly supported kernels, namely, when the considered distributions are convolved with a mollifier (also known as a bump function). Smoothing by means of the Gaussian kernel was extensively studied before for structural and statistical properties (see literature review below), but it remains unclear how to efficiently compute the Gaussian-smoothed Wasserstein distance.[3] Despite recent advancement in computation of continuous-to-continuous OT between distributions with smooth densities [197, 196], these approaches cannot handle the Gaussian-smoothed W_p since they assume compactly supported distributions.[4] Furnishing a smoothed framework that enjoys the structural and statistical virtues as the Gaussian-smoothed Wasserstein distance, while being amenable for efficient computation via the aforementioned approaches is the main motivation of this section. The main use case of compactly supported kernel paradigm is thus when the population distributions are also compactly supported. For the sake of generality, we next define the distance and provide structural properties for arbitrary $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$ distributions (possibly with unbounded support), but restrict to the compactly supported case for the statistical analysis. In Section A.4 of the supplementary material, we provide a thorough account of how to lift the algorithm from [197] to compute our smooth distance.

[3]While the empirical Gaussian-smoothed Wasserstein distance can be evaluated by sampling the kernel and applying computational methods for classic W_p, this approach fails to exploit the smoothness of this framework and sacrifices the statistical advantages pertaining to estimation and inference.

[4]Convolution with a Gaussian kernel does not preserve compact support. In applications where compact support of the convolved distributions is immaterial, the Gaussian kernel is, however, a natural choice.

2.5.1 Background

To set up the smooth Wasserstein distance, we first define a smoothing kernel as follows. Let $\chi \in C^\infty(\mathbb{R}^d)$ be any non-negative function with $\int_{\mathbb{R}^d} \chi(x)dx = 1$ and $\int_{\mathbb{R}^d} \|x\|^p \chi(x)dx < \infty$, for all $1 \le p < \infty$. Then, for any $\sigma > 0$, define $\chi_\sigma = \sigma^{-d}\chi(\cdot/\sigma) \in C^\infty(\mathbb{R}^d)$ and let $\eta_\sigma \in \mathcal{P}(\mathbb{R}^d)$ be a probability measure whose (Lebesgue) density is χ_σ. We call η_σ a *smoothing kernel* of parameter σ, and define the corresponding smooth Wasserstein distance as follows.

Definition 3 (Smooth Wasserstein distances)**.** Let $1 \le p < \infty$ and η_σ be a smoothing kernel. The associated smooth p-Wasserstein distance between $\mu, \nu \in \mathcal{P}_p(\mathbb{R}^d)$ is

$$\mathsf{W}_p^{\eta_\sigma}(\mu, \nu) := \mathsf{W}_p(\mu * \eta_\sigma, \nu * \eta_\sigma).$$

Example 2 (Standard mollifier)**.** A canonical example of a smooth compactly supported function is the standard mollifier

$$\chi(x) = \begin{cases} \frac{1}{C_\chi} \exp\left(-\frac{1}{1-\|x\|^2}\right) & \text{if } \|x\| < 1 \\ 0 & \text{otherwise} \end{cases}, \tag{2.11}$$

where $C_\chi = \int_{\mathbb{R}^d} \chi d\lambda$, from which a compactly supported kernel is readily constructed. Our results, however, are not specialized to the mollifier kernel and hold for any η_σ as described above.

As reviewed next, Gaussian-smoothed Wasserstein distances, i.e., when $\eta_\sigma = \gamma_\sigma := N(0, \sigma^2 I_d)$, have been extensively studied for their structural and statistical properties.

Literature review

Gaussian-smoothed Wasserstein distances were introduced in [102] as a means to mitigate the curse of dimensionality in empirical estimation. Indeed, [102] demonstrated that

$\mathbb{E}[\mathsf{W}_p^{(\gamma_\sigma)}(\hat{\mu}_n,\mu)] = O(n^{-1/2})$, for $p = 1, 2$, in arbitrary dimension provided that μ is sufficiently sub-Gaussian (cf. the recent preprint [22] for sharp bounds on the sub-Gaussian constant for which the rate is parametric when $p = 2$). Structural properties of $\mathsf{W}_1^{(\gamma_\sigma)}$ were explored in [100], showing that it metrizes the classic Wasserstein topology and establishing regularity in σ. These structural and statistical results were later generalized to $\mathsf{W}_p^{(\gamma_\sigma)}$ for any $p > 1$ [152], and asymptotics of the smooth distance as $\sigma \to \infty$ were explored [40]. Relations between $\mathsf{W}_p^{(\gamma_\sigma)}$ and maximum mean discrepancies were studies in [221], and nonparametric mixture model estimation under $\mathsf{W}_p^{(\gamma_\sigma)}$ was considered [114], again demonstrating scalability of error bounds with dimension. The study of limit distributions for empirical $\mathsf{W}_p^{(\gamma_\sigma)}$ was initiated in [101] for $p = 1$ in the one-sample case, extended to the two-sample setting in [172], and generalized to arbitrary $p > 1$ via a non-trivial application of the functional delta method in [105]. These works also considered bootstrap consistency and applications to minimum distance estimation and homogeneity testing. To date, a relatively complete limit distribution theory of $\mathsf{W}_p^{(\gamma_\sigma)}$ in arbitrary dimension is available, as opposed to the rather limited account of classic W_p.

2.5.2 Structural properties

We henceforth consider a compactly supported smoothing kernel η_σ and adopt the shorthand $\mathsf{W}_p^\sigma := \mathsf{W}_p^{\eta_\sigma}$. We start by revisiting structural properties previously established for the Gaussian-smoothed case and demonstrate that they remain valid for W_p^σ.

Proposition 3 (Stability of W_p^σ)**.** For any $1 \leq p < \infty$, $\sigma > 0$, and $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$, we have

$$\mathsf{W}_p^\sigma(\mu,\nu) \leq \mathsf{W}_p(\mu,\nu) \leq \mathsf{W}_p^\sigma(\mu,\nu) + 2\sigma(\mathbb{E}_{\eta_1}[\|X\|^p])^{1/p}.$$

In particular, $\lim_{\sigma\downarrow 0} \mathsf{W}_p^\sigma(\mu,\nu) = \mathsf{W}_p(\mu,\nu)$.

The first bound is due to contractivity of W_p w.r.t. convolution. Constructing a coupling between $\rho \in \mathcal{P}(\mathbb{R}^d)$ and $\rho * \eta_\sigma$ with total cost $\sigma(\mathbb{E}_{\eta_1}[\|X\|^p])^{1/p}$ proves the second. Since η_σ is compactly supported, $\mathrm{spt}(\eta_1)$ is contained in a ball of radius $r > 0$, whereby $(\mathbb{E}_{\eta_1}[\|X\|^p])^{1/p} \leq r$ which contrasts the dimension dependent gap for Gaussian kernels; cf. [100, 152].

As the smooth distance converges to the standard distance as $\sigma \to 0$, it is natural to expect that optimal couplings converge as well. This is stated in the next proposition.

Proposition 4 (Stability of transport plans)**.** For $1 \leq p < \infty$, $\mu, \nu \in \mathcal{P}_p(\mathbb{R}^d)$, and $\sigma_k \downarrow 0$. Let π_k be an optimal coupling for $\mathsf{W}_p^{\sigma_k}(\mu, \nu)$ for each $k \in \mathbb{N}$. Then, there exists an optimal coupling π for $\mathsf{W}_p(\mu, \nu)$ for which $\pi_k \xrightarrow{w} \pi$ along a subsequence.

The proof of this result follows that of Theorem 4 in [100] and [105] with only minor changes and is hence omitted. Note that when the limiting π is unique (e.g., when $p > 1$ and μ has a density), then extraction of a subsequence is not needed.

We next show that W_p^σ is indeed a metric on $\mathcal{P}_p(\mathbb{R}^d)$ that induces the Wasserstein topology.

Proposition 5 (Metric and topological structure)**.** For $1 \leq p < \infty$ and $\sigma > 0$, W_p^σ is a metric on $\mathcal{P}_p(\mathbb{R}^d)$ inducing the same topology as W_p.

The proof of Proposition 5 follows by observing that the characteristic function of η_σ vanishes on at most a null set.

2.5.3 Statistical analysis

This section studies empirical convergence rates and limit distributions for the smooth Wasserstein distances with compactly supported kernels and population distributions. Let $\mathcal{X} \subset \mathbb{R}^d$ be compact, set $\mathcal{X}_\sigma := \mathcal{X} + \overline{B(0,\sigma)}$, and assume for simplicity that the density of η_σ is positive on $B(0,\sigma)$, and identically zero on $\mathbb{R}^d \setminus B(0,\sigma)$. For any $\mu \in \mathcal{P}(\mathcal{X})$, the set $\mathcal{X}_\sigma$ contains the support of the convolved measure $\mu * \eta_\sigma$.

Limit distributions for $p > 1$ under the alternative

Building on the unified framework from Proposition 1, the next theorem establishes asymptotic normality of empirical W_p^σ under the alternative. The null case and the $p = 1$ setting are treated in the sequel.

Theorem 4 (Limit distributions for W_p^σ under the alternative)**.** Set $1 < p < \infty$, $\sigma > 0$, $\mathsf{V}_p^\sigma := [\mathsf{W}_p^\sigma]^p$, and let $\mu, \nu \in \mathcal{P}(\mathcal{X})$ be such that $\mathrm{int}(\mathrm{spt}(\mu * \eta_\sigma))$ is connected. Let φ be an OT potential from $\mu * \eta_\sigma$ to $\nu * \eta_\sigma$ for W_p, which is unique on $\mathrm{int}(\mathrm{spt}(\mu * \eta_\sigma))$ up to additive constants. The following hold.

(i) We have

$$\sqrt{n}\left(\mathsf{V}_p^\sigma(\hat{\mu}_n, \nu) - \mathsf{V}_p^\sigma(\mu, \nu)\right) \xrightarrow{d} N\left(0, v_p^2\right)$$

where $v_p^2 := \mathrm{Var}_\mu(\varphi * \chi_\sigma)$. The asymptotic variance v_p^2 coincides with the semiparametric effiency bound for estimating $\mathsf{V}_p^\sigma(\cdot, \nu)$ at μ. Also, provided that $v_p^2 > 0$, we have

$$\sup_{t \in \mathbb{R}} \left| \mathbb{P}^B\left(\sqrt{n}(\mathsf{V}_p^\sigma(\hat{\mu}_n^B, \nu) - \mathsf{V}_p^\sigma(\hat{\mu}_n, \nu)) \le t\right) - \mathbb{P}(N(0, v_p^2) \le t) \right| \xrightarrow{\mathbb{P}} 0.$$

(ii) If in addition $\nu * \eta_\sigma$ has connected support, then

$$\sqrt{n}\left(\mathsf{V}_p^\sigma(\hat{\mu}_n, \hat{\nu}_n) - \mathsf{V}_p^\sigma(\mu, \nu)\right) \xrightarrow{d} N\left(0, v_p^2 + w_p^2\right),$$

where v_p^2 is as in (i) and $w_p^2 := \mathrm{Var}_\nu(\varphi^c * \chi_\sigma)$. The asymptotic variance $v_p^2 + w_p^2$ coincides with the semiparametric efficiency bound for estimating V_p^σ at (μ, ν). Also, provided that $v_p^2 + w_p^2 > 0$, we have

$$\sup_{t\in\mathbb{R}} \left|\mathbb{P}^B\Big(\sqrt{n}(\mathsf{V}_p^\sigma(\hat{\mu}_n^B, \hat{\nu}_n^B) - \mathsf{V}_p^\sigma(\hat{\mu}_n, \hat{\nu}_n)) \leq t\Big) - \mathbb{P}(N(0, v_p^2 + w_p^2) \leq t)\right| \xrightarrow{\mathbb{P}} 0.$$

The proof of Theorem 4 applies Proposition 1 to the functional $\rho * \eta_\sigma \mapsto \mathsf{W}_p^p(\rho * \eta_\sigma, \nu * \eta_\sigma)$ for $\rho \in \mathcal{P}(\mathcal{X})$ with $\mathrm{spt}(\rho) \subset \mathrm{spt}(\mu)$. To this end, we show that this functional is Lipschitz continuous w.r.t. $\|\cdot\|_{\infty,B}$ for the unit ball B in $L^2(\mathcal{X}_\sigma)$, which follows by duality (2.3) and uniform bounds on the OT potentials (cf. Remark 1.13 in [207]). The differentiability result follows by adapting the Gaussian kernel case (cf. Lemma 3.3 of [105]). To prove weak convergence of the smoothed empirical process $\sqrt{n}(\hat{\mu}_n - \mu) * \eta_\sigma$ in $\ell^\infty(B)$, we employ the CLT in $L^2(\mathcal{X}_\sigma)$ and use a linear isometry from $L^2(\mathcal{X}_\sigma)$ into $\ell^\infty(B)$. Linearity of the derivative yields asymptotic efficiency and bootstrap consistency.

Remark 9 (Connectedness assumption)**.** By Lemma 57 in the supplementary material, the condition from Theorem 4 that $\mathrm{int}(\mathrm{spt}(\mu * \eta_\sigma))$ is connected holds whenever μ itself has connected support.

The ideas from the proof of Theorem 4 coupled with Hilbertian structure of $L^2(\mathcal{X}_\sigma)$ yield rates of convergence in expectation for empirical W_p^σ.

Proposition 6 (Parametric rate)**.** For $1 < p < \infty$, $\sigma > 0$, and $\mu, \nu \in \mathcal{P}(\mathcal{X})$ with $\mu \neq \nu$, we have

$$\mathbb{E}\left[\left|\mathsf{W}_p^\sigma(\hat{\mu}_n, \nu) - \mathsf{W}_p^\sigma(\mu, \nu)\right|\right] \leq 2\|\chi_\sigma\|_\infty \sqrt{\lambda(B(0,\sigma))\lambda(\mathcal{X}_\sigma)}\,\mathrm{diam}(\mathcal{X}_\sigma)^p\big[\mathsf{W}_p^\sigma(\mu, \nu)\big]^{1-p} n^{-1/2}.$$

Limit distributions for $p = 2$ under the null

We derive limit distributions for W_2^σ under the null. Our approach relies on the CLT in Hilbert spaces and is thus limited to $p = 2$. Let C_0^∞ denote the space of infinitely differentiable, compactly supported real functions on $\mathbb{R}^d$.

Definition 4 (Sobolev spaces and their duals)**.** The Sobolev seminorm of a differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ w.r.t. a reference measure $\mu \in \mathcal{P}(\mathbb{R}^d)$ is denoted by $\|f\|_{\dot{H}^{1,2}(\mu)} := \|\nabla f\|_{L^2(\mu)}$. The homogeneous Sobolev space is defined as the completion of $C_0^\infty + \mathbb{R}$ with respect to $\|\cdot\|_{\dot{H}^{1,2}(\mu)}$. The dual Sobolev space $\dot{H}^{-1,2}(\mu)$ is the topological dual of $\dot{H}^{1,2}(\mu)$.

Definition 5 (2-Poincaré inequality)**.** A probability measure $\mu \in \mathcal{P}(\mathbb{R}^d)$ is said to satisfy the 2-Poincaré inequality if there exists $C < \infty$, such that

$$\|f - \mu(f)\|_{L^2(\mu)} \leq C\|\nabla f\|_{L^2(\mu;\mathbb{R}^d)}, \quad f \in C_0^\infty,$$

where $L^2(\mu;\mathbb{R}^k)$ is the space of Borel maps $f : \mathbb{R}^d \to \mathbb{R}^k$ with $\|f\|^2_{L^2(\mu;\mathbb{R}^k)} := \int_{\mathbb{R}^d} \|f\|^2 d\mu < \infty$.

With these definitions in place, we state the limit distribution for W_2^σ.

Theorem 5 (Limit distributions for W_2^σ under the null)**.** Let $\mu \in \mathcal{P}(\mathcal{X})$ be such that $\mu * \eta_\sigma$ satisfies the 2-Poincaré inequality and set $\sigma > 0$. The following hold.

(i) We have

$$\sqrt{n}\mathsf{W}_2^\sigma(\hat{\mu}_n, \mu) \xrightarrow{d} \|\mathbb{G}_\mu\|_{\dot{H}^{-1,2}(\mu * \eta_\sigma)},$$

where $(\mathbb{G}_\mu(f))_{f \in \dot{H}^{1,2}(\mu * \eta_\sigma)}$ is a centered Gaussian process with paths in $\dot{H}^{-1,2}(\mu * \eta_\sigma)$ a.s. and covariance function $\mathrm{Cov}(\mathbb{G}_\mu(f), \mathbb{G}_\mu(g)) = \mathrm{Cov}_\mu(f * \chi_\sigma, g * \chi_\sigma)$.

(ii) Additionally, if $\mu = \nu$, then

$$\sqrt{n}\mathsf{W}_2^\sigma(\hat{\mu}_n, \hat{\nu}_n) \xrightarrow{d} \|\mathbb{G}_\mu - \mathbb{G}'_\mu\|_{\dot{H}^{-1,2}(\mu * \eta_\sigma)},$$

where $\mathbb{G}'_\mu$ is an independent copy of $\mathbb{G}_\mu$.

The proof of Theorem 5 follows a similar approach to the Gaussian kernel case. In contrast to the proof of Proposition 3.1 in [105], to show weak convergence of the smoothed empirical process in $\dot{H}^{-1,2}(\mu * \eta_\sigma)$, we apply the CLT in Hilbert spaces. To this end, we first verify that the smoothed empirical process has paths in $\dot{H}^{-1,2}(\mu * \eta_\sigma)$. This step requires control of the inverse of the density of $\mu * \eta_\sigma$, which decays to zero near the boundary of its support; see the proof of Lemma 58 in the supplementary material. The extension to general $1 < p < \infty$ requires a much finer analysis than provided in the proof of Lemma 58, and thus we focus on the $p = 2$ case.

Remark 10 (Poincaré inequality)**.** In Theorem 5, a sufficient condition for $\mu * \eta_\sigma$ to satisfy the 2-Poincaré inequality is that both μ and η_σ satisfy 2-Poincaré inequalities (see Proposition 1.1 in [210]). The kernel can always be chosen to satisfy 2-Poincaré by simply constructing it from the standard mollifier from Example 2. In that case, η_σ is a log-concave measure (cf. [135, 176]) and hence satisfies the 2-Poincaré inequality [24, 145].

Analogously to Proposition 6, parametric rates for empirical W_2^σ under the null follow from Hilbertian structure of $\dot{H}^{-1,2}(\mu * \eta_\sigma)$ and the ideas from the proof of Theorem 5.

Proposition 7 (Parametric rate)**.** For $\sigma > 0$ and $\mu \in \mathcal{P}(\mathcal{X})$ for which $\mu * \eta_\sigma$ satisfies the 2-Poincaré inequality with constant $C_{\mu,\sigma}$, we have

$$\mathbb{E}\left[\mathsf{W}_2^\sigma(\hat{\mu}_n, \mu)\right] \leq 2C_{\mu,\sigma}\sqrt{(1 \vee \|\chi_\sigma^2\|_\infty)\lambda(\mathcal{X}_\sigma)}n^{-1/2}.$$

Limit distributions for $p = 1$

We now treat the limit distributions for W_1^σ under both the null and the alternative. The Kantorovich-Rubinstein duality for W_1 enables us to do so in the absence of the additional assumptions required when $p > 1$. In what follows, let $\mathsf{Lip}_{1,0}$ denote the set of

1-Lipschitz functions f on $\mathbb{R}^d$ with $f(0) = 0$ and $\mathcal{F}_\sigma = \{f * \chi_\sigma : f \in \mathsf{Lip}_{1,0}\}$. Observe that $\mathsf{W}_1^\sigma(\mu, \nu) = \sup_{f \in \mathcal{F}_\sigma} (\mu - \nu)(f)$ by the Kantorovich-Rubinstein duality.

Theorem 6 (Limit distributions for W_1^σ)**.** Let $\sigma > 0$ and $\mu, \nu \in \mathcal{P}(\mathcal{X})$. There exist independent, tight μ- and ν-Brownian bridge process G_μ and G'_ν in $\ell^\infty(\mathcal{F}_\sigma)$, respectively, such that:

(i) We have

$$\sqrt{n}(\mathsf{W}_1^\sigma(\hat{\mu}_n, \nu) - \mathsf{W}_1^\sigma(\mu, \nu)) \xrightarrow{d} \sup_{f \in M_\sigma} G_\mu(f),$$

where $M_\sigma = \{f \in \overline{\mathcal{F}}_\sigma^\mu : \mu(f - \nu(f)) = \mathsf{W}_1^\sigma(\mu, \nu)\}$ and $\overline{\mathcal{F}}_\sigma^\mu$ is the completion of $\mathcal{F}_\sigma$ for the pseudometric $(f, g) \mapsto \sqrt{\mathrm{Var}_\mu(f - g)}$.

(ii) We have

$$\sqrt{n}(\mathsf{W}_1^\sigma(\hat{\mu}_n, \hat{\nu}_n) - \mathsf{W}_1^\sigma(\mu, \nu)) \xrightarrow{d} \sup_{f \in M'_\sigma} [G_\mu(f) - G'_\nu(f)],$$

where $M'_\sigma = \{f \in \overline{\mathcal{F}}_\sigma^{\mu,\nu} : (\mu - \nu)(f) = \mathsf{W}_1^\sigma(\mu, \nu)\}$ and $\overline{\mathcal{F}}_\sigma^{\mu,\nu}$ is the completion of $\mathcal{F}_\sigma$ for the pseudometric $(f, g) \mapsto \sqrt{\mathrm{Var}_\mu(f - g)} + \sqrt{\mathrm{Var}_\nu(f - g)}$.

Theorem 6 follows by showing that the function class $\mathcal{F}_\sigma$ is Donsker combined with the extended functional delta method for the supremum functional. The proof of Theorem 6 also establishes parametric rates for empirical W_1^σ.

Corollary 4 (Parametric rate)**.** For $\sigma > 0$, $\mu \in \mathcal{P}(\mathcal{X})$, we have $\mathbb{E}[\mathsf{W}_1^\sigma(\hat{\mu}_n, \mu)] = O(n^{-1/2})$.

We conclude this section by referring the reader to Section A.4 of the supplementary material for an account of computational aspects for smooth Wasserstein distances. There, we outline the algorithm from [197], show how to lift it to compute W_2^σ, and discuss limitations of that method.

2.6 Entropic Optimal Transport

EOT is an efficiently-computable convexification of the OT problem. The general machinery of Proposition 1 enables deriving limit theorems for empirical EOT, generalizing previously available statements to allow for dependent data. Our theory also provides new results on semiparametric efficiency of empirical EOT and consistency of the bootstrap estimate.

2.6.1 Background

EOT regularizes OT by the Kullback-Leibler (KL) divergence as

$$\mathsf{S}_c^\epsilon(\mu,\nu) := \inf_{\pi\in\Pi(\mu,\nu)} \int_{\mathbb{R}^d\times\mathbb{R}^d} c(x,y)d\pi(x,y) + \epsilon\,\mathsf{D}_{\mathsf{KL}}(\pi\|\mu\otimes\nu), \tag{2.12}$$

where $\epsilon > 0$ and $\mathsf{D}_{\mathsf{KL}}(\mu\|\nu) := \int \log(d\mu/d\nu)d\mu$ if $\mu \ll \nu$ and $+\infty$ otherwise [179, 130]. We consider the quadratic cost $c(x,y) = \|x-y\|^2/2$, assume that $\epsilon = 1$, and use the shorthand $\mathsf{S}(\mu,\nu) = \mathsf{S}^1_{\|\cdot\|^2/2}(\mu,\nu)$. The assumption that $\epsilon = 1$ comes without loss of generality by a rescaling argument, since $\mathsf{S}^\epsilon_{\|\cdot\|^2/2}(\mu,\nu) = \epsilon\mathsf{S}(\mu_\epsilon,\nu_\epsilon)$, where $\mu_\epsilon = f_{\epsilon\sharp}\mu$ for $f_\epsilon(x) = \epsilon^{-1/2}x$.

To apply Proposition 1 to empirical EOT, we rely on the duality theory for EOT, whereby

$$\mathsf{S}(\mu,\nu) = \sup_{(\varphi,\psi)\in L^1(\mu)\times L^1(\nu)} \int_{\mathbb{R}^d}\varphi d\mu + \int_{\mathbb{R}^d}\psi d\nu - \int_{\mathbb{R}^d\times\mathbb{R}^d} e^{\varphi\oplus\psi-c}d\mu\otimes\nu + 1,$$

with $(\varphi\oplus\psi)(x,y) = \varphi(x)+\psi(y)$. Assuming $\mu,\nu\in\mathcal{P}_2(\mathbb{R}^d)$, the supremum is attained by a pair $(\varphi,\psi)\in L^1(\mu)\times L^1(\nu)$ satisfying the so-called Schrödinger system

$$\begin{aligned}
\int_{\mathbb{R}^d} e^{\varphi(x)+\psi(y')-c(x,y')}d\nu(y') &= 1 \quad \mu\text{-a.e. } x\in\mathbb{R}^d,\\
\int_{\mathbb{R}^d} e^{\varphi(x')+\psi(y)-c(x',y)}d\mu(x') &= 1 \quad \nu\text{-a.e. } y\in\mathbb{R}^d.
\end{aligned} \tag{2.13}$$

We refer to such (φ, ψ) as *optimal EOT potentials* (from μ to ν for φ and vice versa for ψ). Optimal EOT potentials are unique $(\mu \otimes \nu)$-almost everywhere up to additive constants. Conversely, any $(\varphi, \psi) \in L^1(\mu) \times L^1(\nu)$ that admit (2.13) are optimal EOT potentials. See Section 1 in [156] and the references therein for details of the duality results for EOT.

Literature review

The entropic penalty transforms the OT linear optimization problem into a strongly convex one, allowing efficient computation via the Sinkhorn algorithm [53, 4]. While EOT forfeits the metric and topological structure of W_p,[5] it attains fast empirical convergence in certain cases. Specifically, empirical EOT converges as $n^{-1/2}$ for smooth costs and compactly supported distributions [87], or for the squared cost with sub-Gaussian distributions [142].

Limit distributions for EOT (and the Sinkhorn divergence) for $c(x, y) = \|x - y\|^p$ in the discrete support case were provided in [18, 123]. Their approach is to parameterize each marginal by a finite-dimensional simplex vector and find the derivative of the EOT cost w.r.t. the simplex vector to apply the standard delta method; arguably, this approach does not directly extend to general distributions. A CLT for EOT between sub-Gaussian distribution was first derived in [142], showing asymptotic normality of $\sqrt{n}(\mathsf{S}(\hat{\mu}_n, \nu) - \mathbb{E}[\mathsf{S}(\hat{\mu}_n, \nu)])$ and its two-sample analog using the Efron-Stein inequality similar to [64]. The main limitation of this result is that the centering term is the expected empirical EOT, which is undesirable because it does not enable performing inference for $\mathsf{S}(\mu, \nu)$. This limitation was addressed in the recent preprint [65], see the discussion in Remark 11. We provide here an alternative derivation of the CLT that relies on establishing the Hadamard derivatives of the EOT cost w.r.t. the marginals following the

[5] Indeed, e.g., $\mathsf{S}_c^\epsilon(\mu, \mu) \neq 0$; while this can be fixed via centering EOT to obtain the so-called Sinkhorn divergence, it is still not a metric since it lacks the triangle inequality [18].

unified framework from Proposition 1, which automatically leads to asymptotic efficiency of empirical EOT and consistency of the bootstrap estimate, as well as the extension for dependent data. The Hadamard differentiability result (implicit in the proof) may be of independent interest as it pertains to stability analysis of EOT, which has attracted growing interest in the mathematics literature [143, 144, 90, 78, 157].

2.6.2 Statistical analysis

We next state the CLT, asymptotic efficiency, and bootstrap consistency for empirical EOT.

Theorem 7 (CLT, efficiency, and bootstrap consistency for EOT)**.** Suppose that $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$ are sub-Gaussian. Let (φ, ψ) be optimal EOT potentials for (μ, ν). Then, the following hold.

(i) We have $\sqrt{n}(\mathsf{S}(\hat{\mu}_n, \nu) - \mathsf{S}(\mu, \nu)) \stackrel{d}{\to} N(0, \mathfrak{v}_1^2)$ with $\mathfrak{v}_1^2 = \mathrm{Var}_\mu(\varphi)$. The asymptotic variance $\mathfrak{v}_1^2$ coincides with the semiparametric efficiency bound for estimating $\mathsf{S}(\cdot, \nu)$ at μ. Finally, provided that $\mathfrak{v}_1^2 > 0$, we have

$$\sup_{t\in\mathbb{R}} \left| \mathbb{P}^B\Big(\sqrt{n}(\mathsf{S}(\hat{\mu}_n^B, \nu) - \mathsf{S}(\hat{\mu}_n, \nu)) \le t\Big) - \mathbb{P}(N(0, \mathfrak{v}_1^2)) \le t) \right| \stackrel{\mathbb{P}}{\to} 0.$$

(ii) We have $\sqrt{n}(\mathsf{S}(\hat{\mu}_n, \hat{\nu}_n) - \mathsf{S}(\mu, \nu)) \stackrel{d}{\to} N(0, \mathfrak{v}_1^2 + \mathfrak{v}_2^2)$ where $\mathfrak{v}_1^2$ is as in (i) and $\mathfrak{v}_2^2 = \mathrm{Var}_\nu(\psi)$. The asymptotic variance $\mathfrak{v}_1^2 + \mathfrak{v}_2^2$ coincides with the semiparametric efficiency bound for estimating $\mathsf{S}(\cdot, \cdot)$ at (μ, ν). Finally, provided that $\mathfrak{v}_1^2 + \mathfrak{v}_2^2 > 0$, we have

$$\sup_{t\in\mathbb{R}} \left| \mathbb{P}^B\Big(\sqrt{n}(\mathsf{S}(\hat{\mu}_n^B, \hat{\nu}_n^B) - \mathsf{S}(\hat{\mu}_n, \hat{\nu}_n)) \le t\Big) - \mathbb{P}(N(0, \mathfrak{v}_1^2 + \mathfrak{v}_2^2) \le t) \right| \stackrel{\mathbb{P}}{\to} 0.$$

Remark 11 (Comparison with [65])**.** As mentioned in Section 2.6.1, a CLT for one- and two-sample EOT was derived in Theorem 3.6 of [65], whose proof first expands the empirical EOT cost around its expectation and then shows that the bias is negligible. We rederive this result via a markedly different proof technique, relying on the unified framework from Proposition 1, which automatically also implies bootstrap consistency and asymptotic efficiency via Corollary 1 and Proposition 2, both of which were not addressed in [65]. In addition, as Proposition 1 does not assume i.i.d. data, the above result readily extends to dependent data, which falls outside the framework of [65]. For instance, suppose that $\{X_t\}_{t\in\mathbb{Z}}$ is a stationary β-mixing process with compactly supported marginal distribution μ. Then, by Theorem 1 in [71],

$$\sqrt{n}(\hat{\mu}_n - \mu) \stackrel{d}{\to} G \quad \text{in } \ell^\infty(\mathcal{F}_\sigma),$$

where $\mathcal{F}_\sigma$ is the function class given in (A.15) in the supplementary material with $s = \max\{\lfloor d/2 \rfloor + 1, 2\}$ and sufficiently large $\sigma > 0$, while G is a tight centered Gaussian process in $\ell^\infty(\mathcal{F}_\sigma)$ with covariance function $\mathrm{Cov}(G(f), G(g)) = \sum_{t\in\mathbb{Z}} \mathrm{Cov}(f(X_0), g(X_t))$. Conclude from the proof of Theorem 7 that

$$\sqrt{n}(\mathsf{S}(\hat{\mu}_n, \nu) - \mathsf{S}(\mu, \nu)) \stackrel{d}{\to} G(\varphi) \sim N\Big(0, \sum\nolimits_{t\in\mathbb{Z}} \mathrm{Cov}(\varphi(X_0), \varphi(X_t))\Big).$$

Likewise, a CLT result holds for other forms of dependent data, such as exchangeable arrays [56]. Furthermore, for the EOT case, the corresponding Hadamard derivative is linear, so suitable dependent bootstrap methods, such as the block bootstrap for mixing data [31] and the (extended) pigeonhole bootstrap for exchangeable arrays [56], are consistent for the empirical EOT cost, provided that the bootstrap processes satisfy a uniform CLT for $\mathcal{F}_\sigma$.

2.7 Concluding Remarks

This work developed a unified framework for proving limit distribution results for empirical regularized OT distances, semiparametric efficiency of the plug-in empirical estimator, and consistency of the bootstrap. As applications, we focused on three prominent OT regularization methods—smoothing, slicing, and entropic penalty—and provided a comprehensive statistical treatment thereof. We closed existing gaps in the literature (e.g., a limit distribution theory for sliced W_p) and provided several new results concerning empirical convergence rates, asymptotic efficiency, and bootstrap consistency. In particular, for the smooth Wasserstein distance, we explored compactly supported smoothing kernels, which were shown to inherit the structural and statistical properties of the well-studied Gaussian-smoothed framework. The analysis of compactly supported kernels is motivated by computational considerations, as we demonstrated how to lift the efficient algorithm from [197] for computing W_2^2 between smooth densities to the considered smooth OT distance.

Our framework is flexible and can treat a broad class of functionals, potentially well beyond the three examples considered herein. For instance, straightforward adaptations of our arguments for sliced W_p would yield limit distributions, efficiency, and bootstrap consistency of the projection-robust Wasserstein distance from [134], when the projected subspace is of dimension $k \leq 3$ (indeed, the class of projected OT potentials is still Donsker in that case). Going forward, we also plan to explore applicability of the unified framework to empirical OT maps or certain functionals thereof (e.g., inner product with a smooth test function).

Acknowledgments

The authors would like to thank Sloan Nietert for fruitful discussions on smooth Wasserstein distances with compactly supported kernels, and Tudor Manole for bringing their attention to Theorem 4 in [138]. They would also like to thank the anonymous associate editor and reviewers, whose comments helped up improve the paper.

CHAPTER 3

LIMIT LAWS FOR GROMOV-WASSERSTEIN ALIGNMENT WITH APPLICATIONS TO TESTING GRAPH ISOMORPHISMS

This chapter is based on the paper "Limit Laws for Gromov-Wasserstein Alignment with Applications to Testing Graph Isomorphisms" [165], submitted to the *Annals of Statistics*. This is joint work with Ziv Goldfeld and Kengo Kato.

3.1 Introduction

The Gromov-Wasserstein (GW) distance provides a framework, rooted in optimal transport (OT) theory [208, 175], for comparing probability distributions on possibly distinct spaces by aligning them with one another. The (p,q)-GW distance between two metric measure (mm) spaces $(\mathcal{X}_0, \mathsf{d}_0, \mu_0)$ and $(\mathcal{X}_1, \mathsf{d}_1, \mu_1)$ is defined as [141]

$$\mathsf{D}_{p,q}(\mu_0,\mu_1) \coloneqq \inf_{\pi\in\Pi(\mu_0,\mu_1)} \left(\int |\mathsf{d}_0^q(x,x') - \mathsf{d}_1^q(y,y')|^p d\pi\otimes\pi(x,y,x',y') \right)^{\frac{1}{p}}, \tag{3.1}$$

where $\Pi(\mu_0,\mu_1)$ is the set of all couplings of μ_0 and μ_1. Evidently, $\mathsf{D}_{p,q}(\mu_0,\mu_1)$ seeks to minimize distortion of distances over all possible alignment plans (modeled by couplings) of the considered mm spaces. This distance serves as an OT-based L^p relaxation of the classical Gromov-Hausdorff distance and defines a metric on the space of all mm spaces modulo the equivalence defined by isomorphism.[1] The GW framework has seen recent success in tasks involving alignment of heterogeneous datasets, such as single-cell genomics [23, 34, 67], alignment of language models [5], shape matching [125, 140], graph matching [43, 159, 217, 216], heterogeneous domain adaptation [180, 219], and generative modeling [32].

[1]Two mm spaces $(\mathcal{X}_0, \mathsf{d}_0, \mu_0)$ and $(\mathcal{X}_1, \mathsf{d}_1, \mu_1)$ are said to be isomorphic if there exists an isometry $T : \mathrm{spt}(\mu_0) \to \mathrm{spt}(\mu_1)$ for which $\mu_0 \circ T^{-1} = \mu_1$.

Despite its appealing structure and broad applications, a principled statistical and computational framework for GW alignment remained elusive since its inception more than a decade ago [141]. The difficulty in analyzing the GW problem lies in its quadratic dependence on π, leaving techniques developed for the (linear) OT problem not directly applicable. Recently, a variational form of the quadratic GW distance, i.e., with $p = q = 2$, between Euclidean mm spaces $(\mathbb{R}^{d_0}, \|\cdot\|, \mu_0)$ and $(\mathbb{R}^{d_1}, \|\cdot\|, \mu_1)$ was derived, casting it as the infimum over a class of OT problems [223]:

$$\mathsf{D}(\mu_0,\mu_1)^2 \coloneqq \mathsf{D}_{2,2}(\mu_0,\mu_1) = \mathsf{S}_1(\mu_0,\mu_1) + \inf_{\mathbf{A}\in\mathbb{R}^{d_0\times d_1}} \left\{32\|\mathbf{A}\|_{\mathrm{F}}^2 + \mathsf{OT}_{\mathbf{A}}(\mu_0,\mu_1)\right\}. \tag{3.2}$$

Here $\mathsf{S}_1(\mu_0,\mu_1)$ is a constant that depends only on the moments of the marginals and $\mathsf{OT}_{\mathbf{A}}(\mu_0,\mu_1)$ denotes the OT problem with a cost function parametrized by an auxiliary matrix $\mathbf{A}$. This connection to the well-understood OT problem unlocked it as a tool for the study of GW, leading to new algorithms for approximate computation with formal guarantees [164], sharp rates of empirical convergence [223, 110], and advances in GW gradient flows and differential geometry [222]. We seek to further progress the statistical GW theory by studying the asymptotic fluctuations of the empirical GW distance around the population value, providing first limit distribution results (upon suitable scaling) and efficient algorithms for simulating the limits. These advances are leveraged for applications to graph isomorphism testing.

3.1.1 Contributions

Utilizing the so-called *dual formulation* from (3.2), we first explore stability properties of the GW distance under perturbations of the marginal distributions. Combining this with the functional delta method, we establish the first limit theorems for GW distances at the scale of $\sqrt{n}$, under several settings of interest. Specifically, we adopt the methodology for deriving distributional limits developed in [106], which requires showing that the GW functional is right differentiable along certain directions and locally Lipschitz continuous in the sup-norm over a certain function class. This strategy is applied in three cases: (i) discrete GW, when both population distributions are finitely discrete, (ii) semi-discrete GW, when one population is finitely discrete and the other is absolutely continuous and compactly supported, and (iii) entropically regularized GW when the populations are arbitrary 4-sub-Weibull distributions. In all cases, local Lipschitz continuity follows from similar arguments whereas proving right differentiability comprises a more challenging task that requires different approaches tailored to each setting.

In the finitely discrete setting, we employ stability results for the optimal value function in nonlinear optimization [27] to prove the desired right differentiability. The main challenge in doing so stems from the general result requiring a directional regularity condition that does not hold for the GW setting. We circumvent this by exploiting the variational form (3.2) to directly establish strong duality for a linearization of the discrete GW problem. As the obtained derivative is nonlinear in general, the resulting limit distribution is not normal and the naïve bootstrap fails to be consistent (cf. [74, 79]). To address this, we propose an efficient method for directly estimating the limit distribution under the null hypothesis (i.e., the population distributions are isomorphic) and prove its consistency. Remarkably, sampling from the proposed estimator only requires solving a linear program, while computing the test statistic requires the resolution of a nonconvex (in fact, NP-complete) quadratic program. These results serve as the cornerstone of our proposed application: testing if two inhomogeneous random graph models are isomorphic.

In the semi-discrete setting, a central component of our analysis is to derive a simple sufficient condition for the optimal coupling of $\mathsf{OT}_{\mathbf{A}}(\mu_0,\mu_1)$ to be unique and induced by a deterministic map, uniformly in the choice of $\mathbf{A} \in \mathbb{R}^{d_0 \times d_1}$. Consequently, this same condition implies that the GW problem itself admits an alignment plan that is induced by a map, a so-called *Gromov-Monge map*. This adds to the limited number of cases where such alignment maps are known to exist (to wit [189, Theorem 9.21], [204, Theorem 4.1.2 and Proposition 4.2.4], and [75, Theorem 3.2]). Using this condition, the corresponding right derivative can be computed by appealing to the GW dual form, leading to a characterization of distributional limits for semi-discrete GW. As in the discrete case, the limiting law is generally not Gaussian, and the naïve bootstrap is inapplicable.

To address the setting when both distributions are non-discrete, we consider the entropically regularized GW distance between 4-sub-Weibull distributions. The analysis leverages the fact that optimal couplings and dual potentials are unique (in a suitable sense) for each of the entropic OT problems featuring in the dual form. The required stability can then be demonstrated by following a similar argument to the semi-discrete case. Building on the results of [164], we show that the derived limit distribution is normal once the regularization parameter surpasses a given threshold that depends on the marginal moments, and may fail to be normal otherwise. When asymptotic normality holds, the bootstrap enables consistent estimation of the distributional limit.

We apply the derived limit distribution theory to the problem of two-sample testing for graph isomorphism. As our theory account only for Euclidean mm spaces, we propose an embedding for distributions on random graphs with N vertices into the space of distributions on $\mathbb{R}^N$. Crucially, this embedding is constructed such that the embedded distributions are identified under the GW distance if and only if the associated distributions on graphs are isomorphic. This enables us to formulate a consistent test for the graph isomorphism question with the GW distance as the test statistic. To our knowledge, these are the first results in this setting (see literature review below). As the GW test statistic is hard to compute, we propose efficient procedures for estimating it numerically based on the linear programming-based sampler from the discrete GW limiting variable. This methodology is empirically validated in a number of numerical experiments. The experiments also showcase asymptotic normality of empirical entropic GW and consistency of the bootstrap under certain conditions.

3.1.2 Literature Review

The statistical properties of the empirical GW distance are largely unknown, beyond the following convergence rate result (see Theorem 3 [223]). If $(\mu_0, \mu_1) \in \mathcal{P}(\mathcal{X}_0) \times \mathcal{P}(\mathcal{X}_1)$ where $\mathcal{X}_0 \subset \mathbb{R}^{d_0}$ and $\mathcal{X}_1 \subset \mathbb{R}^{d_1}$ are compact, then, for $R = \max_{i \in \{0,1\}} \mathrm{diam}(\mathcal{X}_i)$ and $d = \min_{i \in \{0,1\}} d_i$,

$$\mathbb{E}\left[\left|\mathsf{D}(\hat{\mu}_{0,n}, \hat{\mu}_{1,n})^2 - \mathsf{D}(\mu_0, \mu_1)^2\right|\right] \lesssim_{d_0,d_1} R^4 n^{-\frac{1}{2}} + (1 + R^4) n^{-\frac{2}{\max\{4,d\}}} (\log n)^{\mathbb{1}_{\{d=4\}}},$$

where $\hat{\mu}_{0,n}, \hat{\mu}_{1,n}$ are the empirical measures of n i.i.d. observations from μ_0 and μ_1, respectively. These rates are sharp up to polylog factors and, notably, scale with the dimension of the *lower-dimensional* space. As for the entropically regularized GW distance, Theorem 2 in [223] establishes an $n^{-1/2}$ parametric empirical convergence rate, provided that μ_0 and μ_1 satisfy the aforementioned 4-sub-Weibull condition. That work also showed that entropic GW with parameter $\varepsilon > 0$ approximates the unregularized distance at the level of $O(\varepsilon \log(1/\varepsilon))$, thus approaching it as $\varepsilon \to 0$. In [110], the above empirical convergence rates were shown to adapt to the intrinsic dimension of the population. Formal guarantees for a neural estimator of the GW distance were provided in [211], accounting for empirical estimation and function approximation errors (but not the optimization error).

The existence of Gromov-Monge maps for the semi-discrete GW problem derived herein adds to a few other known cases for which deterministic schemes are optimal. In the Euclidean setting with $\mathsf{D}_{2,2}$, results are available when (i) the distributions are uniform on the same finite number of points [204, Theorem 4.1.2], (ii) μ_0 and μ_1 are absolutely continuous and rotationally invariant about their barycenter [189, Theorem 9.21], and (iii) when the cross-correlation matrix of some optimal alignment plan for $\mathsf{D}(\mu_0, \mu_1)$ is of full rank among other conditions [204, Proposition 4.2.4]. Beyond the quadratic setting, the only other result concerns the inner product GW distance (i.e., when the cost function in (3.1) is $|\langle x, x' \rangle - \langle y, y' \rangle|^2$, with $p = 2$) between populations μ_0, μ_1 that are compactly supported in their respective, with the lower-dimensional one being absolutely continuous.

Our approach to proving limit theorems for GW distances is based on the generic framework put forth in [106] which, in turn, relies on the functional delta method (see Section B.4 for details). The delta method [169, 181] has seen great success for proving limit theorems for Wasserstein distances and regularized OT. Notably, [187] first applied this technique to establish limit theorems for the p-Wasserstein distance on finite metric spaces. Next, [190] extended this approach to the case of countable metric spaces, and [117] provided a generic stability result for the OT cost and provided sufficient conditions for the CLT to hold. This approach also proved effective in studying the distributional limits of regularized optimal transport functionals—under smoothing, slicing, or entropic regularization—and associated objects (e.g., maps and dual potentials) [104, 105, 106].

Testing *equality* of distributions on graphs was first addressed in [109] based on the maximum mean discrepancy. While the experiment in that work also considered equality up to isomorphisms, their theoretical results did not cover that setting. Another approach relies on testing if certain graph statistics, such as the mean graph Laplacian, are equal at the population level based on samples [98]. Other lines of work include the setting where the number of samples grows with the graph size [39, 42] or when the graphs are large and only a small constant number of samples from each distribution is available [93, 92, 94, 119, 133, 191, 192]. The scopes and/or settings of those references differ from the statistical framework and accompanying theory we provide for graph isomorphism testing.

3.2 Background and Preliminary Results

3.2.1 Notation

For a nonempty Borel set $S \subset \mathbb{R}^d$, $\mathcal{P}(S)$ is the set of Borel probability measures on S, $C(S)$ denotes the set of continuous functions on S, and $\ell^\infty(S)$ is the Banach space of bounded real functions on S equipped with the supremum norm $\|\ell\|_{\infty,S} = \sup_{s\in S} |\ell(s)|$. When $S \subset \mathbb{R}^d$ is a compact set, we write $\|S\|_\infty = \sup_{s\in S} \|s\|$, where the latter is the Euclidean norm.

For a probability measure $\rho \in \mathcal{P}(\mathbb{R}^d)$, set $\mathcal{P}_\rho := \left\{\nu \in \mathcal{P}(\mathbb{R}^d) : \mathrm{spt}(\nu) \subset \mathrm{spt}(\rho)\right\}$. If $T : \mathbb{R}^d \to \mathbb{R}^{d'}$ is a measurable map, $T_\sharp\rho \in \mathcal{P}(\mathbb{R}^{d'})$ is the pushforward measure defined by $T_\sharp\rho(A) = \rho(T^{-1}(A))$ for every Borel measurable set $A \subset \mathbb{R}^d$. The centered version of ρ is denoted by $\bar{\rho} = (\mathrm{Id} - \mathbb{E}_\rho[X])_\sharp\rho$. The p-th moment of ρ is denoted $M_p(\rho) = \int \|\cdot\|^p d\rho$, $\mathbf{\Sigma}_\rho = \int zz^\intercal d\rho(z)$ is its cross-correlation matrix, and for a ρ-integrable function $f : \mathbb{R}^d \to \mathbb{R}$, we write $\rho(f) = \int f d\rho$. A probability distribution $\rho \in \mathcal{P}(\mathbb{R}^d)$ is said to be β-sub-Weibull with parameter σ^2 for $\sigma \geq 0$, provided that $\int e^{\|\cdot\|^\beta/(2\sigma^2)} d\rho \leq 2$. We denote by $\mathcal{P}_{\beta,\sigma}(\mathbb{R}^d)$ the set of all β-sub-Weibull distributions with parameter σ^2. Distributions that are 2-sub-Weibull are called sub-

Gaussian, while $X \sim \mu$ being 4-sub-Weibull is equivalent to sub-Gaussianity of $\|X\|^2$. For measures $(\rho_0, \rho_1) \in \mathcal{P}(\mathbb{R}^{d_0}) \times \mathcal{P}(\mathbb{R}^{d_1})$, $\rho_0 \otimes \rho_1$ denotes the product measure. Convergence in distribution and weak convergence are denoted, respectively, by $\xrightarrow{d}$ and $\xrightarrow{w}$.

For matrices $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{d_0 \times d_1}$ and $(i, j) \in [N_0] \times [N_1]$, $\mathbf{A}_{ij}$ is the ij-th entry of $\mathbf{A}$, and $\langle \mathbf{A}, \mathbf{B} \rangle_{\mathrm{F}} = \sum_{(i,j)\in[d_0]\times[d_1]} \mathbf{A}_{ij}\mathbf{B}_{ij}$ is the Frobenius inner product. The induced Frobenius norm is denoted by $\|\cdot\|_{\mathrm{F}}$, with $B_{\mathrm{F}}(r) \subset \mathbb{R}^{d_0 \times d_1}$ designating the closed Frobenius ball with radius $r > 0$. A function $f : \mathbb{R}^{d_0 \times d_1} \to \mathbb{R}$ is said to be Fréchet differentiable at a point $\mathbf{A} \in \mathbb{R}^{d_0 \times d_1}$ if $\lim_{\|\mathbf{H}\|_{\mathrm{F}} \downarrow 0} \frac{|f(\mathbf{A}+\mathbf{H}) - f(\mathbf{A}) - Df_{[\mathbf{A}]}(\mathbf{H})|}{\|\mathbf{H}\|_{\mathrm{F}}} = 0$ for some bounded linear functional $Df_{[\mathbf{A}]} : \mathbb{R}^{d_0 \times d_1} \to \mathbb{R}$ which we call the Fréchet derivative of f at $\mathbf{A}$.

For functions $f_0 : \mathbb{R}^{d_0} \to \mathbb{R}$ and $f_1 : \mathbb{R}^{d_1} \to \mathbb{R}$, their direct sum is $f_0 \oplus f_1 : (x, y) \in \mathbb{R}^{d_0} \times \mathbb{R}^{d_1} \mapsto f_0(x) + f_1(y)$. For any nonnegative integer k and nonempty open set $S \subset \mathbb{R}^d$, $C^k(S)$ is the set of k-times continuously differentiable functions on S. For $N \in \mathbb{N}$, $[N]$ denotes the set $\{1, \dots, N\}$, and $\mathbb{1}_N \in \mathbb{R}^N$ is the vector of all ones.

3.2.2 Optimal Transport

For $(\mu_0, \mu_1) \in \mathcal{P}(\mathbb{R}^{d_0}) \times \mathcal{P}(\mathbb{R}^{d_1})$, and a continuous cost function $c : \mathcal{X}_0 \times \mathcal{X}_1 \to \mathbb{R}$ satisfying $c \geq a \oplus b$ for some continuous functions $(a, b) \in L^1(\mu_0) \times L^1(\mu_1)$, the optimal transport (OT) problem is given by

$$\mathsf{OT}_c(\mu_0, \mu_1) = \inf_{\pi \in \Pi(\mu_0, \mu_1)} \int c d\pi,$$

where $\Pi(\mu_0, \mu_1)$ is the set of all couplings of (μ_0, μ_1). If, in addition, $\int c d\mu_0 \otimes \nu_0 < \infty$, the following duality result holds (cf. e.g. Theorem 5.10 and Remark 5.14 in [208])

$$\mathsf{OT}_c(\mu_0, \mu_1) = \sup_{\substack{(\varphi_0, \varphi_1) \in C(\mathcal{X}_0) \times C(\mathcal{X}_1) \\ \varphi_0 \oplus \varphi_1 \leq c}} \left\{ \int \varphi_0 d\mu_0 + \int \varphi_1 d\mu_1 \right\}. \tag{3.3}$$

Both the primal (minimization) problem and the dual (maximization) problems admit solutions, which we refer to, respectively, as an *OT plan* and *OT potentials* for $\mathsf{OT}_c(\mu_0, \mu_1)$. Of note is that if $(\varphi_0^\star, \varphi_1^\star)$ is a pair of OT potentials for $\mathsf{OT}_c(\mu_0, \mu_1)$, then so is $(\varphi_0^\star + a, \varphi_1^\star - a)$ for any $a \in \mathbb{R}$; we refer to potentials related in this fashion as *versions* of one another. If $\pi = (\mathrm{Id}, T)_\sharp \mu_0$ for a map $T : \mathbb{R}^{d_0} \to \mathbb{R}^{d_1}$, we say that π is induced by the map T and, if π is optimal for $\mathsf{OT}_c(\mu_0, \mu_1)$, T is called an *OT map* (also known as a Monge/Brenier map).

3.2.3 Entropic Optimal Transport

The entropic optimal transport (EOT) problem was introduced in [53] as a convexification of the classical setting. It is defined by regularizing the transportation cost with the (strongly convex) Kullback-Leibler divergence:

$$
\begin{aligned}
\mathsf{OT}_{c,\varepsilon}(\mu_0, \mu_1) &= \inf_{\pi \in \Pi(\mu_0, \mu_1)} \int c d\pi + \varepsilon \mathsf{D}_{\mathrm{KL}}(\pi \| \mu_0 \otimes \mu_1), \text{ where} \\
\mathsf{D}_{\mathrm{KL}}(\alpha \| \beta) &= \begin{cases} \int \log\left(\frac{d\alpha}{d\beta}\right) d\alpha, & \text{if } \alpha \ll \beta, \\ +\infty, & \text{otherwise.} \end{cases}
\end{aligned} \tag{3.4}
$$

The strong convexity of the regularized objective endows (3.4) with many useful properties. First, if $c \in L^1(\mu_0 \otimes \mu_1)$, (3.4) is paired in strong duality with the problem

$$
\sup_{(\varphi_0, \varphi_1) \in L^1(\mu_0) \times L^1(\mu_1)} \left\{ \int \varphi_0 d\mu_0 + \int \varphi_1 d\mu_1 - \varepsilon \int e^{\frac{\varphi_0 \oplus \varphi_1 - c}{\varepsilon}} d\mu_0 \otimes \mu_1 + \varepsilon \right\}. \tag{3.5}
$$

A solution, $(\varphi_0^\star, \varphi_1^\star)$, to (3.5) is called a pair of *EOT potentials* for $\mathsf{OT}_{c,\varepsilon}(\mu_0, \mu_1)$ and is known to be a.s. unique up to additive constants (i.e. any other solution $(\psi_0^\star, \psi_1^\star)$ satisfies $(\psi_0^\star, \psi_1^\star) = (\varphi_0^\star + a, \varphi_1^\star - a)$ $\mu_0 \otimes \mu_1$-a.s. for some $a \in \mathbb{R}$). Moreover, (3.4) has a unique solution $\pi^\star \in \Pi(\mu_0, \mu_1)$, given in terms of the marginal measures and the EOT potentials via

$$
\frac{d\pi^\star}{d\mu_0 \otimes \mu_1} = e^{\frac{\varphi_0^\star \oplus \varphi_1^\star - c}{\varepsilon}}, \quad \mu_0 \otimes \mu_1\text{-a.s.}
$$

This further implies that $(\varphi_0, \varphi_1) \in L^1(\mu_0) \times L^1(\mu_1)$ solves (3.5) if and only if (φ_0, φ_1) solves the Schrödinger system

$$\int e^{\frac{\varphi_0(x)+\varphi_1-c(x,\cdot)}{\varepsilon}} d\mu_1 = 1, \int e^{\frac{\varphi_0+\varphi_1(y)-c(\cdot,y)}{\varepsilon}} d\mu_0 = 1, \tag{3.6}$$

for μ_0-a.e. $x \in \mathbb{R}^{d_0}$ and μ_1-a.e. $y \in \mathbb{R}^{d_1}$. See [155] for a comprehensive introduction to EOT.

3.2.4 Gromov-Wasserstein Distances

Throughout, we focus on the (squared) $(2,2)$-GW distance between the Euclidean mm spaces $(\mathbb{R}^{d_0}, \|\cdot\|, \mu_0)$ and $(\mathbb{R}^{d_1}, \|\cdot\|, \mu_1)$, henceforth identified with the probability measures themselves. The distance between $\mu_0 \in \mathcal{P}(\mathbb{R}^{d_0}), \mu_1 \in \mathcal{P}(\mathbb{R}^{d_1})$ with finite fourth moments is

$$\mathsf{D}(\mu_0, \mu_1) = \inf_{\pi \in \Pi(\mu_0,\mu_1)} \|\Delta\|_{L^2(\pi\otimes\pi)}, \quad \text{for} \quad \Delta(x, y, x', y') := \left| \|x - x'\|^2 - \|y - y'\|^2 \right|. \tag{3.7}$$

A coupling $\pi \in \Pi(\mu_0, \mu_1)$ solving (3.7) always exists [141, Corollary 10.1] and is called a *GW plan*. If a GW plan is induced by a measurable map, $T : \mathbb{R}^{d_0} \to \mathbb{R}^{d_1}$, we call T a Gromov-Monge map. Of note is that D nullifies if and only if there exists an isometry $\iota : \mathrm{spt}(\mu_0) \to \mathrm{spt}(\mu_1)$ for which $\mu_1 = \iota_\sharp\mu_0$. When such a relationship holds, we say that the mm spaces $(\mathbb{R}^{d_0}, \|\cdot\|, \mu_0)$ and $(\mathbb{R}^{d_1}, \|\cdot\|, \mu_1)$ are isomorphic. Furthermore, D defines a metric on the quotient space of all Euclidean mm spaces whose measures have finite fourth moments modulo the aforementioned isomorphism relation.

Recalling that $\bar{\mu}_i = (\mathrm{Id} - \mathbb{E}_{\mu_i}[X])_\sharp \mu_i$, for $i = 0, 1$, denotes the centered measures, it was recently shown in [223] (see Corollary 1 in that work) that D admits the variational form,

$$\mathsf{D}(\mu_0, \mu_1)^2 = \mathsf{D}(\bar{\mu}_0, \bar{\mu}_1)^2 = \mathsf{S}_1(\bar{\mu}_0, \bar{\mu}_1) + \mathsf{S}_2(\bar{\mu}_0, \bar{\mu}_1) \tag{3.8}$$

$$\mathsf{S}_1(\mu_0, \mu_1) = \int \|x - x'\|^4 d\mu_0 \otimes \mu_0(x, x') + \int \|y - y'\|^4 d\mu_1 \otimes \mu_1(y, y') - 4M_2(\mu_0)M_2(\mu_1)$$

$$\mathsf{S}_2(\mu_0, \mu_1) = \inf_{\mathbf{A} \in \mathbb{R}^{d_0\times d_1}} \left\{ 32\|\mathbf{A}\|_{\mathrm{F}}^2 + \mathsf{OT}_{\mathbf{A}}(\mu_0, \mu_1) \right\},$$

where $\mathsf{OT}_{\mathbf{A}}(\mu_0,\mu_1) \coloneqq \mathsf{OT}_{c_{\mathbf{A}}}(\mu_0,\mu_1)$ for $c_{\mathbf{A}} : (x,y) \mapsto -4\|x\|^2\|y\|^2 - 32x^\intercal \mathbf{A} y$ and the equality $\mathsf{D}(\mu_0,\mu_1)^2 = \mathsf{D}(\bar{\mu}_0,\bar{\mu}_1)^2$ is due to translation invariance. In particular, S_1 is an explicit constant whereas S_2 is a minimization problem with objective function,

$$\Phi_{(\mu_0,\mu_1)} : \mathbf{A} \in \mathbb{R}^{d_0\times d_1} \mapsto 32\|\mathbf{A}\|_{\mathrm{F}}^2 + \mathsf{OT}_{\mathbf{A}}(\mu_0,\mu_1). \tag{3.9}$$

Further, if $\pi^\star$ solves (3.7), then $\mathbf{A}^\star = \frac{1}{2}\int xy^\intercal d\pi^\star(x,y)$ solves (3.9) provided that μ_0,μ_1 are centered (see [223, Theorem 1]). Our first result refines the characterization of minimizers of (3.9).

Theorem 8 (On minimizers of (3.9))**.** Let $(\mu_0,\mu_1) \in \mathcal{P}(\mathbb{R}^{d_0}) \times \mathcal{P}(\mathbb{R}^{d_1})$ be compactly supported. The following statements hold:

1. $\Phi_{(\mu_0,\mu_1)}$ is locally Lipschitz continuous and coercive. If all OT plans, $\pi_{\mathbf{A}}$, for $\mathsf{OT}_{\mathbf{A}}(\mu_0,\mu_1)$ admit the same cross-correlation matrix, $\int xy^\intercal d\pi_{\mathbf{A}}(x,y)$, then $\Phi_{(\mu_0,\mu_1)}$ is Fréchet differentiable at $\mathbf{A} \in \mathbb{R}^{d_0\times d_1}$ with $\left(D\Phi_{(\mu_0,\mu_1)}\right)_{[\mathbf{A}]}(\mathbf{B}) = 64\langle \mathbf{A} - \frac{1}{2}\int xy^\intercal d\pi_{\mathbf{A}}(x,y), \mathbf{B}\rangle_{\mathrm{F}}$.
2. If $\mathbf{A}^\star$ minimizes (3.9), then $2\mathbf{A}^\star = \int xy^\intercal d\pi^\star(x,y) \in B_{\mathrm{F}}(\sqrt{M_2(\mu_0)M_2(\mu_1)})$ for some OT plan $\pi^\star$ for $\mathsf{OT}_{\mathbf{A}^\star}(\mu_0,\mu_1)$. If μ_0,μ_1 are centered, then $\pi^\star$ solves (3.7).

Theorem 8, whose proof is provided in Section 3.8.1, shows that all minimizers of (3.9) are contained in $B_{\mathrm{F}}(\frac{1}{2}\sqrt{M_2(\mu_0)M_2(\mu_1)})$ and, given such a minimizer, a solution to (3.7) can be obtained. Although Point 1 appears to directly imply Point 2, since a global minimizer of a differentiable coercive function must be a critical point, we stress that if $\mathsf{OT}_{\mathbf{A}}(\mu_0,\mu_1)$ admits multiple OT plans, $\Phi_{(\mu_0,\mu_1)}$ may fail to be differentiable at $\mathbf{A}$. To formalize this fact, the proof of Theorem 8 characterizes the Clarke subdifferential [49] of $\Phi_{(\mu_0,\mu_1)}$.

Remark 12 (Uniqueness of cross-correlation matrix)**.** The condition that all OT plans for $\mathsf{OT}_{\mathbf{A}}(\mu_0,\mu_1)$ share the same cross-correlation matrix seems nontrivial to verify in general and, in fact, fails in some cases. For instance, example given in Figure 1 of [223] indicates that $\Phi_{(\mu_0,\mu_1)}$ is nonsmooth for simple discrete measures on the line. With this, it appears reasonable to impose the stronger condition that, for every $\mathbf{A} \in \mathbb{R}^{d_0 \times d_1}$, $\mathsf{OT}_{\mathbf{A}}(\mu_0, \mu_1)$ admits a unique optimal plan. However, standard results guaranteeing uniqueness of OT plans generally require injectivity of $\nabla_x c(x, \cdot)$, known as the twist condition, (see e.g. Theorem 10.28 in [208]) among other assumptions. As $\nabla_x c_{\mathbf{A}}(x, y) = -8x\|y\|^2 - 32\mathbf{A}y$, this condition can fail even in anodyne situations (e.g. $0 \in \mathrm{int}(\mathrm{spt}(\mu_0))$ and there exists $y, y' \in \mathrm{spt}(\mu_1) \cap \ker(\mathbf{A})$). The failure of this condition constitutes a roadblock to proving the existence of Gromov-Monge alignment maps for (3.7). Nevertheless, we demonstrate in Section 3.4 that, in the semi-discrete case, a simple condition guarantees uniqueness of OT plans and existence of OT maps for $\mathsf{OT}_{\mathbf{A}}(\mu_0, \mu_1)$. This enables us to prove existence of Gromov-Monge maps for the semi-discrete GW problem, which adds to the few known cases for which optimal alignment maps exist; [189, Theorem 9.21] and [204, Proposition 4.2.4].

The entropic GW distance between the Euclidean mm spaces $(\mathbb{R}^{d_0}, \|\cdot\|, \mu_0)$ and $(\mathbb{R}^{d_1}, \|\cdot\|, \mu_1)$ is obtained by regularizing (3.7) using the Kullback-Leibler divergence, analogously to EOT. Fixing $\varepsilon > 0$, the distance between $\mu_0 \in \mathcal{P}(\mathbb{R}^{d_0}), \mu_1 \in \mathcal{P}(\mathbb{R}^{d_1})$ with finite fourth moments is given by

$$\mathsf{D}_\varepsilon(\mu_0, \mu_1)^2 = \inf_{\pi \in \Pi(\mu_0,\mu_1)} \left\{ \|\Delta\|^2_{L^2(\pi \otimes \pi)} + \varepsilon \mathsf{D}_{\mathrm{KL}}(\pi \| \mu_0 \otimes \mu_1) \right\}. \tag{3.10}$$

This modification was introduced in [160, 186] due to computational considerations, resulting in a heuristic iterative algorithm which involves solving an EOT problem at each iterate. More recently, a gradient-based algorithm for solving the entropic GW problem subject to non-asymptotic convergence guarantees was provided in [164] by leveraging Theorem 23 ahead. We underscore that $\mathsf{D}_\varepsilon(\mu_0, \mu_0) \neq 0$ unless μ_0 is a Dirac measure[2]

[2] $\mathsf{D}_{\mathrm{KL}}(\pi \| \mu_0 \otimes \mu_0) \geq 0$ with equality if and only if $\pi = \mu_0 \otimes \mu_0$ whereas $\int \Delta d\pi \otimes \pi \geq 0$ with equality if and only if π is induced by an isomorphism.

so that D_ε does not define a metric. Nevertheless, D_ε is still invariant to isometric transformations of the marginal spaces and always admits a solution.[3] By analogy with (3.8), the entropic GW problem also admits a variational form (obtained by replacing $\mathsf{OT}_{\mathbf{A}}$ by $\mathsf{OT}_{\mathbf{A},\varepsilon} := \mathsf{OT}_{c_{\mathbf{A}},\varepsilon}$ in (3.8)) and Theorem 8 carries over to the entropic case with the advantage that the measures need only be 4-sub-Weibull and that the objective $\Phi_{(\mu_0,\mu_1)}$ is everywhere Fréchet differentiable; see Section B.1.

3.3 Discrete Gromov-Wasserstein Distance

We first consider the GW distance between finitely discrete measures $\mu_0 \in \mathcal{P}(\mathbb{R}^{d_0})$ and $\mu_1 \in \mathcal{P}(\mathbb{R}^{d_1})$. Throughout, we let $\mathcal{X}_0 = (x^{(i)})_{i=1}^{N_0} = \mathrm{spt}(\mu_0)$ and $\mathcal{X}_1 = (y^{(j)})_{j=1}^{N_1} = \mathrm{spt}(\mu_1)$. Our analysis hinges on the variational formulation of the GW distance and the fact that $\mathsf{D}(\mu_0,\mu_1)^2$ can be identified with the finite-dimensional quadratic program

$$\begin{aligned} \inf_{\mathbf{P}\in\mathbb{R}^{N_0\times N_1}} \ & \langle \mathbf{P}, \Delta(\mathbf{P})\rangle_{\mathrm{F}}, \\ \text{s.t.} \ & \mathbf{P}\mathbb{1}_{N_1} = m_0, \mathbf{P}^\intercal \mathbb{1}_{N_0} = m_1, \\ & \mathbf{P}_{ij} \geq 0, \ \forall (i,j) \in [N_0]\times[N_1]. \end{aligned} \tag{3.11}$$

where $\Delta(\mathbf{P}) \in \mathbb{R}^{N_0\times N_1}$ is the matrix with ij-th entry

$$(\Delta(\mathbf{P}))_{ij} = \sum_{(k,l)\in[N_0]\times[N_1]} \Delta\left(x^{(i)}, y^{(j)}, x^{(k)}, y^{(l)}\right)\mathbf{P}_{kl},$$

and $(m_0, m_1) \in \mathbb{R}^{N_0}\times\mathbb{R}^{N_1}$ are the weight vectors for μ_0 and μ_1, respectively, i.e., $(m_0)_i = \mu_0(\{x^{(i)}\})$ and $(m_1)_j = \mu_1(\{y^{(j)}\})$ for $(i,j) \in [N_0]\times[N_1]$. Evidently, any $\mathbf{P} \in \mathbb{R}^{d_0\times d_1}$ satisfying the constraints in (3.11) can be identified with a coupling $\pi_{\mathbf{P}} \in \Pi(\mu_0,\mu_1)$ via $\pi_{\mathbf{P}}\left(\left\{(x^{(i)}, y^{(j)})\right\}\right) = \mathbf{P}_{ij}$ and it is clear that $\langle \mathbf{P}, \Delta(\mathbf{P})\rangle_{\mathrm{F}} = \int \Delta d\pi_{\mathbf{P}}\otimes\pi_{\mathbf{P}}$.

[3]The proof of Corollary 10.1 in [141] shows that the functional $\varpi \in \mathcal{P}(\mathbb{R}^{d_0}\times\mathbb{R}^{d_1}) \mapsto \int \Delta d\varpi\otimes\varpi$ is continuous in the topology of weak convergence of probability measures. The conclusion follows by noting that $\varpi \mapsto \mathsf{D}_{\mathrm{KL}}(\varpi\|\mu_0\otimes\mu_1)$ is lower semicontinuous and that $\Pi(\mu_0,\mu_1)$ is compact in this topology.

3.3.1 Stability analysis

We explore the stability of the discrete GW functional along perturbations of the marginal distributions. Combining this with the methodology from [106] will then result in distributional limits for empirical GW in the discrete setting. In order to preserve the structure of (3.11) along the perturbations

$$\mu_{0,t} := \mu_0 + t(\nu_0 - \mu_0), \quad \mu_{1,t} := \mu_1 + t(\nu_1 - \mu_1), \quad t \in [0,1]$$

appearing in Proposition 16 ahead, we require that $\nu_i \in \mathcal{P}(\mathcal{X}_i)$, for $i \in \{0,1\}$, is such that $\mathsf{D}(\mu_{0,t}, \mu_{1,t})^2$ can be written as (3.11) with the corresponding weight vectors in place of m_0, m_1. For any $(\nu_0, \nu_1) \in \mathcal{P}(\mathcal{X}_0) \times \mathcal{P}(\mathcal{X}_1)$, we have $\frac{1}{2}\sqrt{M_2(\nu_0)M_2(\nu_1)} \le \frac{1}{2}\|\mathcal{X}_0\|_\infty \|\mathcal{X}_1\|_\infty$, so that all minimizers of $\Phi_{(\nu_0,\nu_1)}$ are contained in $B_{\mathrm{F}}(M)$ for $M := \frac{1}{2}\|\mathcal{X}_0\|_\infty\|\mathcal{X}_1\|_\infty$. Moreover, $M_2(\bar{\nu}_i) = M_2(\nu_i) - \|\mathbb{E}_{\nu_i}[X]\|^2 \le M_2(\nu_i)$, for $i \in \{0,1\}$, so that the minimizers of $\Phi_{(\bar{\nu}_0,\bar{\nu}_1)}$ are also contained in $B_{\mathrm{F}}(M)$.

Stability properties of $\mathsf{D}(\mu_0,\mu_1)^2$ along valid perturbations of the marginals are established by studying a type of linearization of this quadratic programs which is related to $\mathsf{OT}_{\mathbf{A}}(\mu_0,\mu_1)$, as defined in Section 3.2.4, for some choice of $\mathbf{A} \in \mathbb{R}^{d_0 \times d_1}$. To state the subsequent stability property, set

$$\begin{aligned}
\mathcal{F}_i &:= \{f : \mathcal{X}_i \to \mathbb{R} : \|f\|_{\infty,\mathcal{X}_i} \le 1\}, \quad i \in \{0,1\},\\
\mathcal{F}^{\oplus} &:= \{f_0 \oplus f_1 : f_0 \in \mathcal{F}_0, f_1 \in \mathcal{F}_1\}.
\end{aligned}$$

Theorem 9 (Stability for discrete GW)**.** Let $\mathcal{A} = \operatorname{argmin}_{B_{\mathrm{F}}(M)} \Phi_{(\bar{\mu}_0,\bar{\mu}_1)}$. Then, for any pairs $(\nu_0,\nu_1),(\rho_0,\rho_1) \in \mathcal{P}_{\mu_0} \times \mathcal{P}_{\mu_1}$, we have $\left|\mathsf{D}(\nu_0,\nu_1)^2 - \mathsf{D}(\rho_0,\rho_1)^2\right| \le C\|\nu_0 \otimes \nu_1 - \rho_0 \otimes \rho_1\|_{\infty,\mathcal{F}^{\oplus}}$, for some constant $C > 0$ which is independent of $(\nu_0,\nu_1),(\rho_0,\rho_1)$. Further,

$$\begin{aligned}
\frac{d}{dt}\mathsf{D}(\mu_{0,t},\mu_{1,t})^2\Big|_{t=0} = \inf_{\mathbf{A}\in\mathcal{A}} \inf_{\pi \in \bar{\Pi}^\star_{\mathbf{A}}} \sup_{(\varphi_0,\varphi_1)\in\bar{\mathcal{D}}_{\mathbf{A}}} \bigg\{ &\int (f_0 + g_{0,\pi} + \bar{\varphi}_0) d(\nu_0 - \mu_0)\\
&+ \int (f_1 + g_{1,\pi} + \bar{\varphi}_1) d(\nu_1 - \mu_1)\bigg\},
\end{aligned}$$

where, for $x \in \mathbb{R}^{d_0}$,

$$f_0(x) = 2\int \|x - x'\|^4 d\mu_0(x') - 4M_2(\bar{\mu}_1)\|x - \mathbb{E}_{\mu_0}[X]\|^2, g_{0,\pi}(x) = 8x^\intercal \mathbb{E}_{(X,Y)\sim\pi}[\|Y\|^2 X],$$

with $f_1, g_{1,\pi}$ defined symmetrically by interchanging $x \in \mathbb{R}^{d_0}$ and $y \in \mathbb{R}^{d_1}$, X and Y, and μ_0 and μ_1. Further, $\bar{\varphi}_i = \varphi_i(\cdot - \mathbb{E}_{\mu_i}[X])$ for $i \in \{0, 1\}$, and, for $\mathbf{A} \in B_{\mathrm{F}}(M)$, $\bar{\Pi}^\star_{\mathbf{A}} \subset \Pi(\bar{\mu}_0, \bar{\mu}_1)$ is the set of all OT plans, $\pi^\star$, for $\mathsf{OT}_{\mathbf{A}}(\bar{\mu}_0, \bar{\mu}_1)$ satisfying $\mathbf{A} = \frac{1}{2}\int xy^\intercal d\pi^\star(x, y)$, while $\bar{\mathcal{D}}_{\mathbf{A}}$ is the set of all corresponding OT potentials.

The proof of Theorem 9 (see Section 3.8.2) employs a linearization of the GW quadratic program (3.11) along certain perturbations of its marginals. The approach builds on directional differentiability results for the optimal value function in nonlinear programming (see Section 4.3.2 in [27]) while utilizing stability properties of solutions to perturbed (finite-dimensional) quadratic programs [124]. We note that the results from [27] do not apply directly to the problem at hand as the directional regularity condition, used to enforce strong duality for the linearized problem, does not hold due to the structure of the GW coupling set (see Remark 19). Nevertheless, we show that the dual variables correspond to OT potentials for $\mathsf{OT}_{\mathbf{A}}(\bar{\mu}_0, \bar{\mu}_1)$, where $\mathbf{A} = \frac{1}{2}\int xy^\intercal d\pi(x, y)$ for some GW plan π, and use the fact that dual OT potentials can always be replaced with a bounded version thereof without changing the objective to restrict optimization to a bounded constraint set.

Remark 13 (Comparison with OT stability)**.** To compare GW stability to that of OT, let $(\mathcal{X}, \mathsf{d}_{\mathcal{X}})$ be a finite metric space and set $\mu_0, \mu_1 \in \mathcal{P}(\mathcal{X})$ and $(\nu_0, \nu_1) \in \mathcal{P}_{\mu_0} \times \mathcal{P}_{\mu_1}$. Adopting the notation of Theorem 9, Theorem 3.1 in [85] yields

$$\frac{d}{dt}\mathsf{OT}_{\mathsf{d}^p_{\mathcal{X}}}(\mu_{0,t}, \mu_{1,t})\Big|_{t=0} = \sup_{(\varphi_0,\varphi_1)\in\mathcal{D}_p} \left\{\int \varphi_0 d(\nu_0 - \mu_0) + \int \varphi_1 d(\nu_1 - \mu_1)\right\},$$

where $\mathcal{D}_p$ is the set of OT potentials for $\mathsf{OT}_{\mathsf{d}^p_{\mathcal{X}}}(\mu_0, \mu_1)$. This observation was used in [187] to derive limit distributions for the p-Wasserstein distance on a finite space via the delta method. A similar term appears in the GW stability result (Theorem 9), along with (i) a minimization over $\mathcal{A}$, (ii) a minimization over corresponding OT plans, and (iii) the terms f_0 and f_1 which depend only on μ_0 and μ_1. The first occurs due to the minimization over matrices in the variational form, the second is due to the centering required to access the variational form, while the third accounts for the term S_1 in the variational form (3.8).

3.3.2 Distributional limits

The characterization of the stability profile of the discrete GW problem with respect to perturbations of the weights given in Theorem 9 is the driving force behind the subsequent limit theorem once the following assumption is made.[4]

Assumption 1. $(\mu_{0,n}, \mu_{1,n})_{n\in\mathbb{N}} \subset \mathcal{P}_{\mu_0} \times \mathcal{P}_{\mu_1}$ is such that $\sqrt{n}(\mu_{0,n} \otimes \mu_{1,n} - \mu_0 \otimes \mu_1) \xrightarrow{d} \chi_{\mu_0\otimes\mu_1}$ in $\ell^\infty(\mathcal{F}^\oplus)$ where $\chi_{\mu_0\otimes\mu_1}$ is tight.

The canonical statistical setting that fulfills Assumption 1 is when $\hat{\mu}_{0,n}$ and $\hat{\mu}_{1,n}$ are the empirical measures of i.i.d. observations $X_{0,1}, \dots, X_{0,n}$ from μ_0 and $X_{1,1}, \dots, X_{1,n}$ from μ_1, where the two collections of samples are independent and the function classes are Donsker.

Theorem 10 (Limit distributions for discrete GW)**.** In the setting of Theorem 9, if $(\mu_{0,n}, \mu_{1,n})_{n\in\mathbb{N}}$ satisfy Assumption 1, then

$$\begin{aligned}\sqrt{n}&\left(\mathsf{D}(\mu_{0,n}, \mu_{1,n})^2 - \mathsf{D}(\mu_0, \mu_1)^2\right) \\ &\xrightarrow{d} \inf_{\mathbf{A}\in\mathcal{A}} \inf_{\pi\in\bar{\Pi}^\star_{\mathbf{A}}} \sup_{(\varphi_0,\varphi_1)\in\bar{\mathcal{D}}_{\mathbf{A}}} \left\{\chi_{\mu_0\otimes\mu_1}((f_0 + g_{0,\pi} + \bar{\varphi}_0) \oplus (f_1 + g_{1,\pi} + \bar{\varphi}_1))\right\}.\end{aligned}$$

In particular, if $(\mu_{0,n}, \mu_{1,n}) = (\hat{\mu}_{0,n}, \hat{\mu}_{1,n})$, we have $\chi_{\mu_0\otimes\mu_1}(g_0 \oplus g_1) = G_{\mu_0}(g_0) + G_{\mu_0}(g_1)$ for any $g_0 \oplus g_1 \in \mathcal{F}^\oplus$, where G_{μ_i} is a tight μ_i-Brownian bridge in $\ell^\infty(\mathcal{F}_i)$ for $i \in \{0, 1\}$.

[4]Assumption 1 will be used when stating limit laws beyond the discrete case, the function class $\mathcal{F}^\oplus$ should be understood based on the context.

Given Assumption 1 and Theorem 9, Theorem 10 is a direct consequence of the delta method-based framework developed in [106, Proposition 1], included herein for completeness as Proposition 16 in Section B.4. Complete details are provided in Section 3.8.2.

Remark 14 (Extension of limits)**.** In Theorem 10, the distributional limit $\chi_{\mu_0\otimes\mu_1}$ is evaluated at $\xi_0 \oplus \xi_1$ for $\xi_0 := f_0 + g_{0,\pi} + \bar{\varphi}_0$ and $\xi_1 := f_1 + g_{1,\pi} + \bar{\varphi}_1$, respectively, for some choice of OT plan and pair of OT potentials. Though ξ_0 and ξ_1 are not necessarily elements of $\mathcal{F}_0$ and $\mathcal{F}_1$, it is easy to see that $\|\xi_0\|_{\infty,\mathcal{X}_0}$ and $\|\xi_1\|_{\infty,\mathcal{X}_1}$ can be bounded by a constant C which depends only on $\|\mathcal{X}_0\|_\infty$ and $\|\mathcal{X}_1\|_\infty$ (up to choosing the versions of the OT potentials which are bounded and noting that the direct sum is the same for all versions, see Lemma 12). It follows that $(C^{-1}\xi_0, C^{-1}\xi_1) \in \mathcal{F}_0 \times \mathcal{F}_1$ whereby $\sqrt{n}(\mu_{0,n} \otimes \mu_{1,n} - \mu_0 \otimes \mu_1)(\xi_0 \oplus \xi_1) = C\sqrt{n}(\mu_{0,n} \otimes \mu_{1,n} - \mu_0 \otimes \mu_1)(C^{-1}\xi_0 \oplus C^{-1}\xi_1) \xrightarrow{d} C\chi_{\mu_0\otimes\mu_1}(C^{-1}\xi_0 \oplus C^{-1}\xi_1)$. We thus identify $\chi_{\mu_0\otimes\mu_1}(\xi_0 \oplus \xi_1)$ with $C\chi_{\mu_0\otimes\mu_1}(C^{-1}\xi_0 \oplus C^{-1}\xi_1)$.

Remark 15 (Removing the square)**.** Although the limit distribution in Theorem 10 is presented for the squared distance D^2, the asymptotics for D can be obtained by a second application of the delta method with the function $\sqrt{(\cdot)}$, assuming that $\mathsf{D}(\mu_0, \mu_1) \neq 0$.

Remark 16 (Estimating the limit)**.** When the derivative established in Theorem 9 fails to be linear as a function of $(\nu_0 - \mu_0, \nu_1 - \mu_1)$, the limit distribution in Theorem 10 cannot be estimated via the naïve bootstrap [74]. To guarantee that the derivative be linear, it would suffice to show that $\mathcal{A} = \{\mathbf{A}^\star\}$ is a singleton, that the OT plan for $\mathsf{OT}_{\mathbf{A}^\star}(\bar{\mu}_0, \bar{\mu}_1)$ is unique, and that the corresponding OT potentials are unique (up to versions). In this discrete setting under the null, these conditions imply that μ_0 and μ_1 are Dirac delta distributions, rendering the limit trivial. We overcome this predicament by developing an efficient direct method for estimating the limit under the null using linear programming, described in Section 3.3.3 ahead.

Under the null hypothesis $\mathsf{D}(\mu_0, \mu_1) = 0$ (i.e. $\mu_0 = \mu_1$ up to isometries), the obtained limit distribution simplifies as follows.

Corollary 5 (Limit distribution under the null)**.** In the setting of Theorem 10, assume that μ_0 is not a point mass and that $\mathsf{D}(\mu_0, \mu_1) = 0$. If $\bar{\mathcal{T}}$ is the set of all GW maps between $\bar{\mu}_0$ and $\bar{\mu}_1$, then

$$\sqrt{n}\mathsf{D}(\mu_{0,n}, \mu_{1,n})^2 \xrightarrow{d} \inf_{T \in \bar{\mathcal{T}}} \sup_{h \in \mathcal{H}} \left\{ \chi_{\mu_0 \otimes \mu_1}(h(\cdot - \mathbb{E}_{\mu_0}[X]) \oplus -h(T^{-1}(\cdot - \mathbb{E}_{\mu_0}[X]))) \right\} =: L_{\mu_0},$$

$$\mathcal{H} = \{h : \mathrm{spt}(\bar{\mu}_0) \to \mathbb{R} \mid h(x) - h(x') \leq$$
$$2(\|x\|^2 - \|x'\|^2)^2 + 8(x - x')^\intercal \mathbf{\Sigma}_{\bar{\mu}_0}(x - x'), \forall x, x' \in \mathrm{spt}(\bar{\mu}_0) \right\}.$$

If $(\mu_{0,n}, \mu_{1,n}) = (\hat{\mu}_{0,n}, \hat{\mu}_{1,n})$, the above result holds and $L_{\mu_0} = \sqrt{2}\|G_{\mu_0}\|_{\infty, \bar{\mathcal{H}}}$, where $\bar{\mathcal{H}} = \{h(\cdot - \mathbb{E}_{\mu_0}[X]) : h \in \mathcal{H}\}$.

Corollary 5 follows from Theorem 10 upon characterizing the OT potentials under the null and simplifying the resulting expressions. Note that if $h \in \mathcal{H}$, so too is $h - a$, and the value of the direct sum in the limit distribution is invariant to translation. Thus, the direct sum can always be chosen such that it lies in $\mathcal{F}^{\oplus}$ up to a dilation by e.g. setting $h(x) = 0$ for some $x \in \mathrm{spt}(\bar{\mu}_0)$. See Section 3.8.2 for complete details.

3.3.3 Direct estimation of the limit

As the distribution L_{μ_0} in Corollary 5 admits a reasonably simple expression in the case of empirical measures, we may consider estimating it directly as an alternative to the bootstrap. To this end, let $(\bar{x}^{(i)})_{i=1}^{N_n} = \mathrm{spt}(\bar{\hat{\mu}}_{0,n})$ and set

$$\mathcal{H}_n = \{u \in \mathbb{R}^{N_n} \mid u_i - u_j \leq 2(\|\bar{x}^{(i)}\|^2 - \|\bar{x}^{(j)}\|^2)^2$$
$$+ 8(\bar{x}^{(i)} - \bar{x}^{(j)})^\intercal \mathbf{\Sigma}_{\bar{\hat{\mu}}_{0,n}}(\bar{x}^{(i)} - \bar{x}^{(j)}), i \neq j \in [N_n]\}.$$

Algorithm 1 Sampling from L_n

Given samples $X_1, \ldots, X_n$ from μ_0, construct $\hat{\mu}_{0,n}$,

1: Let $(\bar{x}^{(i)})_{i=1}^{N_n} = \mathrm{spt}(\bar{\hat{\mu}}_{0,n})$.

2: Let $b_n \in \mathbb{R}^{N_n(N_n-1)}$ and $\mathbf{A} \in \mathbb{R}^{N_n(N_n-1)\times N_n}$ be such that $\mathcal{H}_n = \{u \in \mathbb{R}^{N_n} : \mathbf{A}u \leq b_n\}$.

3: Sample $Z_n \sim N(0, \mathbf{\Sigma}_{m_n})$ and solve $\sqrt{2}\sup_{u\in\mathcal{H}_n} Z_n^\intercal u \sim L_n$; repeat Step 3 to generate new samples.

Letting m_0 denote the weight vector for μ_0, it is useful to set $u_f = (f(x^{(1)}), \ldots, f(x^{(N_0)}))$ for $f \in \mathcal{F}_0$ so that $\mu_0(f) = m_0^\intercal u_f$. With this, $G_{\mu_0}(f) = Z^\intercal u_f$ for $f \in \mathcal{F}_0$ and $Z \sim N(0, \mathbf{\Sigma}_{m_0})$, where, for a probability vector $r \in \mathbb{R}^N$, $\mathbf{\Sigma}_r \in \mathbb{R}^{N\times N}$ is the covariance matrix of the multinomial distribution with 1 trial and probability vector r, whose entries are given by

$$(\mathbf{\Sigma}_r)_{ij} = \begin{cases} r_i(1-r_i), & \text{if } i = j, \\ \\ -r_i r_j, & \text{if } i \neq j, \end{cases} \tag{3.12}$$

Taking $Z_n \sim N(0, \mathbf{\Sigma}_{m_n})$ and writing m_n for the weight vector associated with $\hat{\mu}_{0,n}$, the above suggests estimating L_{μ_0} as

$$L_n = \sqrt{2}\sup_{u\in\mathcal{H}_n} Z_n^\intercal u, \tag{3.13}$$

Algorithm 1 summarizes the procedure for sampling from L_n. It is worth mentioning that, as soon as $\hat{\mu}_{0,n}$ is supported on more than one point, $\mathcal{H}_n$ has non-empty interior (see Lemma 13), but is unbounded. However, the samples Z_n are almost surely orthogonal to the vector of ones, so that each instance of the linear program admits a bounded solution.[5]

We now establish that this procedure is consistent for estimating the distribution of L_{μ_0}.

Theorem 11. In the setting of Corollary 5 and Assumption 2,

$$\lim_{n\to\infty}\sup_{t\geq 0}\left|\mathbb{P}(L_n \leq t | X_1, X_2, \ldots, X_n) - \mathbb{P}(L_{\mu_0} \leq t)\right| = 0,$$

[5]In practice, it is recommended to add a box constraint on the first variable, $-\delta \leq h(x^{(1)}) \leq \delta$ for any $\delta > 0$, even if the entries of Z_n sum to one with probability one to account for numerical error. This modification results in a bounded feasible region, but does not effect the optimal value obtained, see the proof of Theorem 11.

for almost every realization of $X_1, X_2, \ldots$ provided that μ_0 is not a point mass.

The proof of Theorem 11, included in Section 3.8.2, leverages the stability properties of the optimal value function for linear programs when viewed as a function of the cost and constraint vectors [99]. Using this result we show that, conditionally on $X_1, X_2, \ldots,$ we have $L_n \xrightarrow{d} L_{\mu_0}$ for almost every realization of the data. The result is established by noting that L_{μ_0} has a continuous distribution function. We highlight that this result holds beyond the setting of empirical measure, i.e., when other weakly convergent estimators of the population distributions are employed. Indeed, the proof is carried out under general conditions which encompass both the empirical measure and the estimator used for testing graph isomorphisms, introduced in Section 3.6.

We find it noteworthy that while L_{μ_0} is the weak limit of a sequence of optimal values for random nonconvex quadratic programs (NP-hard), Theorem 11 shows that it can be approximated by solving linear programs. By contrast, the naïve bootstrap estimator may fail to be consistent as the directional derivative provided in Theorem 10 is nonlinear (see Proposition 1 in [74] or Corollary 3.1 in [79]). The m-out-of-n bootstrap [161, 17] may prove consistent, but its computational overhead and the required tuning of m render it computationally costly and unattractive compared to the direct estimator.

Remark 17 (Computational considerations)**.** Estimating the distribution of L_n given $X_1, \ldots, X_n$ requires solving multiple instances of a linear program with a fixed feasible set, but with varying cost vectors. Software for numerical resolution of linear programs, both open-source and commercial, are plentiful (see e.g. [86, 116, 107] for benchmarks of linear programming solvers) and are based on variants of the interior-point method [121], which has polynomial time complexity and the simplex method [54], which has exponential worst case complexity, but performs as well or better than the interior-point method on most problems.

3.4 Semi-Discrete Gromov-Wasserstein Distance

In this section, we consider the semi-discrete setting for the GW distance, where $\mu_0 \in \mathcal{P}(\mathbb{R}^{d_0})$ is supported in $\mathcal{X}_0$, an open ball centered at 0 with finite radius, and $\mu_1 \in \mathcal{P}(\mathbb{R}^{d_1})$ is supported on the points $\mathcal{X}_1 := \left(y^{(i)}\right)_{i=1}^N$. To establish limit distributions for the semi-discrete GW cost, we again require stability of the solution set. However, unlike the discrete case, we can no longer rely on an analysis of finite-dimensional quadratic programs. Rather, we prove that, under mild conditions, the semi-discrete (2, 2)-GW problem is solved by a Gromov-Monge map—a result of independent interest. This, in turn, allows us to deduce the desired stability.

To arrive at the main structural results concerning the existence of Gromov-Monge maps, we start by specializing some results for semi-discrete OT to the family of costs $c_{\mathbf{A}}$ figuring in the variational form of the GW distance (3.8). To that end, recall the so-called semi-dual form of the semi-discrete OT problem,[6]

$$\mathsf{OT}_c(\mu_0,\mu_1) = \sup_{z\in\mathbb{R}^N}\left\{\sum_{i=1}^N z_i\mu_1(\{y^{(i)}\}) + \int \min_{1\le i\le N}\left\{c(\cdot,y^{(i)}) - z_i\right\} d\mu_0\right\}. \tag{3.14}$$

We call a solution to (3.14) an *optimal vector* for $\mathsf{OT}_c(\mu_0,\mu_1)$.

3.4.1 Existence of Gromov-Monge maps

A standard approach for establishing the existence of OT maps for $\mathsf{OT}_c(\mu_0,\mu_1)$ employs the so-called twist condition, which guarantees injectivity of $\nabla_x c(x,\cdot)$ [208, Theorem 10.28]. We leverage a version of the twist condition for semi-discrete OT with a generic

[6]This representation is obtained by recasting the dual OT problem in (3.3) as $\sup_{\varphi_1\in C(\mathcal{X}_1)}\left\{\int \varphi_1^c d\mu_0 + \int \varphi_1 d\mu_1\right\}$, where $\varphi_1^c : x \in \mathcal{X}_0 \mapsto \inf_{y\in\mathcal{X}_1}\{c(x,y) - \varphi_1(y)\}$ is the c-transform of φ_1. In the semi-discrete case, one then identifies the OT potential φ_1 with a vector $z \in \mathbb{R}^N$ to arrive at (3.14).

cost from [122] to identify a primitive condition for $\mathsf{OT}_{\mathbf{A}}(\mu_0, \mu_1)$ to admit a Monge map, uniformly in $\mathbf{A}$. Once this proviso is met, Theorem 8 guarantees the existence of a Gromov-Monge map for the semi-discrete GW problem.

Theorem 12 (Existence of Gromov-Monge maps)**.** Assume without loss of generality that (μ_0, μ_1) are centered and let $\mathbf{A}^\star$ minimize $\Phi_{(\mu_0,\mu_1)}$. If $\mathcal{X}_1$ is such that $y^{(i)} - y^{(j)} \notin \ker(\mathbf{A}^\star)$ for every $i \neq j \in [N]$ with $\|y^{(i)}\| = \|y^{(j)}\|$, then there exists a GW plan for $\mathsf{D}(\mu_0, \mu_1)$ which is induced by the μ_0-a.e. unique map

$$T_{\mathbf{A}^\star} : x \in \mathcal{X} \mapsto y^{(I_{z^{\mathbf{A}^\star}}(x))}, \text{ for } I_{z^{\mathbf{A}^\star}}(x) \in \operatorname*{argmin}_{1 \leq i \leq N} \left(c_{\mathbf{A}^\star}(x, y^{(i)}) - z_i^{\mathbf{A}^\star} \right),$$

where $z_i^{\mathbf{A}^\star}$ is an optimal vector for $\mathsf{OT}_{\mathbf{A}^\star}(\mu_0, \mu_1)$. Further, letting $\mathrm{Lag}_{z^{\mathbf{A}^\star},i}$ denote the set $\{x \in \mathcal{X}_0 : i \in \operatorname{argmin}_{1 \leq i \leq N} \left(c_{\mathbf{A}^\star}(x, y^{(i)}) - z_i^{\mathbf{A}^\star} \right)\}$, $\mathbf{A}^\star = \frac{1}{2} \sum_{i=1}^N \int_{\mathrm{Lag}_{z^{\mathbf{A}^\star},i}} x d\mu_0(x) \left(y^{(i)} \right)^\intercal$. A sufficient condition for the assumption on $\mathcal{X}_1$ to hold at every $\mathbf{A} \in \mathbb{R}^{d_0 \times d_1}$ is that $\|y^{(i)}\| \neq \|y^{(j)}\|$ for every $i \neq j \in [N]$.

The proof of Theorem 12, given in Section 3.8.3, shows that the condition on the support of μ_1 is equivalent to a variant of the standard twist condition. This, in turn, guarantees that the corresponding OT problem admits a unique OT plan that is induced by a map and the result follows using the dual form (3.8). We emphasize that, for given μ_0, μ_1, it is easy to verify the sufficient condition that $\|y^{(i)}\| \neq \|y^{(j)}\|$ for every $i \neq j \in [N]$, whereas the general condition may be difficult to check as $\operatorname{argmin}_{\mathbb{R}^d} \Phi_{(\mu_0,\mu_1)}$ is *a priori* unknown.

Remark 18 (Semi-discrete Gromov-Monge map)Existence (and optimality) of Gromov-Monge maps for the GW problem (3.7) is known only in a few cases. Theorem 4.1.2 in [204] shows that it suffices to optimize over permutations when μ_0, μ_1 are both uniform probability measures supported on the same (finite) number of points, while Theorem 3.2 in [75] treats the inner product GW problem (i.e., when the cost function in (3.7) is set to

$\Delta(x,y,x',y') = \langle x,x'\rangle - \langle y,y'\rangle)$ between compactly supported and absolutely continuous distributions. The only results for the quadratic cost considered herein require additional high-level assumptions, such as that μ_0 and μ_1 be absolutely continuous and rotationally invariant about their barycenter [189, Theorem 9.21] or *a priori* knowledge that the cross-correlation matrix of some GW plan for $\mathsf{D}(\mu_0,\mu_1)$ is of full rank [204, Proposition 4.2.4]. While such conditions are restrictive and hard to verify in practice, they may hold beyond the semi-discrete setting. Theorem 12 presents a new instance of the existence of Gromov-Monge maps for the quadratic GW problem, assuming the semi-discrete setting with the primitive geometric condition on the magnitude of the discrete support points.

3.4.2 Stability and limit distributions

We now derive the asymptotic distribution of the empirical semi-discrete GW distance. To that end, we first study stability of D by leveraging the fact that the same condition guaranteeing existence of Gromov-Monge maps for the semi-discrete case, also imply the uniqueness of the OT plan for $\mathsf{OT}_{\mathbf{A}}(\bar{\mu}_0,\bar{\mu}_1)$ for every $\mathbf{A}\in\operatorname{argmin}_{\mathbb{R}^{d_0\times d_1}}\Phi_{(\mu_0,\mu_1)}$. As in the fully discrete case, we set $M := \frac{1}{2}\|\mathcal{X}_0\|_\infty\|\mathcal{X}_1\|_\infty$ so that, for any $(\nu_0,\nu_1)\in\mathcal{P}(\mathcal{X}_0)\times\mathcal{P}(\mathcal{X}_1)$, all minimizers of $\Phi_{(\bar{\nu}_0,\bar{\nu}_1)}$ are contained in $B_{\mathrm{F}}(M)$

To apply the unified approach described in Proposition 16, we define the function class

$$\mathcal{F}^\oplus := \{f_0\oplus f_1 : (f_0,f_1)\in\mathcal{F}_0\times\mathcal{F}_1\},\ \text{for } \mathcal{F}_0 = B_{C^k(\mathcal{X}_0)} + \mathcal{G}\cup\{0\},$$
$$\mathcal{F}_1 := \{f : \mathcal{X}_1^\circ\to\mathbb{R} : \|f\|_{\infty,\mathcal{X}_1}\le 1\},$$

where $k=\lfloor d/2\rfloor+1$, $B_{C^k(\mathcal{X}_0)}$ denotes the unit ball in $C^k(\mathcal{X}_0)$, and

$$\mathcal{G} = \Big\{x\in\mathcal{X}_0\mapsto \min_{1\le i\le N}\Big\{c_{\mathbf{A}}(x-\xi, y^{(i)}-\zeta)-z_i\Big\}$$
$$: \mathbf{A}\in B_{\mathrm{F}}(M), z\in\mathbb{R}^N, \|z\|_\infty\le K, \zeta\in\mathbb{R}^{d_1}, \|\zeta\|\le\|\mathcal{X}_1\|_\infty, \xi\in\mathcal{X}_0\Big\},$$

These definitions are motivated by the regularity of the OT potentials (Lemma 19).

Theorem 13 (Stability for semi-discrete GW)**.** Let $\mathcal{A} = \operatorname{argmin}_{B_{\mathrm{F}}(M)} \Phi_{(\bar{\mu}_0,\bar{\mu}_1)}$. Then, for any pairs $(\nu_0,\nu_1),(\rho_0,\rho_1) \in \mathcal{P}_{\mu_0} \times \mathcal{P}_{\mu_1}$, $\left|\mathsf{D}(\nu_0,\nu_1)^2 - \mathsf{D}(\rho_0,\rho_1)^2\right| \leq C\|\nu_0 \otimes \nu_1 - \rho_0 \otimes \rho_1\|_{\infty,\mathcal{F}^{\oplus}}$ for some constant $C > 0$ which is independent of $(\nu_0,\nu_1),(\rho_0,\rho_1)$, and, if $\bar{\mu}_0,\bar{\mu}_1$ verify the conditions of Theorem 12 at every $\mathbf{A} \in \mathcal{A}$ and $\operatorname{int}(\operatorname{spt}(\bar{\mu}_0))$ is connected with negligible boundary, we further have

$$\frac{d}{dt}\mathsf{D}(\mu_{0,t},\mu_{1,t})^2\Big|_{t=0} = \inf_{\mathbf{A}\in\mathcal{A}}\left\{\int (f_0 + g_{0,\pi^{\mathbf{A}}} + \bar{\varphi}_0^{\mathbf{A}})d(\nu_0 - \mu_0)\right.$$
$$\left. + \int (f_1 + g_{1,\pi^{\mathbf{A}}} + \bar{\varphi}_1^{\mathbf{A}})d(\nu_1 - \mu_1)\right\},$$

where f_0, f_1, $g_{0,\pi}$, $g_{1,\pi}$, $\mu_{0,t}$, and $\mu_{1,t}$ are as defined in Theorem 9, $\bar{\varphi}_i^{\mathbf{A}} = \varphi_i^{\mathbf{A}}(\cdot - \mathbb{E}_{\mu_i}[X])$ for $i = 0,1$, and, for $\mathbf{A} \in \mathcal{A}$, $\pi^{\mathbf{A}}$ is the unique OT plan for $\mathsf{OT}_{\mathbf{A}}(\bar{\mu}_0,\bar{\mu}_1)$, and $(\varphi_0^{\mathbf{A}},\varphi_1^{\mathbf{A}})$ is any choice of corresponding OT potentials.

The proof of Theorem 13 relies on the decomposition $\mathsf{D}(\bar{\nu}_0,\bar{\nu}_1)^2 = \mathsf{S}_1(\bar{\nu}_0,\bar{\nu}_1) + \mathsf{S}_2(\bar{\nu}_0,\bar{\nu}_1)$. The desired right differentiability and Lipschitz continuity of S_1 is relatively straightforward, whereas that of S_2 requires a more careful analysis. This is since S_2 includes an infimum, and the fact that $\bar{\mu}_{0,t}$ and $\bar{\mu}_{1,t}$ may not share the same support as $\bar{\mu}_0$ and $\bar{\mu}_1$, respectively. This complicates the comparison of OT potentials corresponding to different marginals. We address this issue by first computing the right derivative of $\mathsf{OT}_{\mathbf{A}}(\bar{\nu}_0,\bar{\nu}_1)$ at $(\bar{\mu}_0,\bar{\mu}_1)$ for a fixed $\mathbf{A}$, and then use the fact that $\mathbf{A} \in B_{\mathrm{F}}(M) \mapsto \mathsf{OT}_{\mathbf{A}}(\bar{\mu}_{0,t},\bar{\mu}_{1,t})$ converges uniformly to $\mathbf{A} \in B_{\mathrm{F}}(M) \mapsto \mathsf{OT}_{\mathbf{A}}(\bar{\mu}_0,\bar{\mu}_1)$, which enables evaluating the derivative of the infimum. Complete details are included in Section 3.8.3.

Compared to Theorem 9, Theorem 13 includes neither an infimum over OT couplings nor a supremum over OT potentials. This distinction is a consequence of the uniqueness of OT couplings for $\mathsf{OT}_{\mathbf{A}}(\bar{\mu}_0,\bar{\mu}_1)$ for $\mathbf{A} \in \mathcal{A}$ under the conditions of Theorem 12 and the uniqueness (up to versions) of the OT potentials resulting from the assumption on $\mathrm{spt}(\bar{\mu}_0)$, see Lemma 19. This latter condition is standard in the literature [61, 62].

Armed with the stability result, we obtain the distributional limit for semi-discrete GW by invoking the unified approach from Proposition 16.

Theorem 14 (Limit distributions for semi-discrete GW)**.** In the setting of Theorem 13, let $(\mu_{0,n}, \mu_{1,n})$ satisfy Assumption 1. Then,

$$\sqrt{n}\left(\mathsf{D}(\mu_{0,n}, \mu_{1,n})^2 - \mathsf{D}(\mu_0, \mu_1)^2\right) \xrightarrow{d} \inf_{\mathbf{A}\in\mathcal{A}} \{\chi_{\mu_0\otimes\mu_1}((f_0 + g_{0,\pi} + \bar{\varphi}_0^{\mathbf{A}}) \oplus (f_1 + g_{1,\pi} + \bar{\varphi}_1^{\mathbf{A}}))\}.$$

If $(\mu_{0,n}, \mu_{1,n}) = (\hat{\mu}_{0,n}, \hat{\mu}_{1,n})$, Assumption 1 is met and $\chi_{\mu_0\otimes\mu_1}(g_0 \oplus g_1) = G_{\mu_0}(g_0) + G_{\mu_0}(g_1)$ for any $g_0 \oplus g_1 \in \mathcal{F}^{\oplus}$, where G_{μ_i} is a tight μ_i-Brownian bridge in $\ell^\infty(\mathcal{F}_i)$ for $i \in \{0, 1\}$.

As in the discrete case, we slightly abuse notation by identifying $\chi_{\mu_0\otimes\mu_1}$ with its extension to a certain dilation of $\mathcal{F}^{\oplus}$, see Remark 14 for details. The approach outlined in Remark 15 can also be applied in this case to obtain a limit distribution for D (without the square). Finally, as discussed in Remark 16, the consistency of the naïve bootstrap for the limit in Theorem 14 can be established if $\mathcal{A}$ is a singleton. At present, sufficient conditions for such a result to hold are unknown.

3.5 Entropic Gromov-Wasserstein Distance

To go beyond the fully discrete and semi-discrete settings, we consider the entropic GW distance under the sole assumption that μ_0, μ_1 are 4-sub-Weibull with parameter σ^2. Note that if $(\mu_0, \mu_1) \in \mathcal{P}_{4,\sigma}(\mathbb{R}^{d_0}) \times P_{4,\sigma}(\mathbb{R}^{d_1})$, then $\bar{\mu}_0$ and $\bar{\mu}_1$ are 4-sub-Weibull with parameter $\bar{\sigma}^2 := 8\frac{2+\log 2}{\log 2}\sigma^2$ and $\frac{1}{2}\sqrt{M_2(\bar{\mu}_0)M_2(\bar{\mu}_1)} \leq \sigma$, these auxiliary results are proved in Section 3.8.5. We thus set $M = \sigma$ so that $B_{\mathrm{F}}(M)$ contains all minimizers of $\Phi_{(\bar{\mu}_0,\bar{\mu}_1)}$ (see

Theorem 23). The derivation is similar to the semi-discrete case, as the EOT problem that appears in the entropic GW variational form (see Section B.1) enjoys uniqueness of EOT plans and potentials.

The sub-Weibull assumption stems from the regularity theory for EOT potentials from Lemma 4 of [223] (restated in Lemma 26 ahead), which motivates the choice of the function class as $\mathcal{F}^{\oplus} := \{f_0 \oplus f_1 : (f_0, f_1) \in \mathcal{F}_0 \times \mathcal{F}_1\}$, where

$$\mathcal{F}_i = \left\{ f \in C^{\bar{k}}(\mathbb{R}^{d_i}) : |f| \leq 1 + \|\cdot\|^4, |D^\alpha f| \leq 1 + \|\cdot\|^{3\bar{k}}, \forall \alpha \in \mathbb{N}_0^d \text{ with } 0 < |\alpha| \leq \bar{k} \right\},$$

for $\bar{k} = \lfloor (d_0 \vee d_1)/2 \rfloor + 1$. We have the following stability theorem.

Theorem 15 (Stability for entropic GW)**.** Let $\mathcal{A} = \operatorname{argmin}_{B_{\mathrm{F}}(M)} \Phi_{(\bar{\mu}_0, \bar{\mu}_1)}$. Then, for any pairs $(\nu_0, \nu_1), (\rho_0, \rho_1) \in \mathcal{P}_{4,\sigma}(\mathbb{R}^{d_0}) \times \mathcal{P}_{4,\sigma}(\mathbb{R}^{d_1})$, $\left|\mathsf{D}_\varepsilon(\nu_0, \nu_1)^2 - \mathsf{D}_\varepsilon(\rho_0, \rho_1)^2\right| \leq C\|\nu_0 \otimes \nu_1 - \rho_0 \otimes \rho_1\|_{\infty, \mathcal{F}^{\oplus}}$ for some constant $C > 0$ which is independent of $(\nu_0, \nu_1), (\rho_0, \rho_1)$, and,

$$\begin{aligned}\frac{d}{dt}\mathsf{D}_\varepsilon(\mu_{0,t}, \mu_{1,t})^2\Big|_{t=0} = \inf_{\mathbf{A} \in \mathcal{A}} \Bigg\{ &\int (f_0 + g_{0,\pi^{\mathbf{A}}} + \bar{\varphi}_0^{\mathbf{A}}) d(\nu_0 - \mu_0) \\ &+ \int (f_1 + g_{1,\pi^{\mathbf{A}}} + \bar{\varphi}_1^{\mathbf{A}}) d(\nu_1 - \mu_1) \Bigg\},\end{aligned}$$

where f_i, $g_{i,\pi}$, $\mu_{i,t}$ $\bar{\varphi}_i^{\mathbf{A}}$, for $i = 0, 1$, as well as $\pi^{\mathbf{A}}$ are all defined as in Theorem 13 with $\mathsf{OT}_{\mathbf{A},\varepsilon}(\bar{\mu}_0, \bar{\mu}_1)$ in place of $\mathsf{OT}_{\mathbf{A}}(\bar{\mu}_0, \bar{\mu}_1)$.

The proof of Theorem 15 follows similar lines as that of Theorem 13 with the important distinction that the measures need only be 4-sub-Weibull. The regularity of the EOT potentials and the strong structural properties of solutions to the EOT problem enable treating measures with unbounded support. See Section 3.8.5 for complete details.

Invoking Proposition 16 now leads to the following characterization of distributional limits for the entropic GW cost and bootstrap consistency when $\mathcal{A}$ is a singleton. For $X_{0,1}^B, \ldots, X_{0,n}^B$ and $X_{1,1}^B, \ldots, X_{1,n}^B$ independent collections of i.i.d. samples from $\hat{\mu}_{0,n}$ and $\hat{\mu}_{1,n}$, let $\hat{\mu}^B_{k,n} = \frac{1}{n}\sum_{i=1}^n \delta_{X^B_{k,i}}$ for $k \in \{0,1\}$ denote the bootstrap measures and $\mathbb{P}^B$ the conditional probability given the data.

Theorem 16 (Limit distributions for entropic GW and bootstrap consistency)**.** In the setting of Theorem 15, let $(\mu_{0,n},\mu_{1,n})_{n\in\mathbb{N}}$ satisfy Assumption 1 with $\mathcal{P}_{4,\sigma}(\mathbb{R}^{d_0}) \times \mathcal{P}_{4,\sigma}(\mathbb{R}^{d_1})$ in place of $\mathcal{P}_{\mu_0} \times \mathcal{P}_{\mu_1}$, then

$$\sqrt{n}\left(\mathsf{D}_\varepsilon(\mu_{0,n},\mu_{1,n})^2 - \mathsf{D}_\varepsilon(\mu_0,\mu_1)^2\right) \xrightarrow{d} \inf_{\mathbf{A}\in\mathcal{A}}\{\chi_{\mu_0\otimes\mu_1}((f_0+g_{0,\pi}+\bar{\varphi}_0^{\mathbf{A}})\oplus(f_1+g_{1,\pi}+\bar{\varphi}_1^{\mathbf{A}}))\}.$$

Moreover, if $\varepsilon > 16\sqrt{M_4(\bar{\mu}_0)M_4(\bar{\mu}_1)}$, $\mathcal{A}$ is a singleton and the limit is a point evaluation.

If $(\mu_{0,n},\mu_{1,n}) = (\hat{\mu}_{0,n},\hat{\mu}_{1,n})$, then Assumption 1 is met and $\chi_{\mu_0\otimes\mu_1}(g_0\oplus g_1) = G_{\mu_0}(g_0) + G_{\mu_0}(g_1)$ for any $g_0 \oplus g_1 \in \mathcal{F}^\oplus$, where G_{μ_i} is a tight μ_i-Brownian bridge in $\ell^\infty(\mathcal{F}_i)$ for $i \in \{0,1\}$. Further, if $\mathcal{A} = \{\mathbf{A}\}$ is a singleton, the limit is normal and the bootstrap is consistent

$$\sup_{t\in\mathbb{R}}\left|\mathbb{P}^B\left(\sqrt{n}\left(\mathsf{D}(\hat{\mu}^B_{0,n},\hat{\mu}^B_{1,n})^2 - \mathsf{D}(\hat{\mu}_{0,n},\hat{\mu}_{1,n})^2\right) \le t\right) - \mathbb{P}\left(N(0,v_0^2+v_1^2)\le t\right)\right| \xrightarrow{\mathbb{P}} 0,$$

for $v_0^2 = \mathrm{Var}_{\mu_0}(f_0+g_{0,\pi}+\bar{\varphi}_0^{\mathbf{A}})$ and $v_1^2 = \mathrm{Var}_{\mu_1}(f_1+g_{1,\pi}+\bar{\varphi}_1^{\mathbf{A}})$ provided that $v_0^2+v_1^2 > 0$.

Again, we identify $\chi_{\mu_0\otimes\mu_1}$ with its extension to a certain dilation of $\mathcal{F}^\oplus$, see Remark 14. To obtain a limit distribution for D_ε itself, the technique from Remark 15 can be used. The condition on ε that guarantees that $\mathcal{A}$ is a singleton stems from Theorem 6 in [164], which shows that in that case the objective of the entropic GW variational is strictly convex in $\mathbf{A}$. Beyond this condition, it is unclear when $\mathcal{A}$ is a singleton so that the consistency of the naïve bootstrap cannot be established in general (see Remark 16).

3.6 Testing for Graph Isomorphism

A key property of the GW distance is that it identifies distributions that are related through an isometry. Armed with our limit theorems for GW distances, we consider an application to testing if two random graph models are isomorphic, based on samples from them. As the limit distribution theory accounts for Euclidean mm space, to employ it for graph isomorphism testing, we propose a framework for embedding distributions of random graphs on N vertices into $\mathcal{P}(\mathbb{R}^N)$. Crucially, the embedding is constructed so that the random graph are isomorphic if and only if the GW distance between the embedded distributions nullifies. This allows casting the graph isomorphism testing question as a test for equality of the embedded distributions under the GW distance.

3.6.1 Embedding of random graph distributions

Consider the set $\mathcal{G}_N$ of all weighted undirected graphs on N vertices with nonnegative weights and no self-loops. A graph $G \in \mathcal{G}_N$ is specified by its matrix of weights $\mathbf{W}^G \in \mathbb{R}^{N\times N}$, which is defined as $\mathbf{W}^G_{ij} = \mathbf{W}^G_{ji} = w$ if the nodes i and j are connected by an edge with weight w for $i \neq j \in [N]$. We adopt the convention that a weight of 0 indicates the absence of an edge, which implies $\mathbf{W}^G_{ii} = 0$ for every $i \in [N]$ since there are no self-loops. With this, a distribution $\nu \in \mathcal{P}(\mathcal{G}_N)$ can be specified by fixing $N(N-1)/2$ univariate *edge distributions*, $(\rho_{ij})_{i,j\in[N]:\, i<j} \subset \mathcal{P}([0,\infty))$, and setting $\mathrm{law}(\mathbf{W}^G_{ij}) = \rho_{ij}$ for $G \sim \nu$. Such a distribution treats each edge independently, and we write $\mathcal{P}_{\mathrm{ind}}(\mathcal{G}_N)$ to denote the set of all such distributions to emphasize this fact.

To formulate the comparison of two distributions $\nu_0, \nu_1 \in \mathcal{P}_{\mathrm{ind}}(\mathcal{G}_N)$ with associated edge distributions $(\rho_{0,ij})_{i,j\in[N]:\, i<j}, (\rho_{1,ij})_{i,j\in[N]:\, i<j}$ in terms of the Euclidean GW distance considered herein, we introduce the following embedding $\iota : \mathcal{P}([0,\infty))^{N(N-1)/2} \to \mathcal{P}(\mathbb{R}^N)$,

$$\iota : (\rho_{ij})_{\substack{i,j\in[N]\\ i<j}} \mapsto \frac{1}{N(N+1)(N-1)} \sum_{\substack{i,j=1\\ i\neq j}}^{N} \eta_{\rho_{ij}} + \frac{1}{N+1} \sum_{i=1}^{N} \delta_{-e_i}, \tag{3.15}$$

where $\eta_{\rho_{ij}}(B) = \int \mathbb{1}_B(-e_i + te_j)d\rho_{ij}(t)$, for every Borel set $B \subset \mathbb{R}^N$. To simplify notation, we identify $\nu \in \mathcal{P}_{\text{ind}}(\mathcal{G}_N)$ with its edge distributions, $(\rho_{ij})_{i,j\in[N]:\, i<j}$, and write $\iota(\nu)$ in place of ι applied to the edge distributions.

This construction enables a comparison of ν_0 and ν_1 up to a relabelling of the graph nodes. Precisely, ν_0 and ν_1 are said to be isomorphic if there exists a permutation $\sigma : [N] \to [N]$ for which $\rho_{1,ij} = \rho_{0,\sigma(i)\sigma(j)}$ for every $i \neq j \in [N]$ under the convention that $\rho_{0,ij} = \rho_{0,ji}$ for every $i \neq j$. This notion of isomorphic graph distributions coincides with the standard notion of isomorphic graphs when ν_0 and ν_1 are Dirac distributions. We now establish that ν_0 and ν_1 are isomorphic if and only if $\iota(\nu_0)$ and $\iota(\nu_1)$ are identified under the GW distance.

Proposition 8 (Identification of isomorphic graph distributions).For $\nu_0, \nu_1 \in \mathcal{P}_{\text{ind}}(\mathcal{G}_N)$, $\mu_0 = \iota(\nu_0)$, and $\mu_1 = \iota(\nu_1)$, then $\mathsf{D}(\mu_0, \mu_1) = 0$ if and only if ν_0 and ν_1 isomorphic.

The proof of Proposition 8, included in Section 3.8.6, uses the particular construction of the embedding in (3.15). Notably, the first term in (3.15) compares the individual edge distributions, whereas the second term ensures that the only possible isometry relating μ_0 and μ_1 is a permutation of coordinate axes.

3.6.2 GW testing for graph isomorphism

We are ready to set up the statistical GW-based test for graph isomorphism. Given mutually independent collections of sample graphs $G_{0,1}, \ldots, G_{0,n} \sim \nu_0 \in \mathcal{P}_{\text{ind}}(\mathcal{G}_N)$ and $G_{1,1}, \ldots, G_{1,n} \sim \nu_1 \in \mathcal{P}_{\text{ind}}(\mathcal{G}_N)$, we seek to test between the hypotheses:

$$H_0 : \nu_0 \text{ and } \nu_1 \text{ are isomorphic} \qquad \text{vs.} \qquad H_1 : \nu_0 \text{ and } \nu_1 \text{ are not isomorphic.}$$

Given Proposition 8, this amounts to testing $H_0 : \mathsf{D}(\iota(\nu_0), \iota(\nu_1)) = 0$ versus the alternative $H_1 : \mathsf{D}(\iota(\nu_0), \iota(\nu_1)) > 0$. To formulate the proposed test statistic, let $\hat{\rho}_{k,ij,n} = \frac{1}{n} \sum_{l=1}^{n} \delta_{\mathbf{W}_{ij}^{G_{k,l}}}$, for $k \in \{0,1\}$ and $i,j \in [N]$ with $i < j$, which corresponds to the empirical edge distribution, and let $\nu_{0,n}$ and $\nu_{1,n}$ denote the corresponding distributions on $\mathcal{G}_N$. With this, the test statistic is given by $\sqrt{n}\mathsf{D}(\iota(\nu_{0,n}),\iota(\nu_{1,n}))^2$ and the critical value for the asymptotic level $\alpha \in (0,1)$ is given by $q_{n,1-\alpha}$, the conditional $(1-\alpha)$ quantile of the direct estimator $L_n = \sqrt{2}\sup_{u\in\mathcal{H}_n} Z_n^\intercal u$ for $Z_n \sim N(0,\mathbf{\Sigma}_n)$ as defined in (3.35), where $\mu_{0,n} = \iota(\nu_{0,n})$ and $\mathbf{\Sigma}_n$ is the empirical analogue of $\mathbf{\Sigma}_Z$ defined in (3.59). Using the machinery developed in our study of the empirical discrete GW distance from Section 3.3, we prove in Section 3.8.6 that the proposed test is consistent.

Theorem 17 (Test consistency and asymptotic distribution)**.** Let $\nu_0,\nu_1 \in \mathcal{P}_{\mathrm{ind}}(\mathcal{G}_N)$ be such that their edge distributions, $(\rho_{0,ij})_{i,j=1:\,i<j}^N$ and $(\rho_{1,ij})_{i,j=1:\,i<j}^N$, are supported in $[0,K]$. Then, with $\mu_k = \iota(\nu_k)$ and $\mu_{k,n} = \iota(\nu_{k,n})$ for $k \in \{0,1\}$, we have

$$\mathbb{E}[|\mathsf{D}(\mu_{0,n},\mu_{1,n})^2 - \mathsf{D}(\mu_0,\mu_1)^2|] \lesssim (1+2K^4)n^{-\frac{1}{2}}.$$

Furthermore, if $(\rho_{0,ij})_{i,j=1:\,i<j}^N$ and $(\rho_{0,ij})_{i,j=1:\,i<j}^N$ are finitely supported, then with the notation of Theorem 10, there exists Gaussian processes χ_{μ_0} in $\ell^\infty(\mathcal{F}_0)$ and χ_{μ_1} in $\ell^\infty(\mathcal{F}_1)$ with

$$\begin{aligned}\sqrt{n}&\left(\mathsf{D}(\mu_{0,n},\mu_{1,n})^2 - \mathsf{D}(\mu_0,\mu_1)^2\right)\\ &\xrightarrow{d} \inf_{\mathbf{A}\in\mathcal{A}} \inf_{\pi\in\bar{\Pi}_{\mathbf{A}}^\star} \sup_{(\varphi_0,\varphi_1)\in\bar{\mathcal{D}}_{\mathbf{A}}} \left\{\chi_{\mu_0}(f_0+g_{0,\pi}+\bar{\varphi}_0)+\chi_{\mu_1}(f_1+g_{1,\pi}+\bar{\varphi}_1)\right\}.\end{aligned}$$

If, in addition, $\mathsf{D}(\mu_0,\mu_1) = 0$, then $\sqrt{n}\mathsf{D}(\mu_{0,n},\mu_{1,n})^2 \xrightarrow{d} \sqrt{2}\|\chi_{\mu_0}\|_{\infty,\bar{\mathcal{H}}}$, where $\bar{\mathcal{H}}$ is as defined in Corollary 5 and the test which rejects the null hypothesis, $H_0 : \nu_0$ and ν_1 are isomorphic, if $\sqrt{n}\mathsf{D}(\mu_{0,n},\mu_{1,n})^2 > q_{n,1-\alpha}$ is asymptotically consistent with level α.

3.6.3 Computational framework

To test if two random graph models are isomorphic by using the GW-based approach, the distance $\mathsf{D}(\mu_{0,n},\mu_{1,n})$ should be computed exactly to obtain statistically significant results. As computing this statistic is NP-hard in general, we propose to apply a local subgradient-based solver with a particular warm-start strategy.

Recalling (3.8), $\mathsf{D}(\eta_0,\eta_1)^2 = \mathsf{S}_1(\bar{\eta}_0,\bar{\eta}_1) + \inf_{\mathbf{A}\in\mathbb{R}^{d_0\times d_1}} \Phi_{(\eta_0,\eta_1)}(\mathbf{A})$ for any $(\eta_0,\eta_1) \in \mathcal{P}(\mathbb{R}^{d_0}) \times \mathcal{P}(\mathbb{R}^{d_1})$, whereby $\mathsf{D}(\eta_0,\eta_1)^2$ can be computed by minimizing $\Phi_{(\eta_0,\eta_1)}$. This approach was explored in the context of the entropic GW problem in [164], where the minimization problem is smooth, but possibly nonconvex. We adapt Algorithm 2 in that work to the nonsmooth case by substituting the subgradient of $\Phi_{\bar{\eta}_0,\bar{\eta}_1}$, derived in Lemma 7, in place of the gradient (see Algorithm 2). Though the convergence guarantees presented in [164, Theorem 11] no longer apply in the current setting, in the experiments presented in Section 3.7, Algorithm 2 always returns a point at which a subgradient with norm less than 5×10^{-8} exists when L is set to 128.

A warm-start for Algorithm 2 consists of a particular choice for the initialization $\mathbf{C}_0$. Our proposed warm-start methods are motivated by the fact that, under the null, the set of optimizers for $\Phi_{(\bar{\mu}_0,\bar{\mu}_1)}$ is a subset of all matrices of the form $\mathbf{A}_T = \frac{1}{2}\int xT(x)^\intercal d\bar{\mu}_0(x)$, where $T : \mathbb{R}^N \to \mathbb{R}^N$ is an isometry corresponding to a permutation of the coordinate axes. We write $\mathcal{T}$ for the set of all such isometries. The fact that any limit point of a sequence of minimizers of $\Phi_{(\bar{\mu}_{0,n},\bar{\mu}_{1,n})}$ minimizes $\Phi_{(\bar{\mu}_0,\bar{\mu}_1)}$ with probability 1 (see Section B.2) suggests numerically estimating $\mathsf{D}(\mu_{0,n},\mu_{1,n})^2$ via an *exhaustive search*. Namely, initialize Algorithm 2 at all such matrices $\mathbf{A}_T$, record the output of the subgradient method as Λ_T, and return $\min_{T\in\mathcal{T}} \Lambda_T$ as an estimate for the GW distance. The exhaustive search is accurate under the null for large n, but onerous given that it requires solving $N!$ optimization problems, which limits its usage to small scale problems.

Algorithm 2 Subgradient method for estimating $\mathsf{D}(\eta_0, \eta_1)^2$

Given the initialization $\mathbf{C}_0 \in \mathcal{D}_M$, tolerance $\delta > 0$, and $L > 0$, fix $\beta_k = \frac{1}{2L}$, $\gamma_k = \frac{k}{4L}$, $\tau_k = \frac{2}{k+2}$, $k = 1$.

$\mathbf{A}_1 \leftarrow \mathbf{C}_0$

$\mathbf{G}_1 \leftarrow 64\mathbf{A}_1 + 32 \int xy^\intercal d\pi_{\mathbf{A}_1}(x, y)$ for any OT plan $\pi_{\mathbf{A}_1}$ of $\mathsf{OT}_{\mathbf{A}_1}(\bar{\eta}_0, \bar{\eta}_1)$

while $\|\mathbf{G}_k\|_{\mathrm{F}} \geq \delta$ **do**

  $\mathbf{B}_k \leftarrow \min\left(1, \frac{M}{2\|\mathbf{A}_k - \beta_k \mathbf{G}_k\|_F}\right)(\mathbf{A}_k - \beta_k \mathbf{G}_k)$

  $\mathbf{C}_k \leftarrow \min\left(1, \frac{M}{2\|\mathbf{C}_{k-1} - \gamma_k \mathbf{G}_k\|_F}\right)(\mathbf{C}_{k-1} - \gamma_k \mathbf{G}_k)$

  $\mathbf{A}_{k+1} \leftarrow \tau_k \mathbf{C}_k + (1 - \tau_k)\mathbf{B}_k$

  $\mathbf{G}_{k+1} \leftarrow 64\mathbf{A}_{k+1} + 32 \int xy^\intercal d\pi_{\mathbf{A}_{k+1}}(x, y)$ for any OT plan $\pi_{\mathbf{A}_{k+1}}$ of $\mathsf{OT}_{\mathbf{A}_{k+1}}(\bar{\eta}_0, \bar{\eta}_1)$

  $k \leftarrow k + 1$

return $\mathsf{S}_1(\bar{\eta}_0, \bar{\eta}_1) + \Phi_{(\bar{\eta}_0, \bar{\eta}_1)}(\mathbf{B}_k)$

Relaxed graph matching

To scale up the computation of $\mathsf{D}(\mu_{0,n}, \mu_{1,n})^2$ to larger problems, we proposed an alternative approach to the exhaustive search, applicable for unweighted graphs. Consider the related problem of directly aligning the empirical probabilities $\hat{p}_{0,ij}$ and $\hat{p}_{1,ij}$ of two unweighted graphs by defining

$$\hat{\mathbf{P}}_0 = \begin{pmatrix} 0 & \hat{p}_{0,12} & \hat{p}_{0,13} & \dots & \hat{p}_{0,1N} \\ \hat{p}_{0,21} & 0 & \hat{p}_{0,23} & \dots & \hat{p}_{0,2N} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \hat{p}_{0,N1} & \hat{p}_{0,N2} & \hat{p}_{0,N3} & \dots & 0 \end{pmatrix} \quad \hat{\mathbf{P}}_1 = \begin{pmatrix} 0 & \hat{p}_{1,12} & \hat{p}_{1,13} & \dots & \hat{p}_{1,1N} \\ \hat{p}_{1,21} & 0 & \hat{p}_{1,23} & \dots & \hat{p}_{1,2N} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \hat{p}_{1,N1} & \hat{p}_{1,N2} & \hat{p}_{1,N3} & \dots & 0 \end{pmatrix},$$

and solving the problem $\min_{\mathbf{E}\in\mathcal{E}} \|\mathbf{E}\hat{\mathbf{P}}_0 - \hat{\mathbf{P}}_1\mathbf{E}\|_{\mathrm{F}}^2$, where $\mathcal{E}$ is the set of all permutation matrices on $\mathbb{R}^N$. This is known as the graph matching problem and, whose exact solution generally requires an exhaustive search, though the objective can be evaluated quickly.

A common relaxation of the graph matching problem replaces $\mathcal{E}$ by its convex hull, known as the *Birkhoff polytope*, $\mathrm{conv}(\mathcal{E}) = \left\{\mathbf{E} \in \mathbb{R}^{N\times N} : \mathbf{E} \geq 0, \mathbf{E}\mathbb{1}_{N_0} = \mathbf{E}^\intercal \mathbb{1}_{N_0} = \mathbb{1}_{N_0}\right\}$, which yields the convex relaxation $\min_{\mathbf{R}\in\mathrm{conv}(\mathcal{E})} \|\mathbf{R}\hat{\mathbf{P}}_0 - \hat{\mathbf{P}}_1\mathbf{R}\|_{\mathrm{F}}^2$. As a solution, $\mathbf{R}^\star$, to this relaxed problem need not be a permutation matrix, a projection onto the set of permutation matrices is performed by solving $\max_{\mathbf{E}\in\mathcal{E}}\langle \mathbf{R}^\star, \mathbf{E}\rangle_{\mathrm{F}}$, which is known as the linear assignment problem and can be solved efficiently. The permutation matrix $\mathbf{E}^\star$ obtained by performing these two steps is called a *relaxed graph matching*. The approach is summarized in Algorithm 3 below.

Algorithm 3 Relaxed graph matching

Given $\mu_{0,n}, \mu_{1,n}$, construct $\hat{\mathbf{P}}_0$ and $\hat{\mathbf{P}}_1$,

1: Obtain $\mathbf{R}^\star \in \operatorname{argmin}_{\mathbf{R}\in\mathrm{conv}(\mathcal{E})}\|\mathbf{R}\hat{\mathbf{P}}_0 - \hat{\mathbf{P}}_1\mathbf{R}\|_{\mathrm{F}}^2$ using convex optimization software.

2: Obtain $\mathbf{E}^\star \in \operatorname{argmax}_{\mathbf{E}\in\mathcal{E}}\langle \mathbf{R}^\star, \mathbf{E}\rangle_{\mathrm{F}}$ using a linear assignment problem solver.

3: **return** output of Algorithm 2 initialized at $\mathbf{A}_{T^\star}$, where $T^\star$ is the permutation of axes corresponding to $\mathbf{E}^\star$.

Remarkably, there are settings where this relaxed graph matching is known to coincide with the solution of the original graph matching problem.

Proposition 9 (Theorem 2 in [2]). Fix $\delta, \epsilon > 0$ and suppose that the eigenvalues of $\hat{\mathbf{P}}_0$, $(\lambda_i)_{i=1}^N$, satisfy $\min_{i,j\in[N],\ i\neq j}|\lambda_i - \lambda_j| > \delta$, $\max_{i\in[N]}|\lambda_i| = 1$, and its eigenvectors $(v_i)_{i=1}^N$ satisfy $\epsilon \leq v_i^\intercal \mathbb{1}_N \leq \frac{1}{\epsilon}$ for $i \in [N]$. If $\|\mathbf{E}^\star\hat{\mathbf{P}}_0 - \hat{\mathbf{P}}_1\mathbf{E}^\star\|_{\mathrm{F}} < \frac{\delta^2\epsilon^4}{12N^{1.5}}$, then the graph matching problem and its relaxation are equivalent.

The assumptions of Proposition 9 effectively guarantee that both the relaxed and original problems admit a unique solution (in particular, the only isomorphism relating $\hat{\mathbf{P}}_0$ to itself must be the identity). Observe that the conditions concerning $\hat{\mathbf{P}}_0$ can be verified *a priori*, and both $\hat{\mathbf{P}}_0$ and $\hat{\mathbf{P}}_1$ can be normalized such that the condition $\max_{i\in[N]}|\lambda_i| = 1$ holds. Consequently, upon obtaining $\mathbf{E}^\star$, it suffices to verify that $\|\mathbf{E}^\star\hat{\mathbf{P}}_0 - \hat{\mathbf{P}}_1\mathbf{E}^\star\|_{\mathrm{F}}$ satisfies the final condition to confirm if the relaxation $\mathbf{E}^\star$ solves the graph matching problem.

Although Proposition 9 is only applicable in certain settings, the underlying optimization problems can be solved efficiently thus enabling the matching of moderately sized graphs. Moreover, this approach may still yield good solutions beyond the theoretical guarantees. This graph matching approach is well-suited for unweighted graphs, but appears difficult to generalize to weighted case. By contrast, exhaustive search can easily be applied to weighted graphs, but is much more costly to implement and hence is limited to small-scale problems. A thorough numerical study of the performance of the exhaustive search procedure and Algorithm 3 in provided in Section B.3.

3.7 Numerical Experiments

The experiments in this section serve to verify the different results presented in the paper. We examine settings for graph isomorphism testing and present results that numerically validate the developed limit distribution theory.

3.7.1 Testing for graph isomorphism

The following experiments serve to empirically validate Theorem 17 by studying the type 1 and type 2 errors of the proposed test for graph isomorphism. As the problem is inherently discrete, we operate under the framework of Section 3.3 and focus on the consistency of the direct estimator L_n of the limit, which employs linear programming (see Section 3.3.3). We implement the test which rejects the null hypothesis (ν_0 and ν_1 are isomorphic) if $\sqrt{n}\mathsf{D}(\mu_{0,n}, \mu_{1,n})^2 > q_{n,1-\alpha}$, where $q_{n,1-\alpha}$ is the $(1-\alpha)$ quantile of L_n and α is the significance level.

Unweighted graphs

Figure 3.1 presents the underlying distributions ν_0, ν_1, and ν_2 on the set of binary graphs on 10 vertices. Here, ν_0 and ν_1 are chosen to be isomorphic whereas ν_2 is set to be equal

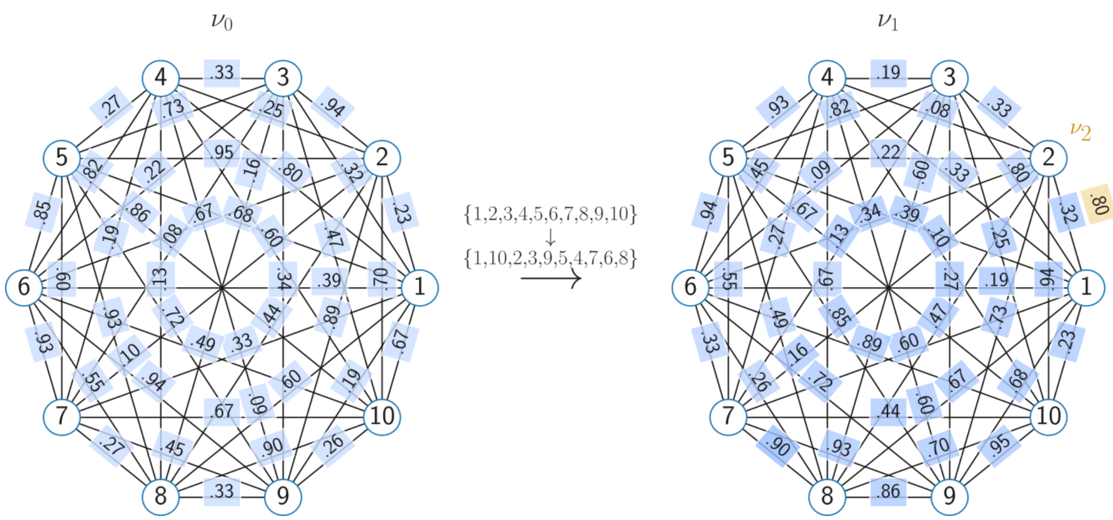


Figure 3.1: Distributions from Section 3.7.1. ν_0 is generated by sampling $(p_{0,ij})_{\substack{i,j\in[N] \\ i<j}}$ from the uniform distribution on $[0, 1]$. ν_1 is obtained from ν_0 by a randomly chosen permutation. The modified probability for ν_2 is highlighted in orange. Edge probabilities are presented with two significant figures for ease of reading.

to ν_1 save for the fact that we set $p_{2,12} = p_{1,13}$, resulting in ν_0 and ν_2 not being isomorphic.

We generate datasets of $n \in \{10, 25, 50, 100, 500, 1000, 5000\}$ graphs from each of these distributions and construct the estimators $\mu_{0,n}, \mu_{1,n}$, and $\mu_{2,n}$ as $\iota(\nu_{0,n}), \iota(\nu_{1,n})$, and $\iota(\nu_{2,n})$. We then compute $\mathsf{D}(\mu_{0,n}, \mu_{1,n})^2$ and $\mathsf{D}(\mu_{0,n}, \mu_{2,n})^2$ approximately via Algorithm 3. Next, we generate 200 samples from L_n using Algorithm 4 (an extension of Algorithm 1 to estimators other than the empirical measures, as employed herein) and estimate $q_{n,1-\alpha}$ for $\alpha \in \{k/9\}_{k=0}^{9}$. With this, we record if $\sqrt{n}\mathsf{D}(\mu_{0,n}, \mu_{1,n})^2 > q_{n,1-\alpha}$, i.e., the test spuriously rejects the null hypothesis, or if $\sqrt{n}\mathsf{D}(\mu_{0,n}, \mu_{2,n})^2 \leq q_{n,1-\alpha}$, whence the test fails to reject the null when the alternative is true. To estimate the type 1 and type 2 errors, we perform 100 repetitions of this procedure and report the fraction of repetitions for which $\sqrt{n}\mathsf{D}(\mu_{0,n}, \mu_{1,n})^2 > q_{n,1-\alpha}$ as the estimated type 1 error and the fraction for which $\sqrt{n}\mathsf{D}(\mu_{0,n}, \mu_{2,n})^2 \leq q_{n,1-\alpha}$ as the estimated type 2 error. These results are compiled in Figure 3.2.

Figure 3.2 validates Theorem 17. Notably, even for a small number of samples,

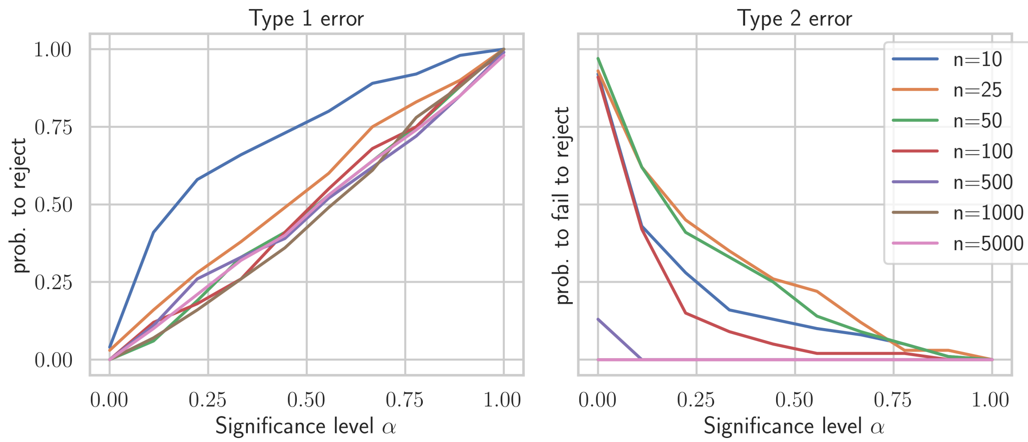


Figure 3.2: These plots compile the type 1 and type 2 error of the proposed test for varying numbers of samples and desired significance levels by following the methodology described in Section 3.7.1.

the graph of the estimated rejection probability as a function of the significance level approaches the graph of the identity function, as expected. In the low sample regime, $n = 10$, the type 1 error is significantly higher than expected whereas the type 2 error is lower than that for $n = 25$ and $n = 50$. This can be explained by the fact that Algorithm 3 is prone to overestimating the GW distance when few samples are available (see Section B.3.1 for additional discussion).

Graphs with finitely many weights

We now consider the case of weighted graphs whose edges may take on the values $\{0, 1, 2, 3\}$. Figure 3.3 depicts isomorphic distributions ν_0, ν_1 on this space of graphs along with a distribution, ν_2, which is identical to ν_1 except for $\rho_{2,12}$ which is set to equal $\rho_{1,13}$ so that ν_0 and ν_2 are not isomorphic. The methodology for estimating the type 1 and type 2 error of this statistic is identical to Section 3.7.1, up to the fact that we use the exhaustive search to estimate the GW distances (given that the relaxed graph matching approach is limited to unweighted graphs). The results are compiled in Figure

3.4, presenting similar trends to those discussed above with the notable exception that the performance of the test in the case $n = 10$ does not deviate heavily from the performance obtained with larger sample sizes.

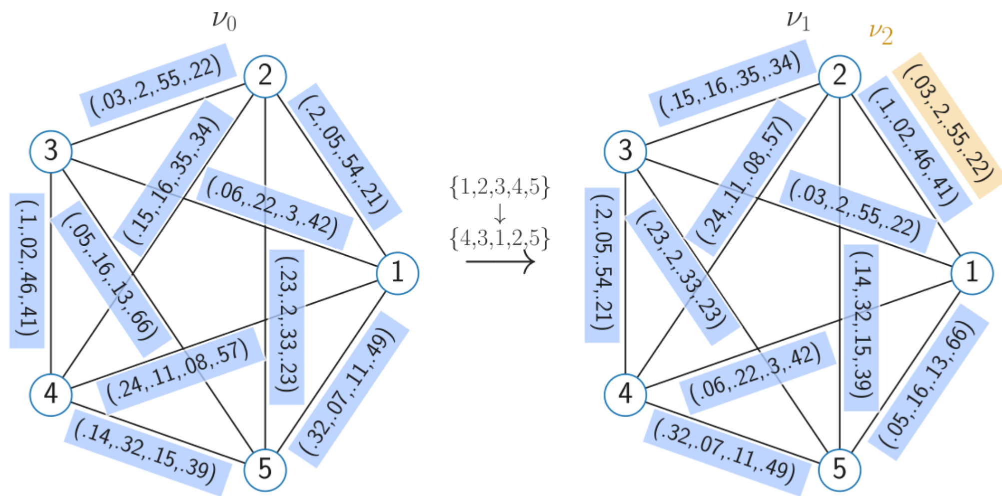


Figure 3.3: Distributions from Section 3.7.1. First, edge weight probabilities for ν_0 were sampled from the Dirichlet distribution, then a random permutation was chosen to generate ν_1 from ν_0 as depicted. Edge weight probabilities are given as vectors whose k-th entry is the probability that the edge takes the weight $k - 1$ (truncated to two significant figures).

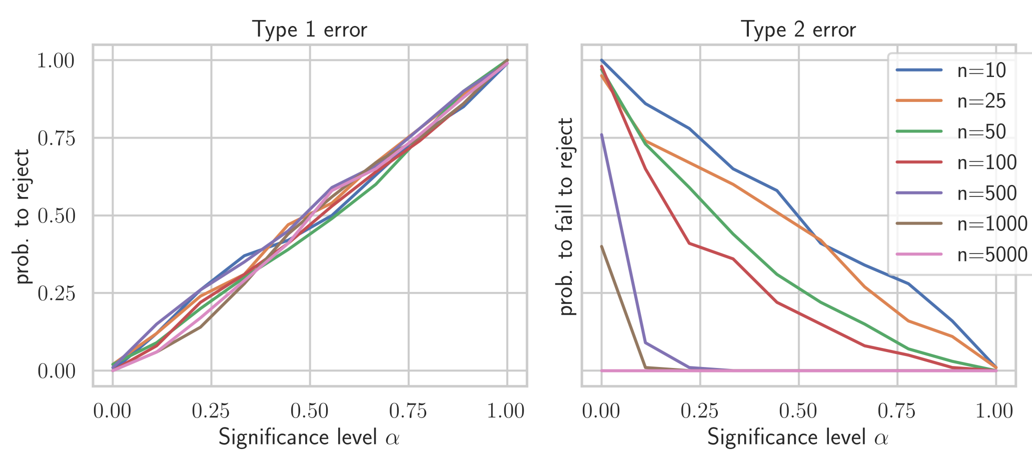


Figure 3.4: These plots compile the type 1 and type 2 error of the proposed test for varying numbers of samples and desired significance levels by following the methodology described in Section 3.7.1.

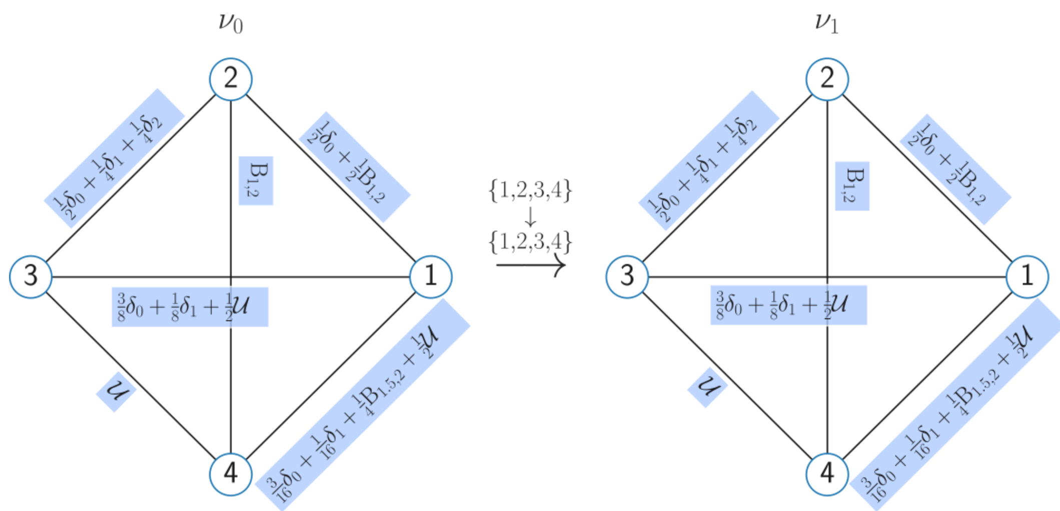


Figure 3.5: Distributions from Section 3.7.2. Here $\mathrm{B}_{a,b}$ denotes the beta distribution with parameters (a, b) whereas $\mathcal{U}$ is the uniform distribution on $[0, 1]$.

3.7.2 Simulating limit distributions

We next set to illustrate the structure of the limit distributions under the different settings studied in this paper.

Graphs with compactly supported weight distributions

This experiment explores the law of $\sqrt{n}\mathsf{D}(\mu_{0,n}, \mu_{1,n})^2$ when the underlying distributions on the edge weights are isomorphic. While Theorem 16 only accounts for finitely supported distributions, we go beyond that setting and consider here continuous, compactly supported distributions, as shown in Figure 3.5. The ground truth permutation relating these distributions is taken to be the identity, and, given n sampled graphs, we estimate $\mathsf{D}(\mu_{0,n}, \mu_{1,n})^2$ using the subgradient method (Algorithm 2) initialized at $\frac{1}{2}\int xx^\intercal d\bar{\mu}_{0,n}(x)$. To generate additional samples from $\mathsf{D}(\mu_{0,n}, \mu_{1,n})^2$, we draw a new set of n graphs and repeat this procedure. As the underlying distributions are isomorphic, $\mathsf{D}(\mu_0, \mu_1) = 0$ and there is no need to center the statistic.

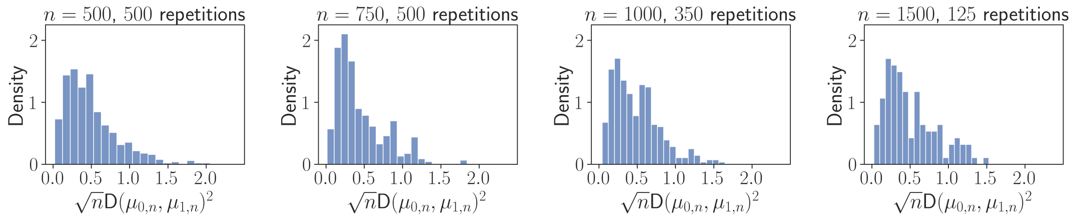


Figure 3.6: Estimated distribution of $\sqrt{n}\mathsf{D}(\mu_{0,n}, \mu_{1,n})^2$ for different values of n.

The results of this experiment, presented in Figure 3.6, appear to indicate that the distribution of $\sqrt{n}\mathsf{D}(\mu_{0,n}, \mu_{1,n})^2$ is stable as n increases, this finding indicates that the derived limit distribution results from Theorem 17 may extend to the compact case. Moreover, the mean rate of convergence provided in that result is in line with these simulations.

Entropic Gromov-Wasserstein distance

As noted in Theorem 16, $\mathsf{D}_\varepsilon(\hat{\mu}_{0,n}, \hat{\mu}_{1,n})^2$ is asymptotically normal under proper scaling and centering and the bootstrap is consistent once $\varepsilon > 16\sqrt{M_4(\bar{\mu}_0)M_4(\bar{\mu}_1)}$. To illustrate this finding numerically, we fix μ_0 to be the uniform distribution on $[-1, 1]^3$ and μ_1 to be the uniform distribution on the unit sphere in $\mathbb{R}^3$ and let $\varepsilon = 16.2\sqrt{\frac{19}{15}} > 16\sqrt{M_4(\bar{\mu}_0)M_4(\bar{\mu}_1)}$, where $\frac{19}{15}$ is the value of the product of the fourth moments. Then, we generate $n = 2500$ samples from both distributions to form the empirical measures $\hat{\mu}_{0,n}, \hat{\mu}_{1,n}$ and compute $\mathsf{D}_\varepsilon(\hat{\mu}_{0,n}, \hat{\mu}_{1,n})^2$ using Algorithm 1 in [164]. It is verified numerically that this choice of ε satisfies $\varepsilon > 16\sqrt{M_4(\bar{\mu}_{0,n})M_4(\bar{\mu}_{1,n})}$ for each computation so that Theorem 11 in [164] guarantees that a nearly optimal value is obtained using this procedure. Finally, the boostrapped empirical measures are constructed by sampling $X^B_{0,1}, \ldots, X^B_{0,n}$ from $\hat{\mu}_{0,n}$, $X^B_{1,1}, \ldots, X^B_{1,n}$ from $\hat{\mu}_{1,n}$, and setting $\hat{\mu}^B_{0,n} := \frac{1}{n}\sum_{i=1}^n \delta_{X^B_{0,i}}$, $\hat{\mu}^B_{1,n} := \frac{1}{n}\sum_{i=1}^n \delta_{X^B_{1,i}}$. The distribution of $\sqrt{n}\left(\mathsf{D}_\varepsilon(\hat{\mu}^B_{0,n}, \hat{\mu}^B_{1,n})^2 - \mathsf{D}_\varepsilon(\hat{\mu}_{0,n}, \hat{\mu}_{1,n})^2\right)$ for fixed $X_1, \ldots, X_n$ is then estimated from 1500 repetitions. The resulting histogram is presented in Figure 3.7 and is seen to

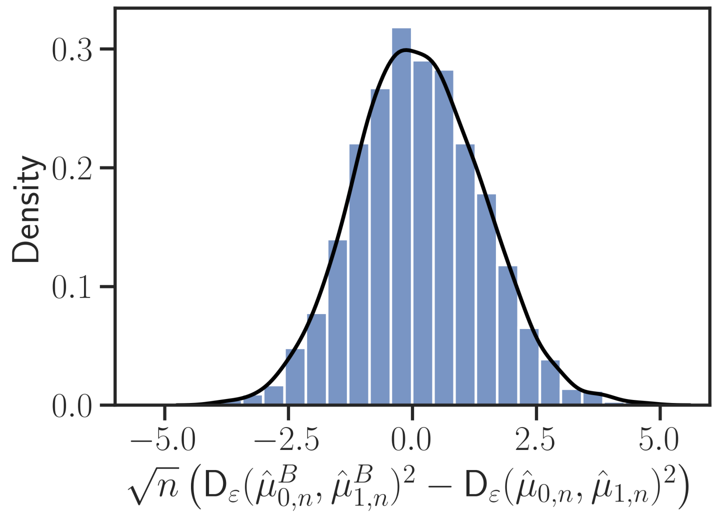


Figure 3.7: Histogram of the bootstrap estimator for the entropic GW distance limit distribution in the setting of Section 3.7.2 along with a kernel density estimator.

be approximately normal.

3.8 Proofs of Main Results

3.8.1 Proof of Theorem 8

To simplify the proof of Theorem 8, we separately prove each statement as its own lemma.

Lemma 4. $\Phi_{(\mu_0,\mu_1)}$ is Fréchet differentiable at $\mathbf{A}\in\mathbb{R}^{d_0\times d_1}$ with derivative

$$\left(D\Phi_{(\mu_0,\mu_1)}\right)_{[\mathbf{A}]}(\mathbf{B}) = 64\langle \mathbf{A}-\frac{1}{2}\int xy^\intercal d\pi_{\mathbf{A}}(x,y),\mathbf{B}\rangle_{\mathrm{F}}$$

provided all optimal couplings for $\mathsf{OT}_{\mathbf{A}}(\mu_0,\mu_1)$ admit the same cross-correlation matrix $\frac{1}{2}\int xy^\intercal d\pi_{\mathbf{A}}(x,y)$.

Proof. It is easy to see that $\|\cdot\|_{\mathrm{F}}^2$ is Fréchet differentiable at $\mathbf{A}$ with derivative $2\langle \mathbf{A}, \cdot\rangle_{\mathrm{F}}$. For any $\mathbf{H} \in \mathbb{R}^{d_0 \times d_1}$, we have that

$$\mathsf{OT}_{\mathbf{A}+\mathbf{H}}(\mu_0, \mu_1) - \mathsf{OT}_{\mathbf{A}}(\mu_0, \mu_1) \leq \int c_{\mathbf{A}+\mathbf{H}} d\pi_{\mathbf{A}} - \int c_{\mathbf{A}} d\pi_{\mathbf{A}} = -32 \int x^\intercal \mathbf{H} y d\pi_{\mathbf{A}}(x, y), \tag{3.16}$$

for any choice of OT plan $\pi_{\mathbf{A}}$ for $\mathsf{OT}_{\mathbf{A}}(\mu_0, \mu_1)$. Similarly,

$$\mathsf{OT}_{\mathbf{A}+\mathbf{H}}(\mu_0, \mu_1) - \mathsf{OT}_{\mathbf{A}}(\mu_0, \mu_1) \geq -32 \int x^\intercal \mathbf{H} y d\pi_{\mathbf{A}+\mathbf{H}}(x, y), \tag{3.17}$$

for any choice of optimal coupling $\pi_{\mathbf{A}+\mathbf{H}}$ for $\mathsf{OT}_{\mathbf{A}+\mathbf{H}}(\mu_0, \mu_1)$. Now, consider an arbitrary sequence $\mathbf{H}_n$ converging to 0. Note that

$$\begin{aligned}
\sup_{\substack{x \in \mathrm{spt}(\mu_0) \\ y \in \mathrm{spt}(\mu_1)}} \left| c_{\mathbf{A}+\mathbf{H}_n}(x, y) - c_{\mathbf{A}}(x, y) \right| &= \sup_{\substack{x \in \mathrm{spt}(\mu_0) \\ y \in \mathrm{spt}(\mu_1)}} |32 x^\intercal \mathbf{H}_n y| \\
&\leq 32 \sup_{\mathrm{spt}(\mu_0)} \|\cdot\| \sup_{\mathrm{spt}(\mu_1)} \|\cdot\| \|\mathbf{H}_n\|_{\mathrm{F}} \to 0,
\end{aligned}$$

hence $c_{\mathbf{A}+\mathbf{H}_n} \to c_{\mathbf{A}}$ uniformly on $\mathrm{spt}(\mu_0) \times \mathrm{spt}(\mu_1)$. It follows from Theorem 5.20 in [208] that, for any subsequence n' of n there exists a further subsequence n'' along which $\pi_{\mathbf{A}+\mathbf{H}_{n''}} \xrightarrow{w} \pi$ for some optimal coupling π for $\mathsf{OT}_{\mathbf{A}}(\mu_0, \mu_1)$. Thus $\int xy^\intercal d\pi_{\mathbf{A}+\mathbf{H}_{n''}}(x, y) \to \int xy^\intercal d\pi(x, y) = \int xy^\intercal d\pi_{\mathbf{A}}(x, y)$ by assumption. As the limit is the same regardless of the choice of subsequence, conclude that $\int xy^\intercal d\pi_{\mathbf{A}+\mathbf{H}}(x, y) \to \int xy^\intercal d\pi_{\mathbf{A}}(x, y)$ as $\mathbf{H} \to 0$, thus

$$\begin{aligned}
&\|\mathbf{H}\|_F^{-1} \left| \mathsf{OT}_{\mathbf{A}+\mathbf{H}}(\mu_0, \mu_1) - \mathsf{OT}_{\mathbf{A}}(\mu_0, \mu_1) + 32 \int x^\intercal \mathbf{H} y d\pi_{\mathbf{A}}(x, y) \right| \\
&\leq 32 \left\| \int xy^\intercal d\pi_{\mathbf{A}+\mathbf{H}}(x, y) - \int xy^\intercal d\pi_{\mathbf{A}}(x, y) \right\|_{\mathrm{F}} \to 0,
\end{aligned}$$

which proves the claim. □

Lemma 5. For any $\pi \in \Pi(\mu_0, \mu_1)$, $\int xy^\intercal d\pi(x, y) \in B_{\mathrm{F}}(\sqrt{M_2(\mu_0) M_2(\mu_1)})$.

Proof. By Jensen's inequality,

$$\left\| \int xy^\intercal d\pi(x, y) \right\|_{\mathrm{F}} \leq \int \|xy^\intercal\|_{\mathrm{F}} \, d\pi(x, y) = \int \|x\| \|y\| d\pi(x, y).$$

This final term is bounded above by $\sqrt{\int \|x\|^2 d\pi(x,y) \int \|y\|^2 d\pi(x,y)} = \sqrt{M_2(\mu_0)M_2(\mu_1)}$ by the Cauchy-Schwarz inequality. $\square$

Lemma 6. $\Phi_{(\mu_0,\mu_1)}$ is locally Lipschitz continuous and coercive.

Proof. Fix a compact set $K \subset \mathbb{R}^{d_0 \times d_1}$. For any $\mathbf{A}, \mathbf{A}' \in K$, it follows from (3.16) and (3.17) that,

$$\begin{aligned}
\mathsf{OT}_{\mathbf{A}'}(\mu_0,\mu_1) - \mathsf{OT}_{\mathbf{A}}(\mu_0,\mu_1) &\le -32 \int x^\intercal(\mathbf{A}' - \mathbf{A}) y \, d\pi_{\mathbf{A}}(x,y) \\
&\le 32\|\mathbf{A}' - \mathbf{A}\|_{\mathrm{F}} \left\| \int xy^\intercal d\pi_{\mathbf{A}}(x,y) \right\|_{\mathrm{F}}, \\
\mathsf{OT}_{\mathbf{A}'}(\mu_0,\mu_1) - \mathsf{OT}_{\mathbf{A}}(\mu_0,\mu_1) &\ge -32\|\mathbf{A}' - \mathbf{A}\|_{\mathrm{F}} \left\| \int xy^\intercal d\pi_{\mathbf{A}'}(x,y) \right\|_{\mathrm{F}},
\end{aligned}$$

that is,

$$\begin{aligned}
&|\mathsf{OT}_{\mathbf{A}'}(\mu_0,\mu_1) - \mathsf{OT}_{\mathbf{A}}(\mu_0,\mu_1)| \\
&\qquad \le 32\|\mathbf{A}' - \mathbf{A}\|_{\mathrm{F}} \left(\left\| \int xy^\intercal d\pi_{\mathbf{A}'}(x,y) \right\|_{\mathrm{F}} \bigvee \left\| \int xy^\intercal d\pi_{\mathbf{A}}(x,y) \right\|_{\mathrm{F}} \right).
\end{aligned}$$

Applying Lemma 5, we obtain

$$|\mathsf{OT}_{\mathbf{A}'}(\mu_0,\mu_1) - \mathsf{OT}_{\mathbf{A}}(\mu_0,\mu_1)| \le 32\sqrt{M_2(\mu_0)M_2(\mu_1)}\|\mathbf{A}' - \mathbf{A}\|_{\mathrm{F}}.$$

Further, $\left|\|\mathbf{A}\|_{\mathrm{F}}^2 - \|\mathbf{A}'\|_{\mathrm{F}}^2\right| = (\|\mathbf{A}\|_{\mathrm{F}} + \|\mathbf{A}'\|_{\mathrm{F}}) \left|\|\mathbf{A}\|_{\mathrm{F}} - \|\mathbf{A}'\|_{\mathrm{F}}\right| \le 2\sup_K \|\cdot\|_{\mathrm{F}} \|\mathbf{A} - \mathbf{A}'\|_{\mathrm{F}}$.

To show coercivity, observe that, for any $\mathbf{A} \in \mathbb{R}^{d_0 \times d_1}$ and $\pi \in \Pi(\mu_0,\mu_1)$,

$$\int -4\|x\|^2\|y\|^2 - 32x^\intercal \mathbf{A} y \, d\pi(x,y) \ge -4\sqrt{M_4(\mu_0)M_4(\mu_1)} - 32\sqrt{M_2(\mu_0)M_2(\mu_1)}\|\mathbf{A}\|_{\mathrm{F}}$$

Hence $32\|\mathbf{A}\|_{\mathrm{F}} + \frac{\mathsf{OT}_{\mathbf{A}}(\mu_0,\mu_1)}{\|\mathbf{A}\|_{\mathrm{F}}} \to \infty$ as $\|\mathbf{A}\|_{\mathrm{F}} \to \infty$ proving coercivity. $\square$

Lemma 7. The Clarke subdifferential of $\Phi_{(\mu_0,\mu_1)}$ at $\mathbf{A} \in \mathbb{R}^{d_0 \times d_1}$ is given by

$$\partial\Phi_{(\mu_0,\mu_1)}(\mathbf{A}) = \left\{ 64\mathbf{A} - 32\int xy^\intercal d\pi_{\mathbf{A}}(x,y) : \pi_{\mathbf{A}} \text{ is an OT plan for } \mathsf{OT}_{\mathbf{A}}(\mu_0,\mu_1) \right\}. \quad (3.18)$$

Proof. We first establish that $\partial\Phi_{(\mu_0,\mu_1)}$ is a subset of the right hand side in (3.18). By Rademacher's theorem, local Lipschitz continuity of $\Phi_{(\mu_0,\mu_1)}$ (Lemma 6) guarantees that it is differentiable on a set Λ of full measure and, by Theorem 8.1 in [48], the Clarke subdifferential of $\Phi_{(\mu_0,\mu_1)}$ at $\mathbf{A} \in \mathbb{R}^{d_0\times d_1}$ can be defined as

$$\partial\Phi_{(\mu_0,\mu_1)}(\mathbf{A}) = \text{conv}\left(\left\{\lim_{\Omega\ni\mathbf{A}_n\to\mathbf{A}}(D\Phi_{(\mu_0,\mu_1)})_{[\mathbf{A}_n]}\right\}\right)$$

where Ω is any subset of Λ for which $\Lambda\backslash\Omega$ is negligible and it is presupposed that the limit converges (here conv denotes the convex hull operation on a set). From Lemma 4, $(D\Phi_{(\mu_0,\mu_1)})_{[\mathbf{A}_n]}$ can be identified with $64\left(\mathbf{A}_n - \frac{1}{2}\int xy^\intercal d\pi_{\mathbf{A}_n}(x,y)\right)$ where $\pi_{\mathbf{A}_n}$ is any OT plan for $\mathsf{OT}_{\mathbf{A}_n}(\mu_0,\mu_1)$ (by assumption all such cross-correlation matrices are identical). As $\mathbf{A}_n \to \mathbf{A}$, it follows from the proof of Lemma 4 that $\pi_{\mathbf{A}_n} \xrightarrow{w} \pi_{\mathbf{A}}$ up to a subsequence, where $\pi_{\mathbf{A}}$ is some OT plan for $\mathsf{OT}_{\mathbf{A}}(\mu_0,\mu_1)$. Thus, $\partial\Phi_{(\mu_0,\mu_1)}(\mathbf{A}) = 64\mathbf{A} - 32\,\text{conv}\left(\left\{\int xy^\intercal d\pi_{\mathbf{A}}(x,y) : \exists\, \pi_{\mathbf{A}_n} \xrightarrow{w} \pi_{\mathbf{A}}, \mathbf{A}_n \in \Omega\right\}\right)$. If $\mathbf{A} \in \Lambda$, $\partial\Phi_{(\mu_0,\mu_1)}(\mathbf{A}) = D\left(\Phi_{(\mu_0,\mu_1)}\right)_{[\mathbf{A}]}$, whereas if $\mathbf{A} \notin \Lambda$, the previous convex hull is a subset of the set of all cross-correlation matrices for some OT plan for $\mathsf{OT}_{\mathbf{A}}(\mu_0,\mu_1)$, noting that any convex combination of OT plans is also an OT plan.

Now, we show that $\partial\Phi_{(\mu_0,\mu_1)}$ is a superset of the right hand side in (3.18). To this end, note that $\mathbf{A} \in \mathbb{R}^{d_0\times d_1} \mapsto -\mathsf{OT}_{\mathbf{A}}(\mu_0,\mu_1)$ is convex by Theorem 5.5 in [167] as it can be expressed as the pointwise supremum of a collection of linear functions of $\mathbf{A}$. By (3.16), it follows that

$$-\mathsf{OT}_{\mathbf{A}'}(\mu_0,\mu_1) - (-\mathsf{OT}_{\mathbf{A}}(\mu_0,\mu_1)) \geq \left\langle 32\int xy^\intercal d\pi_{\mathbf{A}}(x,y), \mathbf{A}' - \mathbf{A}\right\rangle_{\mathrm{F}},$$

for any $\mathbf{A}' \in \mathbb{R}^{d_0\times d_1}$ and any choice of OT plan, $\pi_{\mathbf{A}}$, for $\mathsf{OT}_{\mathbf{A}}(\mu_0,\mu_1)$ so that the subdifferential of $-\mathsf{OT}_{(\cdot)}(\mu_0,\mu_1)$ at $\mathbf{A}$ (in the sense of convex analysis) contains all matrices of the form $32\int xy^\intercal d\pi_{\mathbf{A}}(x,y)$. Note, however, that the Clarke subdifferential coincides with the (convex) subdifferential in this context (see Proposition 2.2.7 in [50]) and that

$\partial\Phi_{(\mu_0,\mu_1)}(\mathbf{A}) = 64\mathbf{A} - \partial(-\mathsf{OT}_{(\cdot)}(\mu_0,\nu_1))(\mathbf{A})$ by Proposition 2.3.1 and Corollary 1 on p.39 of [50], proving the complementary inclusion. $\square$

We now prove point 2, concluding the proof of Theorem 8.

Proof of Theorem 8 (2). By Proposition 2.3.2 in [50], if $\bar{\mathbf{A}}$ is a local minimizer for $\Phi_{(\mu_0,\mu_1)}$, then $0 \in \partial\Phi_{(\mu_0,\mu_1)}(\bar{\mathbf{A}})$, characterized in Lemma 7. Thus, there exists an OT plan $\pi_{\bar{\mathbf{A}}}$ for $\mathsf{OT}_{\bar{\mathbf{A}}}(\mu_0,\mu_1)$ satisfying $2\bar{\mathbf{A}} = \int xy^\intercal d\pi_{\bar{\mathbf{A}}}(x,y) \in B_{\mathrm{F}}(\sqrt{M_2(\mu_0)M_2(\mu_1)})$ by Lemma 5. By coercivity, proved in Lemma 6, at least one local minimizer is globally optimal.

Now, assume that μ_0,μ_1 are centered. It is straightforward to see that if $\mathbf{A}^\star$ solves (3.9), then the associated OT plan $\pi_{\mathbf{A}^\star}$ satisfies

$$\mathsf{S}_1(\mu_0,\mu_1) + \mathsf{S}_2(\mu_0,\mu_1) = \iint \left|\|x-x'\|^2 - \|y-y'\|^2\right|^2 d\pi_{\mathbf{A}^\star} \otimes \pi_{\mathbf{A}^\star}(x,y,x',y'),$$

such that $\pi_{\mathbf{A}^\star}$ is optimal for (3.7) (see Section 5.1 in [223] for details). $\square$

3.8.2 Proofs for Section 3.3

Proof of Theorem 9

We first show that the OT potentials for $\mathsf{OT}_{\mathrm{A}}(\nu_0,\nu_1)$ can be chosen as to satisfy uniform bounds for any choice of $(\nu_0,\nu_1) \in \mathcal{P}_{\mu_0} \times \mathcal{P}_{\mu_1}$ and $\mathbf{A} \in B_{\mathrm{F}}(M)$.

Lemma 8. For any $(\nu_0,\nu_1) \in \mathcal{P}_{\mu_0} \times \mathcal{P}_{\mu_1}$, $\mathbf{A} \in B_{\mathrm{F}}(M)$, and any choice of OT potentials (φ_0,φ_1) for $\mathsf{OT}_{\mathbf{A}}(\nu_0,\nu_1)$, there exists a version, $(\tilde{\varphi}_0,\tilde{\varphi}_1)$, of (φ_0,φ_1) satisfying $\|\tilde{\varphi}_0\|_{\infty,\mathcal{X}_0} \vee \|\tilde{\varphi}_1\|_{\infty,\mathcal{X}_1} \leq K$ where K depends only on $\|\mathcal{X}_0\|_\infty$ and $\|\mathcal{X}_1\|_\infty$.

Proof. In the discrete setting, $\mathsf{OT}_{\mathbf{A}}(\nu_0,\nu_1)$ can be identified with a finite dimensional linear program. By the complementary slackness conditions, any pair OT potentials

(φ_0, φ_1) for $\mathsf{OT}_{\mathbf{A}}(\nu_0, \nu_1)$ satisfies $\varphi_0(x)+\varphi_1(y) = c_{\mathbf{A}}(x, y)$ at all points $(x, y) \in \mathcal{X}_0\times\mathcal{X}_1$ where $\pi(\{(x, y)\}) > 0$ for any choice of OT plan π for $\mathsf{OT}_{\mathbf{A}}(\nu_0, \nu_1)$. As $\pi \in \Pi(\nu_0, \nu_1)$ and $\mathrm{spt}(\nu_i) = \mathcal{X}_i$ for $i = 0, 1$, for every $x \in \mathcal{X}_0$ there exists $y_x \in \mathcal{X}_1$ for which $\varphi_0(x) + \varphi_1(y_x) = c_{\mathbf{A}}(x, y_x)$ and, similarly, for every $y \in \mathcal{X}_1$ there exists $x_y \in \mathcal{X}_0$ for which $\varphi_0(x_y) + \varphi_1(y) = c_{\mathbf{A}}(x_y, y)$. Moreover, $y_x \in \mathrm{argmin}_{\mathcal{X}_1} \{c_{\mathbf{A}}(x, \cdot) - \varphi_1\}$ and $x_y \in \mathrm{argmin}_{\mathcal{X}_0} \{c_{\mathbf{A}}(\cdot, y) - \varphi_0\}$, as $\varphi_0\oplus\varphi_1 \leq c_{\mathbf{A}}$. Whence,

$$-\|c_{\mathbf{A}}\|_{\infty,\mathcal{X}_0\times\mathcal{X}_1} - \sup_{\mathcal{X}_1}\varphi_1 \leq \varphi_0(x) = \inf_{\mathcal{X}_1}\{c_{\mathbf{A}}(x, \cdot) - \varphi_1\} \leq \|c_{\mathbf{A}}\|_{\infty,\mathcal{X}_0\times\mathcal{X}_1} - \sup_{\mathcal{X}_1}\varphi_1,$$
$$-\|c_{\mathbf{A}}\|_{\infty,\mathcal{X}_0\times\mathcal{X}_1} - \sup_{\mathcal{X}_0}\varphi_0 \leq \varphi_1(y) = \inf_{\mathcal{X}_0}\{c_{\mathbf{A}}(\cdot, y) - \varphi_0\} \leq \|c_{\mathbf{A}}\|_{\infty,\mathcal{X}_0\times\mathcal{X}_1} - \sup_{\mathcal{X}_0}\varphi_0.$$

Let $(\tilde\varphi_0, \tilde\varphi_1) = (\varphi_0 + C, \varphi_1 - C)$ for $C = -\sup_{\mathcal{X}_0}\varphi_0 + \|c_{\mathbf{A}}\|_{\infty,\mathcal{X}_0\times\mathcal{X}_1}$ such that $\sup_{\mathcal{X}_0}\tilde\varphi_0 = \|c_{\mathbf{A}}\|_{\infty,\mathcal{X}_0\times\mathcal{X}_1}$. From the prior display, we have $\|\tilde\varphi_1\|_{\infty,\mathcal{X}_1} \leq 2\|c_{\mathbf{A}}\|_{\infty,\mathcal{X}_0\times\mathcal{X}_1}$, $\|\tilde\varphi_0\|_{\infty,\mathcal{X}_0} \leq \|c_{\mathbf{A}}\|_{\infty,\mathcal{X}_0\times\mathcal{X}_1}$.

It suffices, therefore, to bound $\|c_{\mathbf{A}}\|_{\infty,\mathcal{X}_0\times\mathcal{X}_1}$. To this end, for any $(x, y) \in \mathcal{X}_0 \times \mathcal{X}_1$,

$$|c_{\mathbf{A}}(x, y)| = |-4\|x\|^2\|y\|^2 - 32x^\intercal\mathbf{A}y| \leq 4\|\mathcal{X}_0\|_\infty^2\|\mathcal{X}_1\|_\infty^2 + 32\|\mathcal{X}_0\|_\infty\|\mathcal{X}_1\|_\infty M, \tag{3.19}$$

where the inequality is due to the triangle inequality and the fact that $|x^\intercal\mathbf{A}y| = |\langle xy^\intercal, \mathbf{A}\rangle_{\mathrm{F}}| \leq \|xy^\intercal\|_{\mathrm{F}}\|\mathbf{A}\|_{\mathrm{F}} = \|x\|\|y\|\|\mathbf{A}\|_{\mathrm{F}} \leq \|\mathcal{X}_0\|_\infty\|\mathcal{X}_1\|_\infty M$ as $\mathbf{A} \in B_{\mathrm{F}}(M)$. □

Next, we address the failure of the directional regularity condition from [27]. In what follows, we assume without loss of generality that μ_0 and μ_1 are centered. In the case that μ_0, μ_1 are not centered, observe that the limit in (3.22) can be written as

$$\lim_{t\downarrow 0}\frac{\mathsf{D}(\mu_{0,t}, \mu_{1,t})^2 - \mathsf{D}(\mu_0, \mu_1)^2}{t}$$
$$= \lim_{t\downarrow 0}\frac{\mathsf{D}((\mathrm{Id} - \mathbb{E}_{\mu_0}[X])_\sharp\mu_{0,t}, (\mathrm{Id} - \mathbb{E}_{\mu_1}[X])_\sharp\mu_{1,t})^2 - \mathsf{D}(\bar\mu_0, \bar\mu_1)^2}{t},$$

by translation invariance of the Gromov-Wasserstein distance. As the subsequent arguments only require that the perturbed measures and the base measures are supported on the same points and that the base measures are centered, we may apply the same result to the limit on the right hand side to prove the general claim.

Our proof technique is inspired by the arguments presented in Section 4.3.2 of [27]. However, we highlight that the results contained therein are not applicable to the current setting as the so-called directional regularity condition does not hold.

Remark 19 (Failure of directional regularity)**.** Consider the perturbed problem

$$\begin{aligned} \inf_{\mathbf{P}\in\mathbb{R}^{N_0\times N_1}} \quad & \langle \mathbf{P}, \Delta(\mathbf{P})\rangle_{\mathrm{F}}\,, \\ \text{s.t.} \quad & \mathbf{P}\mathbb{1}_{N_1} = m_{0,t}, \\ & \mathbf{P}^{\intercal}\mathbb{1}_{N_0} = m_{1,t}, \\ & \mathbf{P}_{ij} \geq 0,\ \forall (i,j) \in [N_0]\times[N_1], \end{aligned} \tag{3.20}$$

where $(m_{0,t}, m_{1,t}) \in \mathbb{R}^{N_0}\times\mathbb{R}^{N_1}$ are the weight vectors for $\mu_{0,t}$ and $\mu_{1,t}$ for $t \in [0,1]$. To establish stability of the optimal value of (3.20) at $t = 0$, Theorem 4.24 in [27] can be applied provided that the directional regularity condition holds.

According to Theorem 4.9 in [27], the directional regularity condition for the problem (3.20) holds at a point $\mathbf{P} \in \mathbb{R}^{N_0\times N_1}$ which is feasible for (3.20) at $t = 0$ in the direction $d = (n_0 - m_0, n_1 - m_1, 0)$ where n_0, n_1 are the weight vectors for ν_0, ν_1 if and only if

$$0 \in \operatorname{int}\Big(G(\mathbf{P},(m_0,m_1)) + DG_{[(\mathbf{P},(m_0,m_1))]}(\mathbb{R}^{N_0\times N_1}, \{t(n_0-m_0, n_1-m_1) : t\geq 0\}) - K\Big), \tag{3.21}$$

where $G(\mathbf{Q},(r_0,r_1)) = (\mathbf{Q}\mathbb{1}_{N_1} - r_0, \mathbf{Q}^{\intercal}\mathbb{1}_{N_0} - r_1, \mathbf{Q})$ and $K = \{0\}\times\{0\}\times[0,\infty)^{N_0\times N_1}$. It is easy to see that

$$\begin{aligned} & DG_{[(\mathbf{P},(m_0,m_1))]}(\mathbb{R}^{N_0\times N_1}, \{t(n_0-m_0, n_1-m_1) : t\geq 0\}) \\ & \qquad = \Big\{(\mathbf{Q}\mathbb{1}_{N_1} - t(n_0-m_0), \mathbf{Q}^{\intercal}\mathbb{1}_{N_0} - t(n_1-m_1), \mathbf{Q}) : \mathbf{Q}\in\mathbb{R}^{N_0\times N_1}, t\geq 0\Big\}, \end{aligned}$$

so that condition (3.21) reads

$$\begin{aligned} 0 \in \operatorname{int}\big(\{(\mathbf{Q}\mathbb{1}_{N_1} - t(n_0-m_0), \mathbf{Q}^{\intercal}\mathbb{1}_{N_0} - t(n_1-m_1), \mathbf{P}+\mathbf{Q}-\mathbf{S}) \\ : \mathbf{Q}\in\mathbb{R}^{N_0\times N_1}, \mathbf{S}\in[0,\infty)^{N_0\times N_1}, t\geq 0\big\}\big). \end{aligned}$$

We remark that the first and second coordinates of any triple in the above set satisfy

$$\mathbb{1}_{N_1}^\intercal \left(\mathbf{Q}\mathbb{1}_{N_1} - t(n_0 - m_0)\right) = \mathbb{1}_{N_1}^\intercal \mathbf{Q}\mathbb{1}_{N_0} = \mathbb{1}_{N_1}^\intercal \mathbf{Q}^\intercal \mathbb{1}_{N_0} = \mathbb{1}_{N_1}^\intercal \left(\mathbf{Q}^\intercal \mathbb{1}_{N_0} - t(n_1 - m_1)\right),$$

as $n_0 - m_0$ and $n_1 - m_1$ have total mass zero. It follows that this set has empty interior and thus condition (3.21) fails.

For ease of presentation, we separate the proof of Theorem 9 into a number of lemmas.

Lemma 9. Fix $(\nu_0, \nu_1) \in \mathcal{P}_{\mu_0} \times \mathcal{P}_{\mu_1}$ and let $\rho_i = \nu_i - \mu_i$, $\mu_{i,t} = \mu_i + t\rho_i$ for $t \in [0, 1]$ and $i \in \{0, 1\}$. Then,

$$\frac{d}{dt}\mathsf{D}(\mu_{0,t}, \mu_{1,t})^2\Big|_{t=0} = \inf_{\mathbf{A}\in\mathcal{A}} \inf_{\pi\in\bar{\Pi}^\star_{\mathbf{A}}} \sup_{(\varphi_0,\varphi_1)\in\bar{\mathcal{D}}_{\mathbf{A}}} \left\{ \int (f_0 + g_{0,\pi} + \bar{\varphi}_0) d\rho_0 + \int (f_1 + g_{1,\pi} + \bar{\varphi}_1) d\rho_1 \right\}, \tag{3.22}$$

where

$$\begin{aligned} f_0 &= 2\int \|\cdot - x\|^4 d\mu_0(x) - 4M_2(\bar{\mu}_1)\|\cdot - \mathbb{E}_{\mu_0}[X]\|^2, \\ f_1 &= 2\int \|\cdot - y\|^4 d\mu_0(y) - 4M_2(\bar{\mu}_0)\|\cdot - \mathbb{E}_{\mu_1}[X]\|^2, \\ g_{0,\pi} &= 8\mathbb{E}_{(X,Y)\sim\pi}[\|Y\|^2 X]^\intercal(\cdot), \\ g_{1,\pi} &= 8\mathbb{E}_{(X,Y)\sim\pi}[\|X\|^2 Y]^\intercal(\cdot), \end{aligned}$$

$\bar{\varphi}_i = \varphi_i(\cdot - \mathbb{E}_{\mu_i}[X])$ for $i = 0, 1$, and, for $\mathbf{A} \in B_{\mathrm{F}}(M)$, $\bar{\Pi}^\star_{\mathbf{A}}$ is the set of all OT plans for $\mathsf{OT}_{\mathbf{A}}(\bar{\mu}_0, \bar{\mu}_1)$, and $\bar{\mathcal{D}}_{\mathbf{A}}$ is the set of all corresponding OT potentials.

Proof. Assume without loss of generality that μ_0 and μ_1 are centered. We begin by establishing an upper bound on the limit. To this end, let $\bar{\pi} \in \Pi(\mu_0, \mu_1)$ be an arbitrary GW plan for $\mathsf{D}(\mu_0, \mu_1)$, and let γ be a finite signed measure on $\mathcal{X}_0 \times \mathcal{X}_1$ with first marginal $\nu_0 - \mu_0$, second marginal $\nu_1 - \mu_1$, and with $\gamma(\{(x^{(i)}, y^{(j)})\}) \geq 0$ for every $(i, j) \in$

$\mathcal{I}(\bar{\pi}) := \left\{(i,j) \in [N_0] \times [N_1] : \bar{\pi}(\{(x^{(i)}, y^{(j)})\}) = 0\right\}$. With this, $\bar{\pi} + t\gamma \in \Pi(\mu_{0,t}, \mu_{1,t})$ for every $t \in [0,1]$ sufficiently small, whence

$$\begin{aligned}\mathsf{D}(\mu_{0,t},\mu_{1,t})^2 - \mathsf{D}(\mu_0,\mu_1)^2 &\leq \int \Delta d\,(\bar{\pi}+t\gamma)\otimes(\bar{\pi}+t\gamma) - \int \Delta d\bar{\pi}\otimes\bar{\pi} \\ &= 2t\int \Delta d\bar{\pi}\otimes\gamma + t^2 \iint \Delta d\gamma\otimes\gamma,\end{aligned} \tag{3.23}$$

for all such t.

We now tighten the right hand side of (3.23) by optimizing over the choice of γ in the first term of the last line. To this end, define $\mathbf{C} \in \mathbb{R}^{N_0 \times N_1}$ as $\mathbf{C}_{ij} = 2\int \Delta\left(x^{(i)}, y^{(j)}, x, y\right) d\bar{\pi}(x,y)$ for $(i,j) \in [N_0] \times [N_1]$ and consider the linear program

$$\begin{aligned} \inf_{\mathbf{G}\in\mathbb{R}^{N_0\times N_1}} \quad & \langle \mathbf{C}, \mathbf{G}\rangle_{\mathrm{F}} \\ \text{subject to} \quad & \mathbf{G}\mathbb{1}_{N_1} = n_0 - m_0, \\ & \mathbf{G}^\intercal \mathbb{1}_{N_0} = n_1 - m_1, \\ & \mathbf{G}_{ij} \geq 0, \ \forall\, (i,j) \in \mathcal{I}(\bar{\pi}), \end{aligned} \tag{3.24}$$

where $\mathbb{1}_{N_i}$ is the vector of all 1's of length N_i for $i = 0, 1$, and the vectors $n_0, m_0 \in \mathbb{R}^{N_0}$, $n_1, m_1 \in \mathbb{R}^{N_1}$ contain the weights of ν_0, μ_0, ν_1, and μ_1 respectively. The corresponding dual problem is given by

$$\begin{aligned} \sup_{\substack{y\in\mathbb{R}^{N_0}, z\in\mathbb{R}^{N_1} \\ (S_{ij})_{(i,j)\in\mathcal{I}(\bar{\pi})}}} \quad & y^\intercal(n_0 - m_0) + z^\intercal(n_1 - m_1) \\ \text{subject to} \quad & y_i + z_j = \mathbf{C}_{ij}, \ \forall\, (i,j) \notin \mathcal{I}(\bar{\pi}), \\ & y_i + z_j + S_{ij} = \mathbf{C}_{ij}, \ \forall\, (i,j) \in \mathcal{I}(\bar{\pi}), \\ & S_{ij} \geq 0. \end{aligned} \tag{3.25}$$

Since μ_0, μ_1 are centered we have, for $(i, j) \in [N_0] \times [N_1]$, $\mathbf{C}_{ij}$ is given by

$$
\begin{aligned}
&2\int \|x^{(i)} - x\|^4 - 2\|x^{(i)} - x\|^2\|y^{(j)} - y\|^2 + \|y^{(j)} - y\|^4 d\bar{\pi}(x, y) \\
&= 2\int \|x^{(i)} - x\|^4 d\mu_0(x) + 2\int \|y^{(j)} - y\|^4 d\mu_1(y) - 4M_2(\mu_1)\|x^{(i)}\|^2 - 4M_2(\mu_0)\|y^{(j)}\|^2 \\
&+ 8\int \langle x^{(i)}, x\rangle \|y\|^2 d\bar{\pi}(x, y) + 8\int \|x\|^2\langle y^{(j)}, y\rangle d\bar{\pi}(x, y) - 4\int \|y\|^2\|x\|^2 d\bar{\pi}(x, y) \\
&- 4\|x^{(i)}\|^2\|y^{(j)}\|^2 - 16\int \langle x^{(i)}, x\rangle\langle y^{(j)}, y\rangle d\bar{\pi}(x, y),
\end{aligned}
$$

where only the final two terms in the display depend simultaneously on $x^{(i)}$ and $y^{(j)}$. Adopting the auxiliary variables $r \in \mathbb{R}^{N_0}$, $s \in \mathbb{R}^{N_1}$ given by

$$
\begin{aligned}
r_i &= y_i - 2\int \|x^{(i)} - x\|^4 d\mu_0(x) - 8\int \langle x^{(i)}, x\rangle \|y\|^2 d\bar{\pi}(x, y) + 4M_2(\mu_1)\|x^{(i)}\|^2 \\
s_j &= z_j - 2\int \|y^{(j)} - y\|^4 d\mu_1(y) - 8\int \|x\|^2\langle y^{(j)}, y\rangle d\bar{\pi}(x, y) + 4M_2(\mu_0)\|y^{(j)}\|^2 \\
&\quad + 4\int \|y\|^2\|x\|^2 d\bar{\pi}(x, y)
\end{aligned} \tag{3.26}
$$

for every $(i, j) \in [N_0] \times [N_1]$, we obtain the equivalent dual problem

$$
\begin{aligned}
\sup_{\substack{r\in\mathbb{R}^{N_0}, s\in\mathbb{R}^{N_1} \\ (S_{ij})_{(i,j)\in\mathcal{I}(\bar{\pi})}}} \quad & r^\intercal(n_0 - m_0) + s^\intercal(n_1 - m_1) + \varsigma_{\bar{\pi}} \\
\text{subject to} \quad & r_i + s_j = \mathbf{R}_{ij}, \ \forall\ (i, j) \notin \mathcal{I}(\bar{\pi}), \\
& r_i + s_j + S_{ij} = \mathbf{R}_{ij}, \ \forall\ (i, j) \in \mathcal{I}(\bar{\pi}), \\
& S_{ij} \geq 0,
\end{aligned} \tag{3.27}
$$

where $\mathbf{R}_{ij} = -4\|x^{(i)}\|^2\|y^{(j)}\|^2 - 16\int \langle x^{(i)}, x\rangle\langle y^{(j)}, y\rangle d\bar{\pi}(x, y)$ for $(i, j) \in [N_0] \times [N_1]$, and

$$
\begin{aligned}
&\varsigma_{\bar{\pi}} = \\
&\sum_{i=1}^{N_0}\left(2\int \|x^{(i)} - x\|^4 d\mu_0(x) + 8\int \langle x^{(i)}, x\rangle \|y\|^2 d\bar{\pi}(x, y) - 4M_2(\mu_1)\|x^{(i)}\|^2\right)(n_0 - m_0)_i + \\
&\sum_{j=1}^{N_1}\left(2\int \|y^{(j)} - y\|^4 d\mu_1(y) + 8\int \|x\|^2\langle y^{(j)}, y\rangle d\bar{\pi}(x, y) - 4M_2(\mu_0)\|y^{(j)}\|^2\right)(n_1 - m_1)_j\,,
\end{aligned}
$$

the term involving $4\int \|y\|^2\|x\|^2 d\bar{\pi}(x, y)$ is omitted, as $n_1 - m_1$ sums to 0.

If (r,s) is feasible for (3.27), $r_i + s_j \le \mathbf{R}_{ij}$ for every $(i,j) \in [N_0]\times[N_1]$ with equality when $(i,j)\in \mathcal{I}(\bar\pi)$ (i.e. when $(x^{(i)},y^{(j)})\in \mathrm{spt}(\bar\pi)$). Recall from Section 3.2.4 that, as $\bar\pi$ is optimal for $\mathsf{D}(\mu_0,\mu_1)$, it is also optimal for $\mathsf{OT}_{\mathbf{A}_{\bar\pi}}(\mu_0,\mu_1)$, where $\mathbf{A}_{\bar\pi} = \frac{1}{2}\int xy^\intercal d\bar\pi(x,y)$. It follows that $r^\intercal m_0 + s^\intercal m_1 = \int c_{\mathbf{A}_{\bar\pi}} d\bar\pi$ such that (r,s) can be identified with a pair of OT potentials for $\mathsf{OT}_{\mathbf{A}_{\bar\pi}}(\mu_0,\mu_1)$. As the objective in (3.27) is invariant to the transformation $(r,s)\mapsto(r+C,s-C)$ and any choice of OT potential for $\mathsf{OT}_{\mathbf{A}_{\bar\pi}}(\mu_0,\mu_1)$ admits a version which is bounded (see Lemma 8), the feasible set in (3.27) can be restricted to a bounded set. Conclude that (3.27) (and hence (3.25)) admit optimal solutions so that strong duality holds between the problems (3.27) and (3.24) holds (i.e. their optimal values coincide).

Given these deliberations, it follows from (3.23) that

$$\limsup_{t\downarrow 0}\frac{\mathsf{D}(\mu_{0,t},\mu_{1,t})^2-\mathsf{D}(\mu_0,\mu_1)^2}{t} \le \varsigma_{\bar\pi} + \sup_{(\varphi_0,\varphi_1)\in\mathcal{D}_{\mathbf{A}_{\bar\pi}}}\left\{\int\varphi_0 d(\nu_0-\mu_0)+\int\varphi_1 d(\nu_1-\mu_1)\right\},$$

which can be further tightened by minimizing the right hand side over all optimal couplings for $\mathsf{D}(\mu_0,\mu_1)^2$. Alternatively, if $\mathbf{A}\in\mathrm{argmin}_{B_{\mathrm{F}}(M)}\left(\Phi_{(\mu_0,\mu_1)}\right)$, then any optimal coupling, $\pi^\star$, for $\mathsf{OT}_{\mathbf{A}}(\mu_0,\mu_1)$ with $\mathbf{A}=\frac{1}{2}\int xy^\intercal d\pi^\star(x,y)$ is optimal for $\mathsf{D}(\mu_0,\mu_1)^2$ by Theorem 8 (as aforementioned, any solution π to $\mathsf{D}(\mu_0,\mu_1)^2$ also solves $\mathsf{OT}_{\mathbf{A}}(\mu_0,\mu_1)$ for $\mathbf{A}=\frac{1}{2}\int xy^\intercal d\pi(x,y)$) yielding

$$\begin{aligned}\limsup_{t\downarrow 0}\frac{\mathsf{D}(\mu_{0,t},\mu_{1,t})^2-\mathsf{D}(\mu_0,\mu_1)^2}{t} &\le \int f_0 d\rho_0 + \int f_1 d\rho_1 \\ &+\inf_{\mathbf{A}\in\mathcal{A}}\left\{\inf_{\pi\in\bar\Pi^\star_{\mathbf{A}}}\int g_{0,\pi}d\rho_0+\int g_{1,\pi}d\rho_1+\sup_{(\varphi_0,\varphi_1)\in\bar{\mathcal{D}}_{\mathbf{A}}}\int\bar\varphi_0 d\rho_0+\int\bar\varphi_1 d\rho_1\right\},\end{aligned}\tag{3.28}$$

upon expanding the expression for $\varsigma_{\bar\pi}$.

We conclude by proving the complementary inequality. Let $t_n\downarrow 0$ be arbitrary, and let π_{t_n} be any GW plan for $\mathsf{D}(\mu_{0,t_n},\mu_{1,t_n})$. In what follows, we write $\mathbf{P}_{t_n}$ for the matrix with entries $(\mathbf{P}_{t_n})_{ij}=\pi_{t_n}\left(\left\{(x^{(i)},y^{(j)})\right\}\right)$ for $(i,j)\in[N_0]\times[N_1]$ and m_0,n_0,m_1,n_1 for the vectors of weights of $\mu_0, \nu_0, \mu_1, \nu_1$ as in the first half of the proof. With this, $\mathsf{D}(\mu_{0,t}, \mu_{1,t})^2$ can be viewed as a perturbation of the (finite dimensional) quadratic program $\mathsf{D}(\mu_0, \mu_1)^2$ given in (3.11), where the equality constraints are of the form

$$\left\{\mathbf{P} \in \mathbb{R}^{N_0 \times N_1} : \langle \mathbf{\Lambda}_i, \mathbf{P} \rangle_{\mathrm{F}} = b_i(t) \text{ for } i \in [N_0 + N_1]\right\}$$

for some fixed matrices $(\mathbf{\Lambda}_i)_{i=1}^{N_0+N_1} \subset \mathbb{R}^{N_0 \times N_1}$ and a vector $b(t) = (m_0 + t(n_0 - m_0), m_1 + t(n_1 - m_1)) \in \mathbb{R}^{N_0+N_1}$. In this setting, the set of matrix solutions, S_t, of $\mathsf{D}(\mu_{0,t}, \mu_{1,t})^2$ for $t \in [0, 1]$ satisfies

$$S_t \subset S_0 + \gamma \underbrace{\|(m_{0,t}, -m_{0,t}, n_{0,t}, -n_{0,t}) - (m_0, -m_0, n_0, -n_0)\|}_{=t\|(m_0-n_0,n_0-m_0,m_1-n_1,n_1-m_1)\|} B_{\mathrm{F}}(1), \tag{3.29}$$

for all t sufficiently small, where $\gamma > 0$ is a constant and the addition is understood in the sense of Minkowski (see Theorem 3 in [124]).

By the Heine-Borel theorem, there exists a subsequence $t_{n'}$ along which $\mathbf{P}_{t_{n'}} \to \mathbf{P} \in [0, 1]^{N_0 \times N_1}$ and, evidently, $\mathbf{P} \in S_0$ by the previous arguments.[7] It follows from (3.29) that, for every n' sufficiently large, there exists some $\mathbf{P}^0_{n'} \in S_0$ with $\|\mathbf{P}^0_{n'} - \mathbf{P}_{t_{n'}}\|_{\mathrm{F}} = O(t_{n'})$.

Let $(y_{n'}, z_{n'}, (S^{n'}_{ij})_{(i,j) \in \mathcal{I}(\pi_{n'})})$ denote a solution to (3.25), where the set $\mathcal{I}(\bar{\pi})$ is replaced by $\mathcal{I}(\pi_{n'})$ (where $\pi_{n'}$ is the coupling with weights given by $\mathbf{P}^0_{n'}$), and the cost $\mathbf{C}$ is replaced by $\mathbf{C}_{n'}$ with $(\mathbf{C}_{n'})_{ij} = 2 \int \Delta\left(x^{(i)}, y^{(j)}, x, y\right) d\pi_{n'}(x, y)$ for $(i, j) \in [N_0] \times [N_1]$ (note that all of the implications regarding $(y_{n'}, z_{n'}, (S^{n'}_{ij})_{(i,j) \in \mathcal{I}(\pi_{n'})})$ derived in the first half of the proof still hold, as $\pi_{n'}$ is optimal for $\mathsf{D}(\mu_0, \mu_1)^2$). For notational convenience, let $\mathbf{S}_{n'} \in \mathbb{R}^{N_0 \times N_1}$ be such that

$$(\mathbf{S}_{n'})_{ij} = \begin{cases} S^{n'}_{ij} = (\mathbf{C}_{n'})_{ij} - (y_{n'})_i - (z_{n'})_j, & \text{if } (i, j) \in \mathcal{I}(\pi_{n'}), \\ 0, & \text{if } (i, j) \notin \mathcal{I}(\pi_{n'}), \end{cases}$$

and observe that $0 = \langle \mathbf{S}_{n'}, \mathbf{P}^0_{n'} \rangle_{\mathrm{F}} = \langle \mathbf{C}_{n'}, \mathbf{P}^0_{n'} \rangle_{\mathrm{F}} - \langle y_{n'}, m_0 \rangle - \langle z_{n'}, n_0 \rangle$ by definition, and

[7] Indeed, $d(\mathbf{P}, S_0) \leq \|\mathbf{P} - \mathbf{P}_{t_{n'}}\|_{\mathrm{F}} + d(\mathbf{P}_{t_{n'}}, S_0) \to 0$ as $t_{n'} \downarrow 0$, where $d(\cdot, S_0) = \inf_{\mathbf{Q} \in S_0} \{d(\cdot, \mathbf{Q})\}$.

$\langle \mathbf{S}_{n'}, \mathbf{B}\rangle_{\mathrm{F}} \geq 0$ for any $\mathbf{B} \in [0,\infty)^{N_0\times N_1}$ as $S^{n'}_{ij} \geq 0$. As such,

$$\begin{aligned}\langle \mathbf{S}_{n'}, \mathbf{P}_{t_{n'}}\rangle_{\mathrm{F}} &= \langle \mathbf{C}_{n'}, \mathbf{P}_{t_{n'}}\rangle_{\mathrm{F}} - \langle y_{n'}, m_0 + t_{n'}(n_0 - m_0)\rangle - \langle z_{n'}, m_1 + t_{n'}(n_1 - m_1)\rangle \\ &= \langle \mathbf{C}_{n'}, \mathbf{P}_{t'_n} - \mathbf{P}^0_{n'}\rangle_{\mathrm{F}} - t_{n'}\langle y_{n'}, n_0 - m_0\rangle - t_{n'}\langle z_{n'}, n_1 - m_1\rangle \geq 0.\end{aligned} \tag{3.30}$$

Let $a : \mathbf{P} \in \mathbb{R}^{N_0\times N_1} \mapsto \sum_{\substack{i,k\in[N_0]\\ j,l\in[N_1]}} \mathbf{P}_{ij}\Delta(x^{(i)}, y^{(j)}, x^{(k)}, y^{(l)})\mathbf{P}_{kl}$ denote the quadratic form whose minimization underlies the discrete Gromov-Wasserstein distance. As a is smooth with (Fréchet) derivative $Da_{[\mathbf{P}]}(\mathrm{Q}) = 2\sum_{\substack{i,k\in[N_0]\\ j,l\in[N_1]}} \mathbf{P}_{ij}\Delta(x^{(i)}, y^{(j)}, x^{(k)}, y^{(l)})\mathbf{Q}_{kl}$, we have

$$a(\mathbf{P}_{t_{n'}}, \mathbf{P}_{t_{n'}}) - a(\mathbf{P}^0_{n'}, \mathbf{P}^0_{n'}) - \langle \mathbf{C}_{n'}, \mathbf{P}_{t_{n'}} - \mathbf{P}^0_{n'}\rangle_{\mathrm{F}} = O\left(\|\mathbf{P}_{t'_n} - \mathbf{P}^0_{n'}\|^2\right) = O\left(t^2_{n'}\right),$$

so that, from (3.30),

$$\begin{aligned}\mathsf{D}(\mu_{0,t_{n'}}, \mu_{1,t_{n'}})^2 - \mathsf{D}(\mu_0,\mu_1)^2 &\geq a(\mathbf{P}_{t_{n'}}, \mathbf{P}_{t_{n'}}) - a(\mathbf{P}^0_{n'}, \mathbf{P}^0_{n'}) - \langle \mathbf{S}_{n'}, \mathbf{P}_{t_{n'}}\rangle_{\mathrm{F}} \\ &= a(\mathbf{P}_{t_{n'}}, \mathbf{P}_{t_{n'}}) - a(\mathbf{P}^0_{n'}, \mathbf{P}^0_{n'}) - \langle \mathbf{C}_{n'}, \mathbf{P}_{t_{n'}} - \mathbf{P}^0_{n'}\rangle_{\mathrm{F}} \\ &\quad + t_{n'}\langle y_{n'}, n_0 - m_0\rangle + t_{n'}\langle z_{n'}, n_1 - m_1\rangle \\ &= t_{n'}\langle y_{n'}, n_0 - m_0\rangle + t_{n'}\langle z_{n'}, n_1 - m_1\rangle + o(t_{n'}).\end{aligned} \tag{3.31}$$

Recall from (3.26) and (3.27) and the surrounding discussion that $(y_{n'}, z_{n'})$ satisfy

$$\begin{aligned}&\langle y_{n'}, n_0 - m_0\rangle + \langle z_{n'}, n_1 - m_1\rangle \\ &\qquad = \sup_{(\varphi_0,\varphi_1)\in\bar{\mathcal{D}}_{\mathbf{A}(\pi_{n'})}} \left\{\int \varphi_0 d(\nu_0 - \mu_0) + \int \varphi_1 d(\nu_1 - \mu_1)\right\} + \varsigma_{\pi_{n'}} \\ &\qquad \geq \inf_{\pi\in\bar{\Pi}^\star}\left\{\sup_{(\varphi_0,\varphi_1)\in\bar{\mathcal{D}}_{\mathbf{A}(\pi)}} \left\{\int \varphi_0 d(\nu_0 - \mu_0) + \int \varphi_1 d(\nu_1 - \mu_1)\right\} + \varsigma_{\pi}\right\},\end{aligned} \tag{3.32}$$

where $\mathbf{A}(\pi) = \frac{1}{2}\int xy^\intercal d\pi$, and $\bar{\Pi}^\star$ denotes the set of optimal couplings for $\mathsf{D}(\mu_0,\mu_1)^2$. It follows from (3.31) and (3.32) that

$$\begin{aligned}&\liminf_{t_{n'}\downarrow 0} \frac{\mathsf{D}(\mu_{0,t_{n'}}, \mu_{1,t_{n'}})^2 - \mathsf{D}(\mu_0,\mu_1)^2}{t_{n'}} \\ &\qquad \geq \inf_{\pi\in\bar{\Pi}^\star}\left\{\sup_{(\varphi_0,\varphi_1)\in\bar{\mathcal{D}}_{\mathbf{A}(\pi)}} \left\{\int \varphi_0 d(\nu_0 - \mu_0) + \int \varphi_1 d(\nu_1 - \mu_1)\right\} + \varsigma_{\pi}\right\}.\end{aligned}$$

As this final lower bound is independent of the choice of subsequence, and the initial choice of sequence $t_n \downarrow 0$ was arbitrary,

$$\liminf_{t \downarrow 0} \frac{\mathsf{D}(\mu_{0,t}, \mu_{1,t})^2 - \mathsf{D}(\mu_0, \mu_1)^2}{t} \\ \geq \inf_{\pi \in \bar{\Pi}^\star} \left\{ \sup_{(\varphi_0, \varphi_1) \in \bar{\mathcal{D}}_{\mathbf{A}(\pi)}} \left\{ \int \varphi_0 d(\nu_0 - \mu_0) + \int \varphi_1 d(\nu_1 - \mu_1) \right\} + \varsigma_\pi \right\},$$

the right hand side can be rewritten in the same form as the right hand side of (3.28) by expanding ς_π and applying the same reasoning as the first part of the proof, proving the claim. □

Lemma 10. For any pairs $(\nu_0, \nu_1), (\rho_0, \rho_1) \in \mathcal{P}_{\mu_0} \times \mathcal{P}_{\mu_1}$, $\left| \mathsf{D}(\nu_0, \nu_1)^2 - \mathsf{D}(\nu_0, \nu_1)^2 \right| \leq C \| \nu_0 \otimes \nu_1 - \rho_0 \otimes \rho_1 \|_{\infty, \mathcal{F}^\oplus}$ for some constant $C > 0$ which is independent of $(\nu_0, \nu_1), (\rho_0, \rho_1)$.

Proof. Analogously to the proof of Lemma 9, $\mathsf{D}(\nu_0, \nu_1)^2$ and $\mathsf{D}(\nu_0, \nu_1)^2$ represent the optimal values of quadratic programs on $\mathbb{R}^{N_0 \times N_1}$. Let n_0, p_0, n_1, p_1 denote the weight vectors for $\nu_0, \rho_0, \nu_1, \rho_1$ respectively. By Theorem 3 in [124],

$$\begin{aligned} |\mathsf{D}(\nu_0, \nu_1)^2 - \mathsf{D}(\nu_0, \nu_1)^2| &\leq C' \| (n_0, -n_0, n_1, -n_1) - (p_0, -p_0, p_1, -p_1) \| \\ &\leq \sqrt{2} C' \left(\| n_0 - p_0 \| + \| n_1 - p_1 \| \right), \end{aligned} \tag{3.33}$$

for some constant C' which is independent of n_0, n_1, p_0, p_1, as all vectors involved lie in a bounded set and result in nonempty feasible sets for the relevant quadratic programs. As $\mathcal{F}_0, \mathcal{F}_1$ are symmetric in the sense that if $f \in \mathcal{F}_0$, then $-f \in \mathcal{F}_0$, and similarly for $\mathcal{F}_1$,

$$\| \nu_0 \otimes \nu_1 - \rho_0 \otimes \rho_1 \|_{\infty, \mathcal{F}^\oplus} = \sup_{f \in \mathcal{F}_0} |(\nu_0 - \rho_0)(f)| + \sup_{f \in \mathcal{F}_1} |(\nu_1 - \rho_1)(f)| = \| n_0 - p_0 \|_1 + \| n_1 - p_1 \|_1.$$

As all norms on a finite dimensional space are equivalent, this display and (3.33) prove the claim. □

Proof of Theorem 9. Theorem 9 follows directly from Lemmas 9 and 10. □

Proof of Theorem 10

The generic limit theorems follow from Theorem 9 by applying the delta method, see Proposition 16.

To see that the empirical measure satisfies the desired condition, note that, in this setting, $\hat{\mu}_{0,n}$ and μ_0 can be identified with vectors of weights $\hat{m}_{0,n}$ and $m_0 \in \mathbb{R}^{N_0}$. By the central limit theorem,

$$\sqrt{n}\,(\hat{m}_{0,n} - m_0) \xrightarrow{d} N(0, \mathbf{\Sigma}_{m_0}),$$

where $\mathbf{\Sigma}_{m_0}$ is the multinomial covariance matrix

$$(\mathbf{\Sigma}_{\mu_0})_{ij} = \begin{cases} (m_0)_i(1 - (m_0)_i), & \text{if } i = j, \\ -(m_0)_i(m_0)_j, & \text{otherwise.} \end{cases}$$

Thus, $\sqrt{n}(\hat{\mu}_{0,n} - \mu_0) \xrightarrow{d} G_{\mu_0}$ in $\ell^\infty(\mathcal{F}_0)$ where G_{μ_0} is a μ_0-Brownian bridge. An analogous result evidently holds for $\sqrt{n}(\hat{\mu}_{1,n} - \mu_1)$.

As the $X_{0,i}$'s and $X_{1,i}$'s are independent, Example 1.4.6 in [201] and Lemma 3.2.4 [73] imply that

$$\sqrt{n}\,(\hat{\mu}_{0,n} - \mu_0, \hat{\mu}_{1,n} - \mu_1) \xrightarrow{d} (G_{\mu_0}, G_{\mu_1}), \text{ in } \ell^\infty(\mathcal{F}_0) \times \ell^\infty(\mathcal{F}_1).$$

As the map $(l_0, l_1) \in \ell^\infty(\mathcal{F}_0) \times \ell^\infty(\mathcal{F}_1) \mapsto l_0 \otimes l_1 \in \ell^\infty(\mathcal{F}^\oplus)$ is continuous it follows from the continuous mapping theorem that

$$\sqrt{n}\,(\hat{\mu}_{0,n} \otimes \hat{\mu}_{1,n} - \mu_0 \otimes \mu_1) \xrightarrow{d} G_{\mu_0 \otimes \mu_1}, \text{ in } \ell^\infty(\mathcal{F}^\oplus),$$

where $G_{\mu_0 \otimes \mu_1}(f_0 \oplus f_1) = G_{\mu_0}(f_0) + G_{\mu_1}(f_1)$ for every $f_0 \oplus f_1 \in \mathcal{F}^\oplus$. $\square$

Proof of Corollary 5

Throughout we assume without loss of generality that $d_0 \leq d_1$ and let $\iota : x \in \mathbb{R}^{d_0} \to (x, 0, \dots, 0) \in \mathbb{R}^{d_1}$ denote an isometric embedding of $\mathbb{R}^{d_0}$ into $\mathbb{R}^{d_1}$.

Lemma 11. If $T_\sharp \bar{\mu}_0 = \bar{\mu}_1$ for some isometry $T : \mathrm{spt}(\bar{\mu}_0) \to \mathrm{spt}(\bar{\mu}_1)$ then T is inner product preserving.

Proof. Let $S : \iota(\mathrm{spt}(\bar{\mu}_0)) \to \mathrm{spt}(\bar{\mu}_1)$ be an isometry defined via $S = T \circ \iota^{-1}$ and extend S to an isometric self-map on $\mathbb{R}^{d_1}$ which we denote by the same symbol. By the Mazur-Ulam theorem, $S(y) = Uy + b$ for some orthogonal matrix $U \in \mathbb{R}^{d_1} \times \mathbb{R}^{d_1}$ and $b \in \mathbb{R}^{d_1}$. Observe that

$$0 = \mathbb{E}_{\bar{\mu}_1}[X] = \mathbb{E}_{\bar{\mu}_0}[T(X)] = \sum_{i=1}^{N} \bar{\mu}_0(\{x^{(i)}\}) \left(U\iota(x^{(i)}) + b \right) = b + U \sum_{i=1}^{N} \bar{\mu}_0(\{x^{(i)}\}) \iota(x^{(i)}).$$

As $\mathbb{E}_{\bar{\mu}_0}[X] = 0$, $\sum_{i=1}^{N} \bar{\mu}_0(\{x^{(i)}\})\iota(x^{(i)}) = \iota(\mathbb{E}_{\bar{\mu}_0}[X]) = 0$ so that $b = 0$. Conclude that, for any $x, x' \in \mathrm{spt}(\bar{\mu}_0)$, $T(x)^\intercal T(x') = \iota(x)^\intercal U^\intercal U \iota(x') = x^\intercal x'$. $\square$

Lemma 12. Suppose that $T_\sharp \bar{\mu}_0 = \bar{\mu}_1$ for some isometry $T : \mathrm{spt}(\bar{\mu}_0) \to \mathrm{spt}(\bar{\mu}_1)$. Then, for any $\mathbf{A} \in \mathcal{A}$ and $\pi \in \bar{\Pi}^\star_{\mathbf{A}}$, π is induced by an isometric alignment map T_π, $g_{0,\pi}(x) = 8x^\intercal \mathbb{E}_{\bar{\mu}_0}[\|X\|^2 X]$, $g_{1,\pi}(y) = 8y^\intercal \mathbb{E}_{\bar{\mu}_1}[\|Y\|^2 Y]$, and every pair of OT potentials for $\mathsf{OT}_{\mathbf{A}}(\bar{\mu}_0, \bar{\mu}_1)$ is of the form

$$\varphi_0(x) = -2\|x\|^4 - 8x^\intercal \mathbf{\Sigma}_{\bar{\mu}_0} x + h(x), \quad \varphi_1(y) = -2\|y\|^4 - 8y^\intercal \mathbf{\Sigma}_{\bar{\mu}_1} y - h(T_\pi^{-1}(y))$$

for some $h \in \mathcal{H}$.

Proof. Given that $T_\sharp \bar{\mu}_0 = \bar{\mu}_1$, we have that $\mathsf{D}(\bar{\mu}_0, \bar{\mu}_1) = 0$ whereby π is a GW plan if and only if it is induced by a Gromov-Monge map $T_\pi : \mathrm{spt}(\bar{\mu}_0) \to \mathrm{spt}(\bar{\mu}_1)$ which may differ from T. In any case, $g_{0,\pi}(x) = 8x^\intercal \mathbb{E}_{(X,Y)\sim\pi}\left[\|Y\|^2 X\right] = 8x^\intercal \mathbb{E}_{\bar{\mu}_0}\left[\|T_\pi(X)\|^2 X\right] =$

$8x^\intercal \mathbb{E}_{\bar{\mu}_0}\left[\|X\|^2 X\right]$ where we have used the fact that T_π is inner product preserving and hence norm preserving (see Lemma 11). The same arguments apply to $g_{1,\pi}$.

From Theorem 8, $\mathbf{A} \in \mathcal{A}$ if and only if $\mathbf{A} = \frac{1}{2}\int xy^\intercal d\pi(x,y) = \frac{1}{2}\int zT_\pi(z)^\intercal d\bar{\mu}_0(z)$ for some GW plan π. The corresponding OT potentials $(\varphi_{0,\pi}, \varphi_{1,\pi})$ for $\mathsf{OT}_{\mathbf{A}}(\bar{\mu}_0, \bar{\mu}_1)$ satisfy

$$\varphi_{0,\pi}(x) + \varphi_{1,\pi}(y) \leq -4\|x\|^2\|y\|^2 - 32x^\intercal \mathbf{A} y$$

for every $(x,y) \in \operatorname{spt}(\bar{\mu}_0) \times \operatorname{spt}(\bar{\mu}_1)$ with equality π-a.e. Observe that

$$-4\|x\|^2\|y\|^2 = 2\left(\|x\|^2 - \|y\|^2\right)^2 - 2\|x\|^4 - 2\|y\|^4,$$

whereas

$$(x - T_\pi^{-1}(y))^\intercal \mathbf{A}(T_\pi(x) - y) = -x^\intercal \mathbf{A} y + \frac{1}{2}x^\intercal \mathbf{\Sigma}_{\bar{\mu}_0} x + \frac{1}{2}y^\intercal \mathbf{\Sigma}_{\bar{\mu}_1} y - x^\intercal \mathbf{A} y,$$

noting that $x^\intercal \mathbf{A} T_\pi(x) = \frac{1}{2}\int x^\intercal z T_\pi(z)^\intercal T_\pi(x) d\bar{\mu}_0(z) = \frac{1}{2}\int x^\intercal z z^\intercal x d\bar{\mu}_0(z) = \frac{1}{2}x^\intercal \mathbf{\Sigma}_{\bar{\mu}_0} x$ and that

$$T_\pi^{-1}(y)^\intercal \mathbf{A} T_\pi(x) = \frac{1}{2}\int y^\intercal T_\pi(z) z^\intercal x d\bar{\mu}_0(z) = \frac{1}{2}x^\intercal \int z T_\pi(z)^\intercal d\bar{\mu}_0(z) y = x^\intercal \mathbf{A} y.$$

In sum,

$$\begin{aligned}\varphi_{0,\pi}(x) + 2\|x\|^4 + 8x^\intercal \mathbf{\Sigma}_{\bar{\mu}_0} x + \varphi_{1,\pi}(y) + 2\|y\|^4 + 8y^\intercal \mathbf{\Sigma}_{\bar{\mu}_1} y \\ \leq 2\left(\|x\|^2 - \|y\|^2\right)^2 + 16(x - T_\pi^{-1}(y))^\intercal \mathbf{A}(T_\pi(x) - y),\end{aligned}$$

and, setting $\varphi_{0,\pi}(x) = -2\|x\|^4 - 8x^\intercal \mathbf{\Sigma}_{\bar{\mu}_0} x + h(x)$ and $\varphi_{1,\pi}(y) = -2\|y\|^4 - 8y^\intercal \mathbf{\Sigma}_{\bar{\mu}_1} y + g(y)$ for some functions $h : \operatorname{spt}(\bar{\mu}_0) \to \mathbb{R}$ and $g : \operatorname{spt}(\bar{\mu}_1) \to \mathbb{R}$, since $\pi = (\mathrm{Id}, T_\pi)_\sharp \bar{\mu}_0$ we must have that $h(x) + g(T_\pi(x)) = 0$ for every $x \in \operatorname{spt}(\bar{\mu}_0)$ hence $g(y) = -h(T_\pi^{-1}(y))$. In order for $(\varphi_{0,\pi}, \varphi_{1,\pi})$ to be OT potentials we further enforce that

$$h(x) - h(T_\pi^{-1}(y)) \leq 2\left(\|x\|^2 - \|y\|^2\right)^2 + 16(x - T_\pi^{-1}(y))^\intercal \mathbf{A}(T_\pi(x) - y),$$

for every $(x,y) \in \operatorname{spt}(\bar{\mu}_0) \times \operatorname{spt}(\bar{\mu}_1)$. As $T_\pi^{-1}(y) = \operatorname{spt}(\bar{\mu}_0)$ conclude that

$$h(x) - h(x') \leq 2\left(\|x\|^2 - \|x'\|^2\right)^2 + 16\underbrace{(x - x')^\intercal \mathbf{A}(T_\pi(x) - T_\pi(x'))}_{=\frac{1}{2}(x-x')^\intercal \mathbf{\Sigma}_{\bar{\mu}_0}(x-x')},$$

for every $x, x' \in \operatorname{spt}(\bar{\mu}_0)$ whereby $h \in \mathcal{H}$. □

Proof of Corollary 5. Applying Theorem 10 and the result of Lemma 12,

$$\sqrt{n}\mathsf{D}(\hat{\mu}_{0,n},\hat{\mu}_{1,n})^2 \xrightarrow{d} \inf_{\mathbf{A}\in\mathcal{A}}\inf_{\pi\in\bar{\Pi}^\star_{\mathbf{A}}}\sup_{(\varphi_0,\varphi_1)\in\bar{\mathcal{D}}_{\mathbf{A}}}\left\{\chi_{\mu_0\otimes\mu_1}((f_0+g_{0,\pi}+\bar{\varphi}_0)\oplus(f_1+g_{1,\pi}+\bar{\varphi}_1)\right\},$$

where $g_{0,\pi}(x) = 8x^\intercal\mathbb{E}_{\bar{\mu}_0}[\|X\|^2X], g_{1,\pi}(y) = 8y^\intercal\mathbb{E}_{\bar{\mu}_1}[\|Y\|^2Y]$ which are independent of π and $\mathbf{A}$. Furthermore, for any $\mathbf{A}\in\mathcal{A}$ and $\pi\in\bar{\Pi}_{\mathbf{A}}$, π is induced by an (isometric) alignment map $T_\pi : \mathrm{spt}(\bar{\mu}_0)\to\mathrm{spt}(\bar{\mu}_1)$ and every pair of OT potentials for $\mathsf{OT}_{\mathbf{A}}(\bar{\mu}_0,\bar{\mu}_1)$ is of the form

$$\varphi_0(x) = -2\|x\|^4 - 8x^\intercal\mathbf{\Sigma}_{\bar{\mu}_0}x + h(x), \quad \varphi_1(y) = -2\|y\|^4 - 8y^\intercal\mathbf{\Sigma}_{\bar{\mu}_1}y - h(T_\pi^{-1}(y))$$

for some $h\in\mathcal{H}$.

We now set $\psi_0(x) = -2\|x\|^4 - 8x^\intercal\mathbf{\Sigma}_{\bar{\mu}_0}x, \psi_1(y) = -2\|y\|^4 - 8y^\intercal\mathbf{\Sigma}_{\bar{\mu}_1}x$ and show that $f_0+g_0+\bar{\psi}_0$ is constant so that $f_1+g_1+\bar{\psi}_1$ is constant by symmetry. Then, $\sqrt{n}(\mu_{0,n}\otimes\mu_{1,n}-\mu_0\otimes\mu_1)(f_0+g_0+\bar{\varphi}_0\oplus f_1+g_1+\bar{\varphi}_1) = \sqrt{n}(\mu_{0,n}\otimes\mu_{1,n}-\mu_0\otimes\mu_1)(h(\cdot-\mathbb{E}_{\mu_0}[X])\oplus -h(T^{-1}(\cdot-\mathbb{E}_{\mu_1}[X])))$ so that

$$\chi_{\mu_0\otimes\mu_1}(f_0+g_{0,\pi}+\bar{\varphi}_0\oplus f_1+g_{1,\pi}+\bar{\varphi}_1) = \chi_{\mu_0\otimes\mu_1}(h(\cdot-\mathbb{E}_{\mu_0}[X])\oplus -h(T^{-1}(\cdot-\mathbb{E}_{\mu_1}[X]))).$$

To this end, we first expand $f_0(x) = 2\int\|x-x'\|^4d\mu_0(x') - 4\int\|x'-\mathbb{E}_{\mu_0}[X]\|^2\|x-\mathbb{E}_{\mu_0}[X]\|^2d\mu_0(x')$ (recall that $M_2(\bar{\mu}_0) = M_2(\bar{\mu}_1)$ in this case) using the formula

$$\begin{aligned}
&2\int\|x-x'\|^4d\mu_0(x')\\
&= 2\int\left(\|x-\mathbb{E}_{\mu_0}[X]\|^2 + 2(x-\mathbb{E}_{\mu_0}[X])^\intercal(\mathbb{E}_{\mu_0}[X]-x') + \|\mathbb{E}_{\mu_0}[X]-x'\|^2\right)^2d\mu_0(x'),\\
&= 2\int\|x-\mathbb{E}_{\mu_0}[X]\|^4 + 4\left|(x-\mathbb{E}_{\mu_0}[X])^\intercal(\mathbb{E}_{\mu_0}[X]-x')\right|^2 + \|\mathbb{E}_{\mu_0}[X]-x'\|^4d\mu_0(x')\\
&+ 4\int\|x-\mathbb{E}_{\mu_0}[X]\|^2\|\mathbb{E}_{\mu_0}[X]-x'\|^2 + 2(x-\mathbb{E}_{\mu_0}[X])^\intercal(\mathbb{E}_{\mu_0}[X]-x')\|\mathbb{E}_{\mu_0}[X]-x'\|^2\\
&+ 2(x-\mathbb{E}_{\mu_0}[X])^\intercal(\mathbb{E}_{\mu_0}[X]-x')\|x-\mathbb{E}_{\mu_0}[X]\|^2d\mu_0(x'),
\end{aligned}$$

and note that $\bar{\psi}_0(x) = -2\|x-\mathbb{E}_{\mu_0}[X]\|^4 - 8\int\left|(x-\mathbb{E}_{\mu_0}[X])^\intercal(\mathbb{E}_{\mu_0}[X]-x')\right|^2d\mu_0(x')$, and $g_0(x) = 8(x-\mathbb{E}_{\mu_0}[X])^\intercal\mathbb{E}_{\mu_0}[\|X-\mathbb{E}_{\mu_0}[X]\|^2(X-\mathbb{E}_{\mu_0}[X])] + 8\mathbb{E}_{\mu_0}[X]^\intercal\mathbb{E}_{\mu_0}[\|X-\mathbb{E}_{\mu_0}[X]\|^2(X-$

$\mathbb{E}_{\mu_0}[X])]$, whence

$$f_0 + g_0 + \bar{\psi}_0 = \int \|\mathbb{E}_{\mu_0}[X] - x'\|^4 d\mu_0(x') + 8\mathbb{E}_{\mu_0}[X]^\intercal \mathbb{E}_{\mu_0}[\|X - \mathbb{E}_{\mu_0}[X]\|^2 (X - \mathbb{E}_{\mu_0}[X])]$$
$$+ 8 \underbrace{\int \|x - \mathbb{E}_{\mu_0}[X]\|^2 (x - \mathbb{E}_{\mu_0}[X])^\intercal (\mathbb{E}_{\mu_0}[X] - x') d\mu_0(x')}_{=0}, \tag{3.34}$$

proving the claim.

Conclude that

$$\sqrt{n}\mathsf{D}(\mu_{0,n}, \mu_{1,n})^2 \xrightarrow{d} \inf_{\mathbf{A}\in\mathcal{A}} \inf_{\pi\in\bar{\Pi}^\star_{\mathbf{A}}} \sup_{h\in\mathcal{H}} \left\{\chi_{\mu_0\otimes\mu_1}(h(\cdot - \mathbb{E}_{\mu_0}[X]) \oplus -h(T_\pi^{-1}(\cdot - \mathbb{E}_{\mu_1}[X])))\right\}$$
$$= \inf_{T\in\bar{\mathcal{T}}} \sup_{h\in\mathcal{H}} \left\{\chi_{\mu_0\otimes\mu_1}(h(\cdot - \mathbb{E}_{\mu_0}[X]) \oplus -h(T^{-1}(\cdot - \mathbb{E}_{\mu_1}[X])))\right\},$$

proving the first claim.

Specifying this result for the empirical measures, we have from Theorem 10 that $\chi_{\mu_0\otimes\mu_1}(f_0 \oplus f_1) = G_{\mu_0}(f_0) + G_{\mu_1}(f_1)$ for every $f_0 \oplus f_1 \in \mathcal{F}^\oplus$. It remains to show that $G_{\mu_1}(- h(T^{-1}(\cdot - \mathbb{E}_{\mu_1}[X]))) = G_{\mu_0}(- h(\cdot - \mathbb{E}_{\mu_0}[X]))$ for any $T \in \bar{\mathcal{T}}$. To this end, note that if $f_1(\cdot - \mathbb{E}_{\mu_1}[X]), f_1'(\cdot - \mathbb{E}_{\mu_1}[X]) \in \mathcal{F}_1$,

$$\mathrm{Cov}(G_{\mu_1}(f_1(\cdot - \mathbb{E}_{\mu_1}[X])), G_{\mu_1}(f_1'(\cdot - \mathbb{E}_{\mu_1}[X])))$$
$$= \mathrm{Cov}_{\bar{\mu}_0}(f_1 \circ T, f_1' \circ T)$$
$$= \mathrm{Cov}(G_{\mu_0}(f_1(T(\cdot - \mathbb{E}_{\mu_0}[X]))), G_{\mu_0}(f_1'(T(\cdot - \mathbb{E}_{\mu_0}[X])))),$$

thus $\sqrt{n}\mathsf{D}(\hat{\mu}_{0,n}, \hat{\mu}_{1,n})^2 \xrightarrow{d} \sup_{h\in\mathcal{H}} \left\{G_{\mu_0}(h(\cdot - \mathbb{E}_{\mu_0}[X])) + G'_{\mu_0}(-h(\cdot - \mathbb{E}_{\mu_0}[X]))\right\}$, where G'_{μ_0} is an independent copy of G_{μ_0}. As $\mathcal{H}$ is symmetric in the sense that if $h \in \mathcal{H}$, then $-h \in \mathcal{H}$, this limit can be expressed as $\sqrt{2}\|G_{\mu_0}\|_{\infty,\bar{\mathcal{H}}}$. □

Proof of Theorem 11

We will prove that the direct estimation approach proposed in Section 3.3.3 for estimating the limit L_{μ_0} in Theorem 10 is consistent under the following, general, conditions. This enables us to treat both, by the same token, the empirical measures and the measures defined in Section 3.6 for testing graph isomorphisms.

Assumption 2. The first set of assumptions regard the structure of $\chi_{\mu_0 \otimes \mu_1}$ from Assumption 1.

- $\chi_{\mu_0 \otimes \mu_1}(f_0 \oplus f_1) = \chi_{\mu_0}(f_0) + \chi_{\mu_1}(f_1)$ for $f_0 \oplus f_1 \in \mathcal{F}^{\oplus}$, where χ_{μ_0} and χ_{μ_1} are tight mean-zero Gaussian processes on $\mathcal{F}_0$ and $\mathcal{F}_1$,
- For any Gromov-Monge map T between $\bar{\mu}_0$ and $\bar{\mu}_1$, it holds that $\chi_{\mu_1}(g_1(\cdot - \mathbb{E}_{\mu_0}[X])) = \chi_{\mu_0}(g_1(T(\cdot - \mathbb{E}_{\mu_1}[X])))$ whenever $g_1(\cdot - \mathbb{E}_{\mu_0}[X]) \oplus g_1(T(\cdot - \mathbb{E}_{\mu_1}[X])) \in \mathcal{F}^{\oplus}$,
- There exists a mean-zero Gaussian vector, $Z \in \mathbb{R}^{N_0}$, with covariance $\mathbf{\Sigma}_Z$ for which $\chi_{\mu_0}(h) = Z^{\intercal} u_h$ for every $h \in \bar{\mathcal{H}}$, where $u_h = (h(\bar{x}^{(1)}), \dots, h(\bar{x}^{(N_0)})) \in \mathbb{R}^{N_0}$ for $(\bar{x}^{(i)})_{i=1}^{N_0} = \mathrm{spt}(\bar{\mu}_0)$.

The remaining assumptions pertain to the construction of the direct estimator.

- $\mu_{0,n}$ and $\mathbf{\Sigma}_n$ are estimators of μ_0 and $\mathbf{\Sigma}_Z$ constructed from samples $Y_1, \dots, Y_n$ from an auxiliary distribution ν on some (perhaps distinct) space,
- Conditionally on $Y_1, Y_2, \dots$, $\mathbb{E}_{\mu_{0,n}}[X] \to \mathbb{E}_{\mu_0}[X]$, $\mathbf{\Sigma}_{\bar{\mu}_{0,n}} \to \mathbf{\Sigma}_{\bar{\mu}_0}$, and $\mathbf{\Sigma}_n \to \mathbf{\Sigma}_Z$ deterministically given almost every realization of $Y_1, Y_2, \dots$

Under Assumption 2, $L_{\mu_0} = \sqrt{2}\|\chi_{\mu_0}\|_{\infty, \bar{\mathcal{H}}} = \sqrt{2}\sup_{h \in \bar{\mathcal{H}}} Z^{\intercal} u_h$ as follows from the proof of Corollary 5. This suggests setting $(\bar{x}^{(i)})_{i=1}^{N_n} = \mathrm{spt}(\bar{\mu}_{0,n})$ and estimating L_{μ_0} as

$$\begin{aligned} L_n = \sqrt{2} \sup_{u \in \mathcal{H}_n} Z_n^{\intercal} u \text{ for } Z_n \sim N(0, \Sigma_n) \text{ and} \\ \mathcal{H}_n = \{u \in \mathbb{R}^{N_n} \mid u_i - u_j \leq 2(\|\bar{x}^{(i)}\|^2 - \|\bar{x}^{(j)}\|^2)^2 \\ + 8(\bar{x}^{(i)} - \bar{x}^{(j)})^{\intercal} \mathbf{\Sigma}_{\bar{\mu}_{0,n}} (\bar{x}^{(i)} - \bar{x}^{(j)}), i \neq j \in [N_n]\}, \end{aligned} \tag{3.35}$$

Algorithm 4 Generate samples from L_n

Given samples $Y_1, \dots, Y_n$ from ν,

1: Let $(\bar{x}^{(i)})_{i=1}^{N_n} = \mathrm{spt}(\bar{\mu}_{0,n})$.

2: Let $b_n \in \mathbb{R}^{N_n(N_n-1)}$ and $\mathbf{A} \in \mathbb{R}^{N_n(N_n-1)\times N_n}$ be such that $\mathcal{H}_n = \{u \in \mathbb{R}^{N_n} : \mathbf{A}u \le b_n\}$.

3: Sample $Z_n \sim N(0, \mathbf{\Sigma}_n)$ and solve $\sqrt{2}\sup_{u\in\mathcal{H}_n} Z_n^\intercal u \sim L_n$, padding u by zeros if necessary. Repeat step 3 to generate more samples.

Algorithm 4 summarizes the procedure for sampling from L_n.

We now establish the following generalization of Theorem 11

Theorem 18. In the setting of Corollary 5 and Assumption 2, if $\mathbb{P}(Z_n^\intercal \mathbb{1}_{N_0} = 0) = \mathbb{P}(Z^\intercal \mathbb{1}_{N_0} = 0) = 1$ for all n sufficiently large and Z is not a point mass, then

$$\lim_{n\to\infty} \sup_{t\ge 0} \left|\mathbb{P}(L_n \le t | Y_1, Y_2, \dots, Y_n) - \mathbb{P}(L_{\mu_0} \le t)\right| = 0,$$

for almost every realization of $Y_1, Y_2, \dots$

The proof of Theorem 18 depends on the following technical lemmas.

Lemma 13. $\mathcal{H}$, treated as a subset of $\mathbb{R}^{N_0}$, has empty interior if and only if μ_0 is a point mass.

Proof. For any $x, x' \in \mathrm{spt}(\bar{\mu}_0)$,

$$\begin{aligned}(x - x')^\intercal \mathbf{\Sigma}_{\bar{\mu}_0}(x - x') &= \sum_{z\in\mathrm{spt}(\bar{\mu}_0)} \bar{\mu}_0(\{z\})(x - x')^\intercal z z^\intercal (x - x') \\ &= \sum_{z\in\mathrm{spt}(\bar{\mu}_0)} \bar{\mu}_0(\{z\})(z^\intercal x - z^\intercal x')^2,\end{aligned}$$

so that $(x - x')^\intercal \mathbf{\Sigma}_{\bar{\mu}_0}(x - x') = 0$ if and only if $x = x'$. To see this, observe that, for distinct points $x, x' \in \mathrm{spt}(\bar{\mu}_0)$, one of $\|x\|^2 - x^\intercal x'$ or $\|x'\|^2 - x^\intercal x'$ is nonzero as otherwise $\|x\|^2 = \|x'\|^2 = x^\intercal x'$ which holds if and only if $x = x'$. Let $C = 8\min_{\substack{x,x'\in\mathrm{spt}(\bar{\mu}_0)\\ x\neq x'}}(x - x')^\intercal \mathbf{\Sigma}_{\bar{\mu}_0}(x - x')$, then any function, h, on $\mathrm{spt}(\bar{\mu}_0)$ taking values in $[a, a + C]$ for any $a \in \mathbb{R}$ is such that $h \in \mathcal{H}$. □

We proceed with the proof of Theorem 18 by setting some notation. For any $c \in \mathbb{R}^{N_0}$ and any strictly positive vector $b = (b_{ij})_{\substack{i,j \in [N_0] \\ i \neq j}} \in \mathbb{R}_{>0}^{N_0(N_0-1)}$, let $\mathrm{LP}(c, b)$ denote the linear program

$$\begin{aligned} \sup_{u \in \mathbb{R}^{N_0}} \quad & \langle c, u \rangle \\ \text{subject to} \quad & u_i - u_j \leq b_{ij}, \forall i, j \in [N_0], i \neq j, \end{aligned}$$

$v(c, b)$ denote the corresponding optimal value, and $D = \mathbb{R}^{N_0} \times \mathbb{R}_{>0}^{N_0(N_0-1)}$ denote the set of all such pairs (c, b). Remark that if u satisfies the constraints of $\mathrm{LP}(c, b)$ then so too does $u + a$ for any $a \in \mathbb{R}^{N_0}$. As such, $\mathrm{LP}(c, b)$ is unbounded and, unless $c^\intercal \mathbb{1}_{N_0} = 0$, $v(c, b) = \infty$. On the other hand, if $c^\intercal \mathbb{1}_{N_0} = 0$ we have that, for any feasible u, $\tilde{u} = u - u_1 + a$ achieves the same objective value as u for any fixed $a \in \mathbb{R}$. Observe that $\tilde{u}_1 = a$ and $\tilde{u}_i \in [a - b_{1i}, b_{i1} + a]$ for every $i \neq 1$. We thus consider the auxiliary problem $\mathrm{LP}_\delta(c, b)$ given by

$$\begin{aligned} \sup_{u \in \mathbb{R}^{N_0}} \quad & \langle c, u \rangle \\ \text{subject to} \quad & u_i - u_j \leq b_{ij}, \forall i, j \in [N_0], i \neq j, \\ & u_1 \leq \delta, -u_1 \leq \delta, \end{aligned}$$

for any $\delta > 0$. For any strictly positive b, this problem is feasible and bounded hence solvable and, if $c^\intercal \mathbb{1}_{N_0} = 0$, its optimal value function $v_\delta(c, b)$ coincides with $v(c, b)$.

It is useful to note that this orthogonality condition is satisfied if $c \sim N(0, \mathbf{\Sigma}_p)$, where $\mathbf{\Sigma}_p$ is the covariance matrix in (3.12) and p is a vector of probabilities.

Lemma 14. Fix a nonnegative vector $p \in \mathbb{R}^{N_0}$ whose entries sum to 1. If $Z \sim N(0, \mathbf{\Sigma}_p)$, then $Z^\intercal \mathbb{1}_{N_0} = 0$ almost surely.

Proof. The k-th entry of the vector $\mathbf{\Sigma}_p \mathbb{1}_{N_0}$ is given by $p_k(1 - p_k) - \sum_{j \neq k} p_k p_j = p_k(1 - 1) = 0$. Conclude that $Z^\intercal \mathbb{1}_{N_0}$ has mean and variance zero so that it takes the value 0 with probability one. $\square$

By the previous discussion and the assumption that $\mathbb{P}(Z_n^\intercal \mathbb{1}_{N_0} \neq 0) = 0$ for all n sufficiently large, $v(Z_n, b_n) = v_\delta(Z_n, b_n)$ is almost surely finite for large n, where $b_n = (b_{ij,n})_{\substack{i,j\in[N_0]\\ i\neq j}}$ is the vector with

$$b_{ij,n} = 2(\|\bar{x}^{(i)}\|^2 - \|\bar{x}^{(j)}\|^2)^2 + 8(\bar{x}^{(i)} - \bar{x}^{(j)})^\intercal \mathbf{\Sigma}_{\bar{\mu}_{0,n}}(\bar{x}^{(i)} - \bar{x}^{(j)}), i, j \in [N_0], i \neq j,$$

which is known to be positive once $\mu_{0,n}$ is not a point mass (see Lemma 13), here $\bar{x}^{(i)} = x^{(i)} - \mathbb{E}_{\mu_{0,n}}[X]$ for $i = 1, \dots, N_0$. We define $\bar{b}$ by analogy with μ_0 in place of $\mu_{0,n}$ and remark that the same implications hold for $\text{LP}(Z, \bar{b})$ with $Z \sim N(0, \mathbf{\Sigma}_Z)$.

We now establish some technical lemmas which will be useful in the sequel.

Lemma 15. Fix $\delta > 0$, $c \in \mathbb{R}^{N_0}$ and $\beta \in \mathbb{R}_{>0}^{N_0(N_0-1)}$ satisfying $\beta_{ij} \in (\underline{\beta}, \overline{\beta})$ for some $0 < \underline{\beta} < \overline{\beta}$ for $i, j \in [N_0], i \neq j$. Then any dual solution, $p^\star$, to $\text{LP}_\delta(c, \beta)$ satisfies $\|p^\star\|_1 \leq \frac{\|c\|_1(\delta+\overline{\beta})}{\min\{\underline{\beta},\delta\}}$ and any primal solution $u^\star$ satisfies $\|u^\star\|_\infty \leq \delta + \overline{\beta}$.

Proof. The dual problem to $\text{LP}_\delta(c, \beta)$ is given by

$$\begin{aligned} &\inf_{p\in\mathbb{R}^{N_0(N_0-1)+2}} && \langle(\beta, \delta, \delta), p\rangle \\ &\text{subject to} && \mathbf{A}^\intercal p = c, \ p \geq 0, \end{aligned}$$

where $\mathbf{A}$ is the constraint matrix from the primal problem. By definition, $p^\star$ satisfies $v_\delta(c, \beta) = \langle(\beta, \delta, \delta), p^\star\rangle \geq \min(\beta, \delta, \delta)\|p^\star\|_1$ so that $\|p^\star\|_1 \leq \frac{v_\delta(c,\beta)}{\min\{\underline{\beta},\delta\}}$.

By the discussion at the beginning of this section, a solution to the primal problem, $u^\star$, satisfies $|u_1^\star| \leq \delta$ and $u_i^\star \in [-\delta - \underline{\beta}, \delta + \overline{\beta}]$ so that $\|u^\star\|_\infty \leq \delta + \overline{\beta}$. Conclude from Hölder's inequality that $v_\delta(c, \beta) \leq \|c\|_1(\delta + \overline{\beta})$. □

Lemma 16. Fix $\delta > 0$, $0 < \underline{\beta} < \overline{\beta}$, and let $D_{(\underline{\beta},\overline{\beta})} = \{\beta \in \mathbb{R}^{N_0(N_0-1)} : \beta_{ij} \in (\underline{\beta}, \overline{\beta}), i, j \in [N_0], i \neq j\}$. Then, if $\beta_n \to \beta \in D_{(\underline{\beta},\overline{\beta})}$, $v_\delta(\cdot, \beta_n) \to v_\delta(\cdot, \beta)$ uniformly on compact sets.

Proof. Applying Theorem 5.2 in [99] along with Lemma 15, we have that the map $(c,\beta) \in \mathbb{R}^{N_0} \times D_{[\underline{\beta},\bar{\beta}]} \mapsto v_\delta(c,\beta)$ is locally Lipschitz. As such, for any compact set $K \subset \mathbb{R}^{N_0}$, there exists a constant L_K for which

$$|v_\delta(c,\beta_n) - v_\delta(c,\beta)| \leq L_K \|\beta_n - \beta\|$$

for every $c \in K$ and n sufficiently large that $\beta_n \in D_{(\underline{\beta},\bar{\beta})}$, proving the claim. □

Proof of Theorem 18. By assumption, we have that, conditionally on $Y_1, Y_2, \ldots$, $\mathbb{E}_{\mu_{0,n}}[X] \to \mathbb{E}_{\mu_0}[X]$, $\mathbf{\Sigma}_{\bar{\mu}_{0,n}} \to \mathbf{\Sigma}_{\bar{\mu}_0}$, and $\mathbf{\Sigma}_n \to \mathbf{\Sigma}_Z$ deterministically given almost every realization of $Y_1, Y_2, \ldots$ so that $b_n \to b$ and $Z_n \sim N(0, \Sigma_n) \xrightarrow{d} Z \sim N(0, \Sigma_Z)$ conditionally on the data. It follows from the proof of Lemma 13 that $b \in D_{(\underline{b},\bar{b})}$ for some $0 < \underline{b} < \overline{b}$. By Lemma 16, $v_\delta(\cdot, b_n) \to v_\delta(\cdot, b)$ uniformly on compact sets conditionally on the data so that, applying the continuous mapping theorem, conditionally on $Y_1, Y_2, \ldots$, $v_\delta(Z_n, b_n) \xrightarrow{d} v_\delta(Z, b)$ given almost every realization of $Y_1, Y_2, \ldots$ As $Z^\intercal \mathbb{1}_{N_0} = 0$ and $Z_n^\intercal \mathbb{1}_{N_0} = 0$ with probability 1 for n sufficiently large, v_δ can be replaced by v so that, conditionally on $Y_1, Y_2, \ldots, L_n \xrightarrow{d} L_{\mu_0}$ given almost every realization of $Y_1, Y_2, \ldots$ The first claim follows by applying Lemma 2.11 in [200], noting that the distribution function of L_{μ_0} is continuous as follows from Proposition 12.1 in [57].

□

Proof of Theorem 11. If $\mu_{0,n} = \hat{\mu}_{0,n}$ and $\mathbf{\Sigma}_n = \mathbf{\Sigma}_{m_n}$, it follows from the law of large numbers that, conditionally on $Y_1, Y_2, \ldots, \mathbb{E}_{\mu_{0,n}}[X] \to \mathbb{E}_{\mu_0}[X]$, $\mathbf{\Sigma}_{\bar{\mu}_{0,n}} \to \mathbf{\Sigma}_{\bar{\mu}_0}$, and $\mathbf{\Sigma}_{m_n} \to \mathbf{\Sigma}_m$, given almost every realization of $Y_1, Y_2, \ldots$, where m is the weight vector for μ_0. By applying Lemma 14, the orthogonality condition for Z_n and Z always hold. The remaining assumptions regarding the structure of $\chi_{\mu_0 \otimes \mu_1}$ are verified in the proof of Corollary 5 save for the fact that $G_{\mu_0}(h) = Z^\intercal u_h$ for every $h \in \bar{\mathcal{H}}$ where $Z \sim N(0, \mathbf{\Sigma}_m)$, but this follows directly from the fact that G_{μ_0} is a μ_0-Brownian bridge process. □

3.8.3 Proofs for Section 3.4

Throughout, we will assume that the following variant of the twist condition holds.

Assumption 3. $\mu_0 \in \mathcal{P}(\mathcal{X}_0)$ is absolutely continuous with respect to the Lebesgue measure and, for every $i \neq j$ with $i, j \in \{1, \ldots, N\}$ and $t \in \mathbb{R}$, we have that

$$\mu_0\left(\left(c(\cdot, y^{(i)}) - c(\cdot, y^{(j)})\right)^{-1}(t)\right) = 0.$$

Specializing to $c_{\mathbf{A}}(x, y) = -4\|x\|^2\|y\|^2 - 32x^\intercal \mathbf{A} y$ from (3.8), the condition above is related to $\mathcal{X}_1$ and holds under the following primitive condition.

Lemma 17 (Necessary and sufficient condition on $\mathcal{X}_1$)**.** Let $\mu_0 \in \mathcal{P}(\mathcal{X}_0)$ be absolutely continuous with respect to the Lebesgue measure. Then, Assumption 3 is satisfied with $c_{\mathbf{A}}$ at $\mathbf{A} \in \mathbb{R}^{d_0 \times d_1}$ if and only if $\mathcal{X}_1$ is such that $y^{(i)} - y^{(j)} \notin \ker(\mathbf{A})$ for every $i \neq j \in \{1, \ldots, N\}$ with $\|y^{(i)}\| = \|y^{(j)}\|$. In particular, if $\|y^{(i)}\| \neq \|y^{(j)}\|$ for every $i \neq j$, then Assumption 3 holds for any $\mathbf{A} \in \mathbb{R}^{d_0 \times d_1}$.

Proof. Fix $i \neq j \in \{1, \ldots, N\}$ and $\mathbf{A} \in \mathbb{R}^{d_0} \times \mathbb{R}^{d_1}$. Then, $c_{\mathbf{A}}(x, y^{(i)}) - c_{\mathbf{A}}(x, y^{(j)}) = -4\|x\|^2(\|y^{(i)}\|^2 - \|y^{(j)}\|^2) - 32x^\intercal \mathbf{A}(y^{(i)} - y^{(j)})$ so that $c_{\mathbf{A}}(x, y^{(i)}) - c_{\mathbf{A}}(x, y^{(j)})$ is constant as a function of x on an open set $U \subset \mathbb{R}^{d_0}$ if and only if $-8x(\|y^{(i)}\|^2 - \|y^{(j)}\|^2) = 32\mathbf{A}(y^{(i)} - y^{(j)})$ for every $x \in U$. Observe that the prior equality holds on a nonnegligible set with respect to the Lebesgue measure if and only $y^{(i)} - y^{(j)} \in \ker(\mathbf{A})$ and $\|y^{(i)}\| = \|y^{(j)}\|$. Given that μ_0 is absolutely continuous with respect to the Lebesgue measure Assumption 3 holds at $\mathbf{A}$ if and only if $y^{(i)} - y^{(j)} \notin \ker(\mathbf{A})$ whenever $\|y^{(i)}\| = \|y^{(j)}\|$. $\square$

Lemma 17 provides a necessary and sufficient condition for Assumption 3 to hold at every $\mathbf{A} \in \mathbb{R}^{d_0 \times d_1}$ simultaneously. This is crucial for our study of limit distributions of the empirical semi-discrete GW distance to guarantee, e.g., uniqueness of OT plans for

$\mathsf{OT}_{\mathbf{A}}(\mu_0, \mu_1)$ uniformly in $\mathbf{A}$. This uniqueness will lead to a substantial simplification for the resulting limit, as compared to the discrete case (Theorem 10).

Under Assumption 3, solutions of $\mathsf{OT}_{\mathbf{A}}(\mu_0, \mu_1)$ admit the following characterization.

Lemma 18 (On solutions of $\mathsf{OT}_{\mathbf{A}}(\mu_0, \mu_1)$)**.** Fix $\mathbf{A} \in \mathbb{R}^{d_0 \times d_1}$ and let $z^{\mathbf{A}}$ be an optimal vector for $\mathsf{OT}_{\mathbf{A}}(\mu_0, \mu_1)$. Under Assumption 3, the OT plan is unique and is induced by the map

$$T_{\mathbf{A}} : x \in \mathcal{X} \mapsto y^{(I_{z^{\mathbf{A}}}(x))}, \text{ where } I_{z^{\mathbf{A}}}(x) \in \operatorname*{argmin}_{1 \leq i \leq N} \left(c_{\mathbf{A}}(x, y^{(i)}) - z_i^{\mathbf{A}} \right), \tag{3.36}$$

which is uniquely defined μ_0-a.e.

Proof. We first address uniqueness of the OT plan. For any $z \in \mathbb{R}^N$, let $\phi_z : x \in \mathcal{X}_0 \mapsto \min_{1 \leq i \leq N} \left\{ c_{\mathbf{A}}(x, y^{(i)}) - z_i \right\}$. If $x \in \mathcal{X}_0$ is such that $\phi_z(x) + \phi_z^c(y^{(i)}) = c_{\mathbf{A}}(x, y^{(i)})$ and $\phi_z(x) + \phi_z^c(y^{(j)}) = c_{\mathbf{A}}(x, y^{(j)})$ for some $i \neq j \in [N]$, then $c_{\mathbf{A}}(x, y^{(j)}) - c_{\mathbf{A}}(x, y^{(i)}) = \phi_z^c(y^{(j)}) - \phi_z^c(y^{(i)})$. Assumption 3 guarantees that the previous equality occurs on a μ_0-negligible set so that we can apply Theorem 5.30 in [208] whereby the OT plan for $\mathsf{OT}_{\mathbf{A}}(\mu_0, \mu_1)$ is unique and induced by a map.

We now derive the expression for the optimal map. Let $z^{\mathbf{A}}$ be an optimal vector for $\mathsf{OT}_{\mathbf{A}}(\mu_0, \mu_1)$ and $\varphi_0^{\mathbf{A}}$ be the corresponding OT potential. For convenience, set $\zeta(y^{(i)}) = z_i^{\mathbf{A}}$ for $i = 1, \ldots, N$. By strong duality,

$$\int \varphi_0^{\mathbf{A}}(x) + \zeta(y) - c_{\mathbf{A}}(x, y) d\pi_{\mathbf{A}}(x, y) = 0.$$

As $\varphi_0^{\mathbf{A}}(x) + \zeta(y) \leq c_{\mathbf{A}}(x, y)$ for every $(x, y) \in \mathcal{X}_0 \times \mathcal{X}_1$ by definition, it follows that $\varphi_0^{\mathbf{A}}(x) + \zeta(y) = c_{\mathbf{A}}(x, y)$ $\pi_{\mathbf{A}}$-a.e. As $\pi_{\mathbf{A}}$ is induced by a map $T_{\mathbf{A}} : \mathcal{X}_0 \to \mathcal{X}_1$, $\varphi_0^{\mathbf{A}}(x) = c_{\mathbf{A}}(x, T_{\mathbf{A}}(x)) - \zeta(T_{\mathbf{A}}(x))$ μ_0-a.e. Since $\varphi_0^{\mathbf{A}}(x) = \min_{1 \leq i \leq N} \left\{ c_{\mathbf{A}}(x, y^{(i)}) - \zeta(y^{(i)}) \right\}$, it holds that $T_{\mathbf{A}}(x) = y^{I_{z^{\mathbf{A}}}(x)}$ (recalling that $I_{z^{\mathbf{A}}}(x) \in \operatorname{argmin}_{1 \leq i \leq N} \left\{ c_{\mathbf{A}}(x, y^{(i)}) - \zeta(y^{(i)}) \right\}$). Assumption 3 further guarantees that $I_{z^{\mathbf{A}}}(x)$ is a singleton for μ_0-a.e. $x \in \mathcal{X}_0$.

$\square$

Theorem 12 is a direct consequence of Lemmas 17 and 18.

Proof of Theorem 13

In this section, the centering required to access the variational form of the GW distance (3.8) will play a prominent role. Of note is that if $(\nu_0, \nu_1) \in \mathcal{P}(\mathcal{X}_0) \times \mathcal{P}(\mathcal{X}_1)$, we generally do not have that $\mathrm{spt}(\bar{\nu}_0)$ or $\mathrm{spt}(\bar{\nu}_1)$ are contained in $\mathcal{X}_0$ and $\mathcal{X}_1$ respectively. In order to compare OT potentials arising in the relevant variational forms, we thus extend them to sets containing $\mathcal{X}_0 - \mathbb{E}_{\nu_0}[X]$ and $\mathcal{X}_1 - \mathbb{E}_{\nu_1}[X]$.

Definition 6 (Extended OT potentials)**.** For $i = 0, 1$, let $\mathcal{X}_i^\circ$ be open balls centered at 0 with radius $r > 2\|\mathcal{X}_i\|_\infty$. Fix arbitrary $(\nu_0, \nu_1) \in \mathcal{P}(\mathcal{X}_0) \times \mathcal{P}(\mathcal{X}_1)$ and $\mathbf{A} \in \mathbb{R}^{d_0 \times d_1}$, let $z^{\mathbf{A}}$ be an optimal vector for $\mathsf{OT}_{\mathbf{A}}(\bar{\nu}_0, \bar{\nu}_1)$. Then, the extended OT potentials for $\mathsf{OT}_{\mathbf{A}}(\bar{\nu}_0, \bar{\nu}_1)$ are given by

$$\varphi_0^{\mathbf{A}} : x \in \mathcal{X}_0^\circ \mapsto \min_{1 \le i \le N} \left\{ c_{\mathbf{A}}(x, y^{(i)} - \mathbb{E}_{\nu_1}[X]) - z_i^{\mathbf{A}} \right\},$$
$$\varphi_1^{\mathbf{A}} : y \in \mathcal{X}_1^\circ \mapsto \inf_{x \in \mathcal{X}_0^\circ} \left\{ c_{\mathbf{A}}(x, y) - \varphi_0^{\mathbf{A}}(x) \right\}.$$

As $\varphi_1^{\mathbf{A}}$ is simply the $c_{\mathbf{A}}$-transform of $\varphi_0^{\mathbf{A}}$, $(\varphi_0^{\mathbf{A}}, \varphi_1^{\mathbf{A}})$ is still a pair of OT potentials for $\mathsf{OT}_{\mathbf{A}}(\bar{\nu}_0, \bar{\nu}_1)$. The definitions of $\mathcal{X}_0^\circ$ and $\mathcal{X}_1^\circ$ are motivated by the fact that if $\nu_i \in \mathcal{P}(\mathcal{X}_i)$, then $\|\mathbb{E}_{\nu_i}[X]\| \le \mathbb{E}_{\nu_i}[\|X\|] \le \|\mathcal{X}_i\|_\infty$ for $i \in \{0, 1\}$, so that $\mathcal{X}_i - \mathbb{E}_{\nu_i}[X] \subset \mathcal{X}_i^\circ$ and thus $\bar{\nu}_i \in \mathcal{P}(\mathcal{X}_i^\circ)$.

We first establish some useful properties for the extended OT potentials.

Lemma 19 (Properties of extended OT potentials)**.** Fix $(\nu_0, \nu_1) \in \mathcal{P}(\mathcal{X}_0) \times \mathcal{P}(\mathcal{X}_1)$ and $\mathbf{A} \in B_{\mathrm{F}}(M)$. Let $z^{\mathbf{A}}$ be an optimal vector for $\mathsf{OT}_{\mathbf{A}}(\bar{\nu}_0, \bar{\nu}_1)$ and $(\varphi_0^{\mathbf{A}}, \varphi_1^{\mathbf{A}})$ be the extended OT potentials. Then,

1. $\varphi_0^{\mathbf{A}}$ and $\varphi_1^{\mathbf{A}}$ are Lipschitz continuous with a shared constant, L, depending only on

$M, \|\mathcal{X}_0^\circ\|_\infty, \|\mathcal{X}_1^\circ\|_\infty$.

2. the triple $(z^{\mathbf{A}}, \varphi_0^{\mathbf{A}}, \varphi_1^{\mathbf{A}})$ can be chosen such that $\|z^{\mathbf{A}}\|_\infty \vee \|\varphi_0^{\mathbf{A}}\|_{\infty,\mathcal{X}_0^\circ} \vee \|\varphi_1^{\mathbf{A}}\|_{\infty,\mathcal{X}_1^\circ} \leq K$, where K depends only on $M, \|\mathcal{X}_0^\circ\|_\infty, \|\mathcal{X}_1^\circ\|_\infty$.

3. if $\mathrm{int}(\mathrm{spt}(\nu_0))$ is connected with (Lebesgue) negligible boundary and ν_0 is absolutely continuous with respect to the Lebesgue measure, then the pair $(\varphi_0^{\mathbf{A}}, \varphi_1^{\mathbf{A}})$ is unique up to additive constants on $\mathrm{spt}(\nu_0) \times \mathrm{spt}(\nu_1)$.

Proof. For (1), fix $x, x' \in \mathcal{X}_0^\circ$ and observe that

$$\begin{aligned}\left|\varphi_0^{\mathbf{A}}(x) - \varphi_0^{\mathbf{A}}(x')\right| &\leq \max_{1\leq i\leq N} \left|c_{\mathbf{A}}\left(x, y^{(i)} - \mathbb{E}_{\nu_1}[X]\right) - c_{\mathbf{A}}\left(x', y^{(i)} - \mathbb{E}_{\nu_1}[X]\right)\right| \\ &= \max_{1\leq i\leq N} \Big|-4\|x\|^2\|y^{(i)} - \mathbb{E}_{\nu_1}[X]\|^2 - 32x^\intercal\mathbf{A}(y^{(i)} - \mathbb{E}_{\nu_1}[X]) \\ &+4\|x'\|^2\|y^{(i)} - \mathbb{E}_{\nu_1}[X]\|^2 + 32x'^\intercal\mathbf{A}(y^{(i)} - \mathbb{E}_{\nu_1}[X])\Big| \\ &\leq 4|\|x\|^2 - \|x'\|^2| \max_{i=1}^{N} \|y^{(i)} - \mathbb{E}_{\nu_1}[X]\|^2 \\ &+ 32\|x - x'\|M \max_{i=1}^{n} \|y^{(i)} - \mathbb{E}_{\nu_1}[X]\| \\ &\leq L_0\|x - x'\|,\end{aligned}$$

as $|\|x\|^2 - \|x'\|^2| = |(\|x\| + \|x'\|)(\|x\| - \|x'\|)| \leq 2\|\mathcal{X}_0^\circ\|_\infty\|x - x'\|$ due to the reverse triangle inequality. Since $\|y^{(i)} - \mathbb{E}_{\nu_1}[X]\| \leq \|\mathcal{X}_1^\circ\|_\infty$, L_0 may be chosen independently of $\mathbf{A}, \nu_0, \nu_1$.

The same approach can be used to derive a Lipschitz constant for $\varphi_1^{\mathbf{A}}$; the maximum of the two Lipschitz constants serves as a shared Lipschitz constant for both OT potentials.

For (2), fix a solution $(z^{\mathbf{A}}, \varphi_0^{\mathbf{A}}, \varphi_1^{\mathbf{A}})$ for $\mathsf{OT}_{\mathbf{A}}(\bar{\nu}_0, \bar{\nu}_1)$ and assume that $\mathrm{spt}(\nu_1) = \mathcal{X}_1$ such that

$$\int \varphi_0^{\mathbf{A}} d\bar{\mu}_0 + \sum_{i=1}^{N} z_i^{\mathbf{A}} \bar{\mu}_1(\{y^{(i)}\}) = \int \varphi_0^{\mathbf{A}} d\bar{\mu}_0 + \int \varphi_1^{\mathbf{A}} d\bar{\mu}_1 = \mathsf{OT}_{\mathbf{A}}(\bar{\mu}_0, \bar{\mu}_1)$$

which, coupled with the condition $\varphi_0^{\mathbf{A}} \oplus \varphi_1^{\mathbf{A}} \leq c_{\mathbf{A}}$, implies that $\varphi_0^{\mathbf{A}} \oplus \varphi_1^{\mathbf{A}} = c_{\mathbf{A}}$ $\pi^{\mathbf{A}}$-a.e. for any choice of optimal coupling $\pi^{\mathbf{A}}$ for $\mathsf{OT}_{\mathbf{A}}(\bar{\mu}_0, \bar{\mu}_1)$ and similarly for $(z^{\mathbf{A}}, \varphi_0^{\mathbf{A}})$. In particular, if

$\mathrm{spt}(\nu_1) = \mathcal{X}_1$ there exists some $x^{(i)} \in \mathrm{spt}(\bar{\mu}_0)$ for which $\varphi_0^{\mathbf{A}}(x^{(i)}) + z_i^{\mathbf{A}} = c_{\mathbf{A}}(x^{(i)}, y^{(i)} - \mathbb{E}_{\nu_1}[X])$ for every $i \in [N]$ and $\varphi_0^{\mathbf{A}}(x) \leq c_{\mathbf{A}}(x, y^{(i)} - \mathbb{E}_{\nu_1}[X]) - z_i^{\mathbf{A}}$ for any other $x \in \mathcal{X}_0^\circ$. Thus,

$$\varphi_1^{\mathbf{A}}(y^{(i)} - \mathbb{E}_{\nu_1}[X]) = \inf_{x \in \mathcal{X}_0^\circ} \left\{ c_{\mathbf{A}}(x, y^{(i)} - \mathbb{E}_{\nu_1}[X]) - \varphi_0^{\mathbf{A}}(x) \right\} \geq z_i^{\mathbf{A}}$$

with equality when $x = x^{(i)}$ i.e. $\varphi_1^{\mathbf{A}}(y^{(i)} - \mathbb{E}_{\nu_1}[X]) = z_i^{\mathbf{A}}$ for every $i \in [N]$. Now, let $C = -\max_{1\leq i\leq N} z_i^{\mathbf{A}}$ and set $(\tilde{z}^{\mathbf{A}}, \tilde{\varphi}_0^{\mathbf{A}}, \tilde{\varphi}_1^{\mathbf{A}}) = (z^{\mathbf{A}}+C, \varphi_0^{\mathbf{A}}-C, \varphi_1^{\mathbf{A}}+C)$ such that $\max_{1\leq i\leq N} \tilde{z}_i^{\mathbf{A}} = 0$. With this normalization, for any $(x, y) \in \mathcal{X}_0^\circ \times \mathcal{X}_1^\circ$,

$$-\|c_{\mathbf{A}}\|_{\infty, \mathcal{X}_0^\circ \times \mathcal{X}_1^\circ} \leq \tilde{\varphi}_0^{\mathbf{A}}(x) = \min_{1\leq i\leq N} \left\{ c_{\mathbf{A}}(x, y^{(i)} - \mathbb{E}_{\nu_1}[X]) - \tilde{z}_i^{\mathbf{A}} \right\} \leq \|c_{\mathbf{A}}\|_{\infty, \mathcal{X}_0^\circ \times \mathcal{X}_1^\circ},$$

$$\begin{aligned} -\|c_{\mathbf{A}}\|_{\infty, \mathcal{X}_0^\circ \times \mathcal{X}_1^\circ} - \sup_{x \in \mathcal{X}_0^\circ} \tilde{\varphi}_0^{\mathbf{A}}(x) \leq \tilde{\varphi}_1^{\mathbf{A}}(y) &= \inf_{x \in \mathcal{X}_0^\circ} \left\{ c_{\mathbf{A}}(x, y) - \tilde{\varphi}_0^{\mathbf{A}}(x) \right\} \\ &\leq \|c_{\mathbf{A}}\|_{\infty, \mathcal{X}_0^\circ \times \mathcal{X}_1^\circ} - \sup_{x \in \mathcal{X}_0^\circ} \tilde{\varphi}_0^{\mathbf{A}}(x), \end{aligned}$$

so that $\|\tilde{z}^{\mathbf{A}}\|_\infty \vee \|\tilde{\varphi}_0^{\mathbf{A}}\|_{\infty, \mathcal{X}_0^\circ} \vee \|\tilde{\varphi}_1^{\mathbf{A}}\|_{\infty, \mathcal{X}_1^\circ} \leq 2\|c_{\mathbf{A}}\|_{\infty, \mathcal{X}_0^\circ \times \mathcal{X}_1^\circ}$ where, by the previous arguments, $\|\tilde{z}^{\mathbf{A}}\|_\infty \leq \|\tilde{\varphi}_1^{\mathbf{A}}\|_{\infty, \mathcal{X}_1^\circ}$. $\|c_{\mathbf{A}}\|_{\infty, \mathcal{X}_0^\circ \times \mathcal{X}_1^\circ}$ can be bounded as $4\|\mathcal{X}_0^\circ\|_\infty^2 \|\mathcal{X}_1^\circ\|_\infty^2 + 32M\|\mathcal{X}_0^\circ\|_\infty \|\mathcal{X}_1^\circ\|_\infty$ via the Cauchy-Schwarz inequality, proving the claim.

If $\mathrm{spt}(\nu_1) \neq \mathcal{X}_1$, one may set $z_i^{\mathbf{A}} = 0$ if $y^{(i)} \notin \mathrm{spt}(\nu_1)$ and the previous arguments can be applied to achieve the same bound, though the correspondence between $\tilde{\varphi}_1^{\mathbf{A}}$ and $z^{\mathbf{A}}$ will only hold on $\mathrm{spt}(\bar{\nu}_1)$.

For (3), suppose that $(\tilde{\varphi}_0^{\mathbf{A}}, \tilde{\varphi}_1^{\mathbf{A}})$ is another pair of extended OT potentials.By Rademacher's theorem, $\varphi_0^{\mathbf{A}}$ and $\tilde{\varphi}_0^{\mathbf{A}}$ are differentiable almost everywhere on $\mathcal{X}_0^\circ$ (in particular on $\mathrm{int}(\mathrm{spt}(\nu_0))$) so that Proposition 1.15 in [175] implies that $\nabla\varphi_0^{\mathbf{A}} = \nabla\tilde{\varphi}_0^{\mathbf{A}}$ almost everywhere on $\mathrm{int}(\mathrm{spt}(\nu_0))$. As $\varphi_0^{\mathbf{A}}$ and $\tilde{\varphi}_0^{\mathbf{A}}$ are Lipschitz continuous, Theorem 2.6 in [61] implies that $\varphi_0^{\mathbf{A}} - \tilde{\varphi}_0^{\mathbf{A}} = a$ on $\mathrm{int}(\mathrm{spt}(\nu_0))$ for some constant $a \in \mathbb{R}$. This equality extends to $\mathrm{spt}(\nu_0)$ again due to Lipschitz continuity.

As for $\varphi_1^{\mathbf{A}}$ and $\tilde{\varphi}_1^{\mathbf{A}}$, it holds that $\varphi_0^{\mathbf{A}}(x) + \varphi_1^{\mathbf{A}}(y) = c_{\mathbf{A}}(x, y)$ $\pi^{\mathbf{A}}$-almost everywhere, where

$\pi^{\mathbf{A}}$ is any OT plan for $\mathsf{OT}_{\mathbf{A}}(\nu_0, \nu_1)$. $(\tilde{\varphi}_0^{\mathbf{A}}, \tilde{\varphi}_1^{\mathbf{A}})$ satisfy the same equality so that

$$\varphi_0^{\mathbf{A}}(x) + \varphi_1^{\mathbf{A}}(y) = \tilde{\varphi}_0^{\mathbf{A}}(x) + a + \varphi_1^{\mathbf{A}}(y) = \tilde{\varphi}_0^{\mathbf{A}}(x) + \tilde{\varphi}_1^{\mathbf{A}}(y)$$

$\pi^{\mathbf{A}}$- almost everywhere i.e. $\tilde{\varphi}_1^{\mathbf{A}}(y) = \varphi_1^{\mathbf{A}}(y) + a$ ν_1-almost everywhere, proving the claim. □

To compute the relevant directional derivative, we fix $\nu_0 \otimes \nu_1 \in \mathcal{P}_{\mu_0} \times \mathcal{P}_{\mu_1}$, set $\mu_{i,t} = \mu_i + t(\nu_i - \mu_i)$ for $i = 0, 1$ and $t \in [0, 1]$, and decompose the limit as follows

$$\begin{aligned}
&\lim_{t \downarrow 0} \frac{\mathsf{D}(\mu_{0,t}, \mu_{1,t})^2 - \mathsf{D}(\mu_0, \mu_1)^2}{t} \\
&\qquad = \lim_{t \downarrow 0} \frac{\mathsf{D}(\bar{\mu}_{0,t}, \bar{\mu}_{1,t})^2 - \mathsf{D}(\bar{\mu}_0, \bar{\mu}_1)^2}{t} \\
&\qquad = \lim_{t \downarrow 0} \frac{\mathsf{S}_1(\bar{\mu}_{0,t}, \bar{\mu}_{1,t}) - \mathsf{S}_1(\bar{\mu}_0, \bar{\mu}_1)}{t} + \frac{\mathsf{S}_2(\bar{\mu}_{0,t}, \bar{\mu}_{1,t}) - \mathsf{S}_2(\bar{\mu}_0, \bar{\mu}_1)}{t}
\end{aligned}$$

as follows from the variational form. The following lemmas compute each limit separately.

Throughout, all extended OT potentials are chosen as to satisfy the estimates from Lemma 19.

Lemma 20. As $t \downarrow 0$, we have that

$$\begin{aligned}
\frac{\mathsf{S}_1(\bar{\mu}_{0,t}, \bar{\mu}_{1,t}) - \mathsf{S}_1(\bar{\mu}_0, \bar{\mu}_1)}{t} \to\ & 2 \iint \|x - x'\|^4 d\mu_0(x) d(\nu_0 - \mu_0)(x') \\
&+ 2 \iint \|y - y'\|^4 d\mu_1(y) d(\nu_1 - \mu_1)(y') \\
&- 4 \iint \|x - \mathbb{E}_{\mu_0}[X]\|^2 \|y - \mathbb{E}_{\mu_1}[X]\|^2 d(\nu_0 - \mu_0)(x) d\mu_1(y) \\
&- 4 \iint \|x - \mathbb{E}_{\mu_0}[X]\|^2 \|y - \mathbb{E}_{\mu_1}[X]\|^2 d\mu_0(x) d(\nu_1 - \mu_1)(y).
\end{aligned}$$

Proof. For any probability measure η on $\mathbb{R}^{d_i}$ $(i = 0, 1)$ with finite fourth moment,

$$\iint \|x - x'\|^4 d\bar{\eta}(x) d\bar{\eta}(x) = \iint \|x - x'\|^4 d\eta(x) d\eta(x),$$

hence, for $i = 0, 1$,

$$\iint \|x - x'\|^4 d\bar{\mu}_{i,t}(x) d\bar{\mu}_{i,t}(x') - \iint \|x - x'\|^4 d\bar{\mu}_i(x) d\bar{\mu}_i(x')$$
$$= 2t \iint \|x - x'\|^4 d\mu_i(x) d(\nu_i - \mu_i)(x') + t^2 \iint \|x - x'\|^4 d(\nu_i - \mu_i)(x) d(\nu_i - \mu_i)(x'). \tag{3.37}$$

On the other hand, letting $f_i(t) = \mathbb{E}_{\mu_i}[X] + t\left(\mathbb{E}_{\nu_i}[X] - \mathbb{E}_{\mu_i}[X]\right)$ for $i = 0, 1$ and $t \in [0, 1]$,

$$\begin{aligned}
&\iint \|x\|^2 \|y\|^2 d\bar{\mu}_{0,t}(x) d\bar{\mu}_{1,t}(y) \\
&\qquad = \iint \|x - f_0(t)\|^2 \|y - f_1(t)\|^2 d\mu_{0,t}(x) d\mu_{1,t}(y) \\
&\qquad = \iint \|x - f_0(t)\|^2 \|y - f_1(t)\|^2 d\mu_0(x) d\mu_1(y) \\
&\qquad + t \iint \|x - f_0(t)\|^2 \|y - f_1(t)\|^2 d(\nu_0 - \mu_0)(x) d\mu_1(y) \\
&\qquad + t \iint \|x - f_0(t)\|^2 \|y - f_1(t)\|^2 d\mu_0(x) d(\nu_1 - \mu_1)(y) \\
&\qquad + t^2 \iint \|x - f_0(t)\|^2 \|y - f_1(t)\|^2 d(\nu_0 - \mu_0)(x) d(\nu_1 - \mu_1)(y).
\end{aligned}$$

As $\|x - f_i(t)\|^2 = \|x - \mathbb{E}_{\mu_i}[X]\|^2 - 2t\langle x - \mathbb{E}_{\mu_i}[X], \mathbb{E}_{\nu_i}[X] - \mathbb{E}_{\mu_i}[X]\rangle + t^2 \|\mathbb{E}_{\nu_i}[X] - \mathbb{E}_{\mu_i}[X]\|^2$ for $i = 0, 1$ and the term involving inner products integrates to 0 w.r.t. μ_i,

$$\begin{aligned}
\iint \|x\|^2 \|y\|^2 d\bar{\mu}_{0,t}(x) d\bar{\mu}_{1,t}(y) &= \iint \|x\|^2 \|y\|^2 d\bar{\mu}_0(x) d\bar{\mu}_1(y) \\
&+ t \iint \|x - \mathbb{E}_{\mu_0}[X]\|^2 \|y - \mathbb{E}_{\mu_1}[X]\|^2 d(\nu_0 - \mu_0)(x) d\mu_1(y) \\
&+ t \iint \|x - \mathbb{E}_{\mu_0}[X]\|^2 \|y - \mathbb{E}_{\mu_1}[X]\|^2 d\mu_0(x) d(\nu_1 - \mu_1)(y) \\
&+ o(t).
\end{aligned}$$

The above display and (3.37) prove the claim. $\square$

Lemma 21. For any pairs $(\nu_0, \nu_1), (\rho_0, \rho_1) \in \mathcal{P}_{\mu_0} \times \mathcal{P}_{\mu_1}$, $|\mathsf{S}_1(\bar{\nu}_0, \bar{\nu}_1) - \mathsf{S}_1(\bar{\rho}_0, \bar{\rho}_1)| \le C\|\nu_0 \otimes \nu_1 - \rho_0 \otimes \rho_1\|_{\infty, \mathcal{F}^{\oplus}}$ for some constant $C > 0$ which is independent of $(\nu_0, \nu_1), (\rho_0, \rho_1)$.

Proof. For $i = 0, 1$, we have that

$$\begin{aligned}\iint \|x - x'\|^4 d\bar{\nu}_i(x) d\bar{\nu}_i(x') &= \iint \|x - x'\|^4 d\nu_i(x) d\nu_i(x') \\ &= \iint \|x - x'\|^4 d(\nu_i - \rho_i)(x) d\nu_i(x') \\ &+ \iint \|x - x'\|^4 d(\nu_i - \rho_i)(x) d\rho_i(x') \\ &+ \iint \|x - x'\|^4 d\bar{\rho}_i(x) d\bar{\rho}_i(x'),\end{aligned}$$

such that

$$\begin{aligned}&\left| \sum_{i=0}^{1} \iint \|x - x'\|^4 d\bar{\nu}_i(x) d\bar{\nu}_i(x') - \iint \|x - x'\|^4 d\bar{\rho}_i(x) d\bar{\rho}_i(x') \right| \\ &\qquad = \sum_{i=0}^{1} (\nu_i - \rho_i) \left(\int \|\cdot - x'\|^4 d\nu_i(x') + \int \|\cdot - x'\|^4 d\rho_i(x') \right).\end{aligned}$$

Observe that $\xi_0 : x \in \mathcal{X}_0 \mapsto \int \|x - x'\|^4 d\nu_0(x') + \int \|x - x'\|^4 d\rho_0(x')$ is smooth with uniformly bounded derivatives of all orders so that there exists a positive constant C_0 which is independent of ν_0, ρ_0 for which $C_0 \xi_0 \in \mathcal{F}_0$. Similarly, $\xi_1 : x \in \mathcal{X}_1 \mapsto \int \|x - x'\|^4 d\nu_1(x') + \int \|x - x'\|^4 d\rho_1(x')$ is uniformly bounded so that there exists a positive constant C_1 which is independent of ν_1, ρ_1 for which $C_1 \xi_1 \in \mathcal{F}_1$. Conclude that

$$\begin{aligned}&\left| \sum_{i=0}^{1} \iint \|x - x'\|^4 d\bar{\nu}_i(x) d\bar{\nu}_i(x') - \iint \|x - x'\|^4 d\bar{\rho}_i(x) d\bar{\rho}_i(x') \right| \\ &\qquad = \left| C_0^{-1} (\nu_0 - \rho_0)(C_0 \xi_0) + C_1^{-1} (\nu_1 - \rho_1)(C_1 \xi_1) \right| \\ &\qquad \leq C_0^{-1} \left| (\nu_0 - \rho_0)(C_0 \xi_0) \right| + C_1^{-1} \left| (\nu_1 - \rho_1)(C_1 \xi_1) \right| \\ &\qquad \leq C_0^{-1} \sup_{f_0 \in \mathcal{F}_0} \left| (\nu_0 - \rho_0)(f_0) \right| + C_1^{-1} \sup_{f_1 \in \mathcal{F}_1} \left| (\nu_1 - \rho_1)(f_1) \right| \\ &\qquad \leq (C_0^{-1} + C_1^{-1}) \|\nu_0 \otimes \nu_1 - \rho_0 \otimes \rho_1\|_{\infty, \mathcal{F}^{\oplus}},\end{aligned} \tag{3.38}$$

where we have used the fact that $0 \in \mathcal{F}_1$ so that

$$\sup_{f_0 \in \mathcal{F}_0} \left| (\nu_0 - \rho_0)(f_0) \right| = \sup_{f_0 \in \mathcal{F}_0} \left| (\nu_0 - \rho_0)(f_0) + (\nu_1 - \rho_1)(0) \right| \leq \|\nu_0 \otimes \nu_1 - \rho_0 \otimes \rho_1\|_{\infty, \mathcal{F}^{\oplus}}, \tag{3.39}$$

the same implications hold for $\sup_{f_1 \in \mathcal{F}_1} |(\nu_1 - \rho_1)(f_1)|$.

As for the other term,

$$\begin{aligned}\int \|x\|^2 d\bar{\nu}_0(x) \int \|x\|^2 d\bar{\nu}_1(x) &= \int \|x - \mathbb{E}_{\nu_0}[X]\|^2 d\nu_0(x) \int \|x - \mathbb{E}_{\nu_1}[X]\|^2 d\nu_1(x) \\ &= \int \|x - \mathbb{E}_{\nu_0}[X]\|^2 d(\nu_0 - \rho_0)(x) \int \|x - \mathbb{E}_{\nu_1}[X]\|^2 d\nu_1(x) \\ &+ \int \|x - \mathbb{E}_{\nu_0}[X]\|^2 d\rho_0(x) \int \|x - \mathbb{E}_{\nu_1}[X]\|^2 d(\nu_1 - \rho_1)(x) \\ &+ \int \|x - \mathbb{E}_{\nu_0}[X]\|^2 d\rho_0(x) \int \|x - \mathbb{E}_{\nu_1}[X]\|^2 d\rho_1(x),\end{aligned}$$

where $\zeta_0 : x \in \mathcal{X}_0 \mapsto \|x - \mathbb{E}_{\mu_0}[X]\|^2$ is smooth with uniformly bounded derivatives of all orders so that there exists a positive constant C_0' which is independent of ν_0, ρ_0 for which $C_0'\zeta_0 \in \mathcal{F}_0$. Similarly, $\zeta_1 : x \in \mathcal{X}_1 \mapsto \int \|x - \mathbb{E}_{\nu_1}[X]\|^2$ is uniformly bounded so that there exists a positive constant C_1' which is independent of ν_1, ρ_1 for which $C_1'\zeta_1 \in \mathcal{F}_1$. Conclude that

$$\begin{aligned}&\left|\int \|x\|^2 d\bar{\nu}_0(x) \int \|x\|^2 d\bar{\nu}_1(x) - \int \|x\|^2 d\bar{\rho}_0(x) \int \|x\|^2 d\bar{\rho}_1(x)\right| \\ &= \Bigg| M_2(\bar{\nu}_1) \int \|x - \mathbb{E}_{\nu_0}[X]\|^2 d(\nu_0 - \rho_0) \\ &\qquad + \int \|x - \mathbb{E}_{\nu_0}[X]\|^2 d\rho_0(x) \int \|x - \mathbb{E}_{\nu_1}[X]\|^2 d(\nu_1 - \rho_1)(x) \Bigg| \\ &\leq M_2(\bar{\nu}_1)\,|(\nu_0 - \rho_0)(\zeta_0)| + \int \|x - \mathbb{E}_{\nu_0}[X]\|^2 d\rho_0(x)\,|(\nu_1 - \rho_1)(\zeta_1)| \\ &\leq \left(M_2(\bar{\nu}_1)(C_0')^{-1} + \int \|x - \mathbb{E}_{\nu_0}[X]\|^2 d\rho_0(x)(C_1')^{-1} \right) \|\nu_0 \otimes \nu_1 - \rho_0 \otimes \rho_1\|_{\infty, \mathcal{F}^{\oplus}},\end{aligned}$$

where the final inequality is due to (3.39). This final constant can be bounded independently of $\rho_0, \rho_1, \nu_0, \nu_1$ so that the above display and (3.38) yield the desired result. □

Lemma 22. If Assumption 3 holds at $\mathbf{A} \in \mathbb{R}^{d_0 \times d_1}$ and $\mathrm{int}(\mathrm{spt}(\bar{\mu}_0))$ is connected with negligible boundary, then

$$\begin{aligned}\lim_{t \downarrow 0} \frac{\mathsf{OT}_{\mathbf{A}}(\bar{\mu}_{0,t}, \bar{\mu}_{1,t}) - \mathsf{OT}_{\mathbf{A}}(\bar{\mu}_0, \bar{\mu}_1)}{t} &= \int \bar{\varphi}_0^{\mathbf{A}} + 8\mathbb{E}_{(X,Y)\sim\pi^{\mathbf{A}}}[\|Y\|^2 X]^{\intercal}(\cdot) d(\nu_0 - \mu_0) \\ &+ \int \bar{\varphi}_1^{\mathbf{A}} + 8\mathbb{E}_{(X,Y)\sim\pi^{\mathbf{A}}}[\|X\|^2 Y]^{\intercal}(\cdot) d(\nu_1 - \mu_1).\end{aligned}$$

Proof. Fix $\mathbf{A} \in \mathbb{R}^{d_0 \times d_1}$ and consider the limit

$$\lim_{t \downarrow 0} \frac{\mathsf{OT}_{\mathbf{A}}(\bar{\mu}_{0,t}, \bar{\mu}_{1,t}) - \mathsf{OT}_{\mathbf{A}}(\bar{\mu}_0, \bar{\mu}_1)}{t} = \lim_{t \downarrow 0} \Bigg(\frac{\mathsf{OT}_{\mathbf{A}}(\bar{\mu}_{0,t}, \bar{\mu}_{1,t}) - \mathsf{OT}_{\mathbf{A}}(\tilde{\mu}_{0,t}, \tilde{\mu}_{1,t})}{t} + \frac{\mathsf{OT}_{\mathbf{A}}(\tilde{\mu}_{0,t}, \tilde{\mu}_{1,t}) - \mathsf{OT}_{\mathbf{A}}(\bar{\mu}_0, \bar{\mu}_1)}{t} \Bigg)$$

where $(\tilde{\mu}_{0,t}, \tilde{\mu}_{1,t}) = \Big((\mathrm{Id} - \mathbb{E}_{\mu_0}[X])_\sharp \mu_{0,t}, (\mathrm{Id} - \mathbb{E}_{\mu_1}[X])_\sharp \mu_{1,t} \Big)$. We compute each limit on the right hand side separately.

First, let $(\varphi_0^{\mathbf{A}}, \varphi_1^{\mathbf{A}})$ and $(\varphi_{0,t}^{\mathbf{A}}, \varphi_{1,t}^{\mathbf{A}})$ be OT potentials for $\mathsf{OT}_{\mathbf{A}}(\bar{\mu}_0, \bar{\mu}_1)$ and $\mathsf{OT}_{\mathbf{A}}(\tilde{\mu}_{0,t}, \tilde{\mu}_{1,t})$ satisfying the bounds from Lemma 19 (though $(\tilde{\mu}_{0,t}, \tilde{\mu}_{1,t})$ are not centered versions of measures in $\mathcal{P}(\mathcal{X}_0) \times \mathcal{P}(\mathcal{X}_1)$, they are supported in $\mathcal{P}(\mathcal{X}_0^\circ) \times \mathcal{P}(\mathcal{X}_1^\circ)$ so that all relevant results still hold for the OT potentials) and observe that

$$\begin{aligned} \mathsf{OT}_{\mathbf{A}}(\tilde{\mu}_{0,t}, \tilde{\mu}_{1,t}) - \mathsf{OT}_{\mathbf{A}}(\bar{\mu}_0, \bar{\mu}_1) &\leq \int \varphi_{0,t}^{\mathbf{A}} d\tilde{\mu}_{0,t} + \int \varphi_{1,t}^{\mathbf{A}} d\tilde{\mu}_{1,t} - \int \varphi_{0,t}^{\mathbf{A}} d\bar{\mu}_0 - \int \varphi_{1,t}^{\mathbf{A}} d\bar{\mu}_1 \\ &= t \left(\int \bar{\varphi}_{0,t}^{\mathbf{A}} d(\nu_0 - \mu_0) + \int \bar{\varphi}_{1,t}^{\mathbf{A}} d(\nu_1 - \mu_1) \right), \end{aligned} \tag{3.40}$$

where, for a pair of functions (g_0, g_1), $(\bar{g}_0, \bar{g}_1) = (g_0(\cdot - \mathbb{E}_{\mu_0}[X]), g_1(\cdot - \mathbb{E}_{\mu_1}[X]))$.

Now, fix an arbitrary sequence $t_n \downarrow 0$ and a subsequence $t_{n'} \downarrow 0$. By the Arzèla-Ascoli theorem, there is a further subsequence $t_{n''} \downarrow 0$ along which $(\varphi_{0,t_{n''}}^{\mathbf{A}}, \varphi_{1,t_{n''}}^{\mathbf{A}})$ converges uniformly on $\mathcal{X}_0^\circ \times \mathcal{X}_1^\circ$ to a pair of continuous functions (φ_0, φ_1)[8] and, as $(\tilde{\mu}_{0,t}, \tilde{\mu}_{1,t})$ converge weakly to $(\bar{\mu}_0, \bar{\mu}_1)$, Proposition 5.20 in [208] and the dominated convergence theorem imply that

$$\int \bar{\varphi}_{0,t_{n''}}^{\mathbf{A}} d\mu_{0,t_{n''}} + \int \bar{\varphi}_{1,t_{n''}}^{\mathbf{A}} d\mu_{1,t_{n''}} \to \int \bar{\varphi}_0 d\mu_0 + \int \bar{\varphi}_1 d\mu_1$$

$$\mathsf{OT}_{\mathbf{A}}(\tilde{\mu}_{0,t_{n''}}, \tilde{\mu}_{1,t_{n''}}) \to \mathsf{OT}_{\mathbf{A}}(\bar{\mu}_0, \bar{\mu}_1).$$

As both terms on the left are equal, conclude that $\int \varphi_0 d\bar{\mu}_0 + \int \varphi_1 d\bar{\mu}_1 = \mathsf{OT}_{\mathbf{A}}(\bar{\mu}_0, \bar{\mu}_1)$. As the pointwise inequality $\varphi_{0,t_{n''}}^{\mathbf{A}} \oplus \varphi_{1,t_{n''}}^{\mathbf{A}} \leq c_{\mathbf{A}}$ is preserved in the limit, conclude that

[8] Indeed, these OT potentials extend uniquely to the closure of their domains whilst preserving the uniform equicontinuity and equiboundedness properties.

(φ_0, φ_1) is optimal for $\mathsf{OT}_{\mathbf{A}}(\bar{\mu}_0, \bar{\mu}_1)$ and that $(\varphi_0, \varphi_1) = (\varphi_0^{\mathbf{A}} + a, \varphi_1^{\mathbf{A}} - a)$ for some $a \in \mathbb{R}$ by Lemma 19 . It follows that

$$\lim_{t_{n''} \downarrow 0} \int \bar{\varphi}^{\mathbf{A}}_{0,t_{n''}} d(\nu_0 - \mu_0) + \int \bar{\varphi}^{\mathbf{A}}_{1,t_{n''}} d(\nu_1 - \mu_1) = \int \bar{\varphi}^{\mathbf{A}}_0 d(\nu_0 - \mu_0) + \int \bar{\varphi}^{\mathbf{A}}_1 d(\nu_1 - \mu_1),$$

as $\nu_0 - \mu_0$ and $\nu_1 - \mu_1$ have total mass zero. As this final limit is independent of the choice of original subsequence, (3.40) reads

$$\limsup_{t \downarrow 0} \frac{\mathsf{OT}_{\mathbf{A}}(\tilde{\mu}_{0,t}, \tilde{\mu}_{1,t}) - \mathsf{OT}_{\mathbf{A}}(\bar{\mu}_0, \bar{\mu}_1)}{t} \leq \int \bar{\varphi}^{\mathbf{A}}_0 d(\nu_0 - \mu_0) + \int \bar{\varphi}^{\mathbf{A}}_1 d(\nu_1 - \mu_1).$$

On the other hand,

$$\frac{\mathsf{OT}_{\mathbf{A}}(\tilde{\mu}_{0,t}, \tilde{\mu}_{1,t}) - \mathsf{OT}_{\mathbf{A}}(\bar{\mu}_0, \bar{\mu}_1)}{t} \geq \int \bar{\varphi}^{\mathbf{A}}_0 d(\nu_0 - \mu_0) + \int \bar{\varphi}^{\mathbf{A}}_1 d(\nu_1 - \mu_1), \tag{3.41}$$

which readily implies that

$$\lim_{t \downarrow 0} \frac{\mathsf{OT}_{\mathbf{A}}(\tilde{\mu}_{0,t}, \tilde{\mu}_{1,t}) - \mathsf{OT}_{\mathbf{A}}(\bar{\mu}_0, \bar{\mu}_1)}{t} = \int \bar{\varphi}^{\mathbf{A}}_0 d(\nu_0 - \mu_0) + \int \bar{\varphi}^{\mathbf{A}}_1 d(\nu_1 - \mu_1). \tag{3.42}$$

As for the other limit, note that if $\tilde{\pi} \in \Pi(\tilde{\mu}_{0,t}, \tilde{\mu}_{1,t})$, then $\tau^{-t}_{\sharp}\tilde{\pi} \in \Pi(\bar{\mu}_{0,t}, \bar{\mu}_{1,t})$, and if $\bar{\pi} \in \Pi(\bar{\mu}_{0,t}, \bar{\mu}_{1,t})$, then $\tau^{t}_{\sharp}\bar{\pi} \in \Pi(\tilde{\mu}_{0,t}, \tilde{\mu}_{1,t})$, where $\tau^t = (\tau^t_0, \tau^t_1) = (\mathrm{Id} + t(\mathbb{E}_{\nu_0}[X] - \mathbb{E}_{\mu_0}[X]), \mathrm{Id} + t(\mathbb{E}_{\nu_1}[X] - \mathbb{E}_{\mu_1}[X]))$ for $t \in [-1, 1]$. Hence, letting $\tilde{\pi}_t$ and $\bar{\pi}_t$ be OT plans for $\mathsf{OT}_{\mathbf{A}}(\tilde{\mu}_{0,t}, \tilde{\mu}_{1,t})$ and $\mathsf{OT}_{\mathbf{A}}(\bar{\mu}_{0,t}, \bar{\mu}_{1,t})$, we have

$$\begin{aligned} \mathsf{OT}_{\mathbf{A}}(\bar{\mu}_{0,t}, \bar{\mu}_{1,t}) - \mathsf{OT}_{\mathbf{A}}(\tilde{\mu}_{0,t}, \tilde{\mu}_{1,t}) &\leq \int c_{\mathbf{A}} \circ \tau^{-t} - c_{\mathbf{A}} d\tilde{\pi}_t, \\ \mathsf{OT}_{\mathbf{A}}(\bar{\mu}_{0,t}, \bar{\mu}_{1,t}) - \mathsf{OT}_{\mathbf{A}}(\tilde{\mu}_{0,t}, \tilde{\mu}_{1,t}) &\geq \int c_{\mathbf{A}} - c_{\mathbf{A}} \circ \tau^{t} d\bar{\pi}_t, \end{aligned} \tag{3.43}$$

by expanding the expressions for the cost functions, we obtain

$$\begin{aligned} c_{\mathbf{A}} \circ \tau^t(x, y) = &-4(\|x\|^2 + 2\langle x, t(\mathbb{E}_{\nu_0}[X] - \mathbb{E}_{\mu_0}[X])\rangle + t^2\|\mathbb{E}_{\nu_0}[X] - \mathbb{E}_{\mu_0}[X]\|^2) \\ &\times (\|y\|^2 + 2\langle y, t(\mathbb{E}_{\nu_1}[X] - \mathbb{E}_{\mu_1}[X])\rangle + t^2\|\mathbb{E}_{\nu_1}[X] - \mathbb{E}_{\mu_1}[X]\|^2) \\ &- 32(x + t(\mathbb{E}_{\nu_0}[X] - \mathbb{E}_{\mu_0}[X]))^\intercal \mathbf{A}(y + t(\mathbb{E}_{\nu_1}[X] - \mathbb{E}_{\mu_1}[X])) \\ &= c_{\mathbf{A}}(x, y) - 8t\langle x, \mathbb{E}_{\nu_0}[X] - \mathbb{E}_{\mu_0}[X]\rangle\|y\|^2 - 8t\langle y, \mathbb{E}_{\nu_1}[X] - \mathbb{E}_{\mu_1}[X]\rangle\|x\|^2 \\ &- 32t\left(\mathbb{E}_{\nu_0}[X] - \mathbb{E}_{\mu_0}[X]\right)^\intercal \mathbf{A}y - 32tx^\intercal\mathbf{A}\left(\mathbb{E}_{\nu_1}[X] - \mathbb{E}_{\mu_1}[X]\right) + O(t^2), \end{aligned} \tag{3.44}$$

such that

$$\limsup_{t\downarrow 0} \frac{\mathsf{OT}_{\mathbf{A}}(\bar{\mu}_{0,t},\bar{\mu}_{1,t}) - \mathsf{OT}_{\mathbf{A}}(\tilde{\mu}_{0,t},\tilde{\mu}_{1,t})}{t} \leq \int 8\langle x, \mathbb{E}_{\nu_0}[X] - \mathbb{E}_{\mu_0}[X]\rangle \|y\|^2 + 8\langle y, \mathbb{E}_{\nu_1}[X] - \mathbb{E}_{\mu_1}[X]\rangle \|x\|^2 d\pi^{\mathbf{A}}(x,y),$$

$$\liminf_{t\downarrow 0} \frac{\mathsf{OT}_{\mathbf{A}}(\bar{\mu}_{0,t},\bar{\mu}_{1,t}) - \mathsf{OT}_{\mathbf{A}}(\tilde{\mu}_{0,t},\tilde{\mu}_{1,t})}{t} \geq \int 8\langle x, \mathbb{E}_{\nu_0}[X] - \mathbb{E}_{\mu_0}[X]\rangle \|y\|^2 + 8\langle y, \mathbb{E}_{\nu_1}[X] - \mathbb{E}_{\mu_1}[X]\rangle \|x\|^2 d\pi^{\mathbf{A}}(x,y),$$

noting that $\tilde{\pi}_t$ and $\bar{\pi}_t$ both converge weakly to the unique (under Assumption 3) OT plan $\pi^{\mathbf{A}}$ for $\mathsf{OT}_{\mathbf{A}}(\bar{\mu}_0,\bar{\mu}_1)$ as follows from Theorem 5.20 in [208], Lemma 5.2.1 in [6], and the fact that $\bar{\mu}_0$ and $\bar{\mu}_1$ are mean-zero. This result, along with (3.42), proves the claim. □

Lemma 23. Let $\mathcal{A} = \operatorname{argmin}_{B_{\mathrm{F}}(M)} \Phi_{(\bar{\mu}_0,\bar{\mu}_1)}$. If $\bar{\mu}_0, \bar{\mu}_1$ satisfy Assumption 3 at every $\mathbf{A} \in \mathcal{A}$ and $\operatorname{int}(\operatorname{spt}(\bar{\mu}_0))$ is connected with negligible boundary,

$$\frac{d}{dt}\mathsf{S}_2(\bar{\mu}_{0,t},\bar{\mu}_{1,t})|_{t=0} = \inf_{\mathbf{A}\in\mathcal{A}} \left\{ \int g_{0,\pi^{\mathbf{A}}} + \bar{\varphi}_0^{\mathbf{A}} d(\nu_0 - \mu_0) + \int g_{1,\pi^{\mathbf{A}}} + \bar{\varphi}_1^{\mathbf{A}} d(\nu_1 - \mu_1) \right\},$$

where, for $(x,y) \in \mathbb{R}^{d_0} \times \mathbb{R}^{d_1}$,

$$g_{0,\pi}(x) = 8\mathbb{E}_{(X,Y)\sim\pi}[\|Y\|^2 X]^{\intercal} x,$$

$$g_{1,\pi}(y) = 8\mathbb{E}_{(X,Y)\sim\pi}[\|X\|^2 Y]^{\intercal} y,$$

$\bar{\varphi}_i^{\mathbf{A}} = \varphi_i^{\mathbf{A}}(\cdot - \mathbb{E}_{\mu_i}[X])$ for $i = 0, 1$, and, for $\mathbf{A} \in \mathcal{A}$, $\pi^{\mathbf{A}}$ is the unique OT plan for $\mathsf{OT}_{\mathbf{A}}(\bar{\mu}_0,\bar{\mu}_1)$, and $(\varphi_0^{\mathbf{A}}, \varphi_1^{\mathbf{A}})$ is any choice of corresponding OT potentials.

Proof. Let $\mathbf{A}^\star \in \mathcal{A}$ be arbitrary. From Lemma 22, we have that

$$\begin{aligned}
&\mathsf{S}_2(\bar{\mu}_{0,t},\bar{\mu}_{1,t}) - \mathsf{S}_2(\bar{\mu}_0,\bar{\mu}_1) \\
&\qquad \leq \mathsf{OT}_{\mathbf{A}^\star}(\bar{\mu}_{0,t},\bar{\mu}_{1,t}) - \mathsf{OT}_{\mathbf{A}^\star}(\bar{\mu}_0,\bar{\mu}_1) \\
&\qquad = t\left(\int g_{0,\pi^{\mathbf{A}^\star}} + \bar{\varphi}_0^{\mathbf{A}^\star} d(\nu_0 - \mu_0) + \int g_{1,\pi^{\mathbf{A}^\star}} + \bar{\varphi}_1^{\mathbf{A}^\star} d(\nu_1 - \mu_1) \right) + o(t),
\end{aligned}$$

such that

$$\limsup_{t\downarrow 0} \frac{\mathsf{S}_2(\bar{\mu}_{0,t},\bar{\mu}_{1,t}) - \mathsf{S}_2(\bar{\mu}_0,\bar{\mu}_1)}{t} \leq \inf_{\mathbf{A}\in\mathcal{A}} \left\{ \int g_{0,\pi^{\mathbf{A}}} + \bar{\varphi}_0^{\mathbf{A}} d(\nu_0 - \mu_0) + \int g_{1,\pi^{\mathbf{A}}} + \bar{\varphi}_1^{\mathbf{A}} d(\nu_1 - \mu_1) \right\}. \tag{3.45}$$

For the opposite limit, observe that the solutions of $\mathsf{S}_2(\bar\mu_{0,t},\bar\mu_{1,t})$ lie in $B_{\mathrm{F}}(M)$ for every $t\in[0,1]$ as follows from Theorem 8 and let $\mathbf{A}_t$ be optimal for each corresponding problem. By Lemma 19 and equations (3.40), (3.41), (3.43), and (3.44),

$$\left|\mathsf{OT}_{\mathbf{A}}(\bar\mu_{0,t},\bar\mu_{1,t})-\mathsf{OT}_{\mathbf{A}}(\bar\mu_0,\bar\mu_1)\right|=O(t),$$

where the $O(t)$ term can be chosen to be independent of $\mathbf{A}$. Whence, $\mathbf{A}\in B_{\mathrm{F}}(M)\mapsto \|\mathbf{A}\|_{\mathrm{F}}^2+\mathsf{OT}_{\mathbf{A}}(\bar\mu_{0,t},\bar\mu_{1,t})$ converges uniformly to $\mathbf{A}\in B_{\mathrm{F}}(M)\mapsto\|\mathbf{A}\|_{\mathrm{F}}^2+\mathsf{OT}_{\mathbf{A}}(\bar\mu_0,\bar\mu_1)$ as $t\downarrow 0$. Extending these functions by $+\infty$ outside of $B_{\mathrm{F}}(M)$, we may apply Proposition 7.15 and Theorem 7.33 from [168] to obtain that any cluster point of $(\mathbf{A}_t)_{t\downarrow 0}$ is an element of $\mathscr{A}$ keeping in mind that these functions are continuous on $B_{\mathrm{F}}(M)$ (see Theorem 8).

Now, let $t_n\downarrow 0$ be arbitrary and fix a subsequence $t_{n'}$. By the Bolzano-Weierstrass theorem, there exists a further subsequence $t_{n''}\downarrow 0$ along which $\mathbf{A}_{t_{n''}}\to\mathbf{A}\in\mathscr{A}$. From equations (3.41) and (3.43),

$$\begin{aligned}\frac{\mathsf{S}_2(\bar\mu_{0,t_{n''}},\bar\mu_{1,t_{n''}})-\mathsf{S}_2(\bar\mu_0,\bar\mu_1)}{t_{n''}}&\geq\frac{\mathsf{OT}_{\mathbf{A}_{t_{n''}}}(\bar\mu_{0,t_{n''}},\bar\mu_{1,t_{n''}})-\mathsf{OT}_{\mathbf{A}_{t_{n''}}}(\bar\mu_0,\bar\mu_1)}{t_{n''}}\\&\geq\int\bar\varphi_0^{\mathbf{A}_{t_{n''}}}d(\nu_0-\mu_0)+\int\bar\varphi_1^{\mathbf{A}_{t_{n''}}}d(\nu_1-\mu_1)\\&+\int c_{\mathbf{A}_{t_{n''}}}-c_{\mathbf{A}_{t_{n''}}}\circ\tau^{t_{n''}}d\bar\pi_{t_{n''}},\end{aligned}$$

where $(\varphi_0^{\mathbf{A}_{t_{n''}}},\varphi_1^{\mathbf{A}_{t_{n''}}})$ are OT potentials for $\mathsf{OT}_{\mathbf{A}_{t_{n''}}}(\bar\mu_0,\bar\mu_1)$ and $\bar\pi_{t_{n''}}$ is an OT plan for $\mathsf{OT}_{\mathbf{A}_{t_{n''}}}(\bar\mu_{0,t_{n''}},\bar\mu_{1,t_{n''}})$. It remains to show that $\bar\pi_{t_{n''}}$ converges weakly to $\pi^{\mathbf{A}}$ weakly and that $(\bar\varphi_0^{\mathbf{A}_{t_{n''}}},\bar\varphi_1^{\mathbf{A}_{t_{n''}}})$ converges to $(\bar\varphi_0^{\mathbf{A}}+a,\bar\varphi_1^{\mathbf{A}}-a)$ uniformly on $\mathcal{X}_0\times\mathcal{X}_1$ along a further subsequence and some choice of $a\in\mathbb{R}$ to obtain that

$$\liminf_{t\downarrow 0}\frac{\mathsf{S}_2(\bar\mu_{0,t},\bar\mu_{1,t})-\mathsf{S}_2(\bar\mu_0,\bar\mu_1)}{t}\geq\int g_{0,\pi^{\mathbf{A}}}+\bar\varphi_0^{\mathbf{A}}d(\nu_0-\mu_0)+\int g_{1,\pi^{\mathbf{A}}}+\bar\varphi_1^{\mathbf{A}}d(\nu_1-\mu_1),$$

recalling the expansion of the translated cost function from (3.44). The above display, together with (3.45), yields the desired result.

Note that $c_{\mathbf{A}_{t_{n''}}}$ converges uniformly to $c_{\mathbf{A}}$ on $\mathcal{X}_0\times\mathcal{X}_1$ and $(\bar\mu_{0,t_{n''}},\bar\mu_{1,t_{n''}})$ converges weakly to $(\bar{\mu}_0, \bar{\mu}_1)$ whereby $\pi^{\mathbf{A}_{t_{n'''}}}$ converges to $\pi^{\mathbf{A}}$ weakly along a further subsequence, $t_{n'''}$ (see Theorem 5.20 in [208]).

Applying the Arzelà-Ascoli theorem, we may also assume that $(\varphi_0^{\mathbf{A}_{t_{n'''}}}, \varphi_1^{\mathbf{A}_{t_{n'''}}})$ converges uniformly to some pair of continuous functions $(\bar{\varphi}_0, \bar{\varphi}_1)$ on $\mathcal{X}_0^\circ \times \mathcal{X}_1^\circ$ (see also Footnote 8). Applying the dominated convergence theorem and the weak convergence of the couplings,

$$\int \varphi_0^{\mathbf{A}_{t_{n'''}}} d\bar{\mu}_{0,t_{n'''}} + \int \varphi_1^{\mathbf{A}_{t_{n'''}}} d\bar{\mu}_{1,t_{n'''}} \to \int \varphi_0 d\bar{\mu}_0 + \int \varphi_1 d\bar{\mu}_1,$$

$$\mathsf{OT}_{\mathbf{A}_{t_{n'''}}}(\bar{\mu}_{0,t_{n'''}}, \bar{\mu}_{1,t_{n'''}}) \to \mathsf{OT}_{\mathbf{A}}(\bar{\mu}_0, \bar{\mu}_1).$$

As the two terms on the left hand side are equal, $\int \varphi_0 d\bar{\mu}_0 + \int \varphi_1 d\bar{\mu}_1 = \mathsf{OT}_{\mathbf{A}}(\bar{\mu}_0, \bar{\mu}_1)$ and, as the inequalities $\varphi_0^{\mathbf{A}_{t_{n'''}}} \oplus \varphi_1^{\mathbf{A}_{t_{n'''}}} \leq c_{\mathbf{A}_{t_{n'''}}}$ are preserved in the limit, (φ_0, φ_1) is a pair of OT potentials for $\mathsf{OT}_{\mathbf{A}}(\bar{\mu}_0, \bar{\mu}_1)$ and hence coincide with $(\varphi_0^{\mathbf{A}}, \varphi_1^{\mathbf{A}})$ up to constants, concluding the proof. □

Lemma 24. For any pairs $(\nu_0, \nu_1), (\rho_0, \rho_1) \in \mathcal{P}_{\mu_0} \times \mathcal{P}_{\mu_1}$, $|\mathsf{S}_2(\bar{\nu}_0, \bar{\nu}_1) - \mathsf{S}_2(\bar{\rho}_0, \bar{\rho}_1)| \leq C\|\nu_0 \otimes \nu_1 - \rho_0 \otimes \rho_1\|_{\infty, \mathcal{F}^\oplus}$ for some constant $C > 0$ which is independent of $(\nu_0, \nu_1), (\rho_0, \rho_1)$.

Proof. Observe that

$$|\mathsf{S}_2(\bar{\nu}_0, \bar{\nu}_1) - \mathsf{S}_2(\bar{\rho}_0, \bar{\rho}_1)| \leq \sup_{\mathbf{A} \in B_{\mathrm{F}}(M)} |\mathsf{OT}_{\mathbf{A}}(\bar{\nu}_0, \bar{\nu}_1) - \mathsf{OT}_{\mathbf{A}}(\bar{\rho}_0, \bar{\rho}_1)|\,. \tag{3.46}$$

Let $(\varphi_{0,\bar{\nu}}^{\mathbf{A}}, \varphi_{1,\bar{\nu}}^{\mathbf{A}})$ and $(\varphi_{0,\bar{\rho}}^{\mathbf{A}}, \varphi_{1,\bar{\rho}}^{\mathbf{A}})$ be extended OT potentials for $\mathsf{OT}_{\mathbf{A}}(\bar{\nu}_0, \bar{\nu}_1)$ and $\mathsf{OT}_{\mathbf{A}}(\bar{\rho}_0, \bar{\rho}_1)$ respectively satisfying the estimates from Lemma 19. We have that

$$\mathsf{OT}_{\mathbf{A}}(\bar{\nu}_0, \bar{\nu}_1) - \mathsf{OT}_{\mathbf{A}}(\bar{\rho}_0, \bar{\rho}_1) \leq \int \varphi_{0,\bar{\nu}}^{\mathbf{A}} d(\bar{\nu}_0 - \bar{\rho}_0) + \int \varphi_{1,\bar{\nu}}^{\mathbf{A}} d(\bar{\nu}_1 - \bar{\rho}_1),$$

$$\mathsf{OT}_{\mathbf{A}}(\bar{\nu}_0, \bar{\nu}_1) - \mathsf{OT}_{\mathbf{A}}(\bar{\rho}_0, \bar{\rho}_1) \geq \int \varphi_{0,\bar{\rho}}^{\mathbf{A}} d(\bar{\nu}_0 - \bar{\rho}_0) + \int \varphi_{1,\bar{\rho}}^{\mathbf{A}} d(\bar{\nu}_1 - \bar{\rho}_1).$$

Fix $i \in \{0, 1\}$ and observe that, for any L-Lipschitz function $f_i : \mathcal{X}_i^\circ \to \mathbb{R}$,

$$\begin{aligned}\int f_i d(\bar{\nu}_i - \bar{\rho}_i) &= \int f_i(\cdot - \mathbb{E}_{\nu_i}[X]) d\nu_i - \int f_i(\cdot - \mathbb{E}_{\rho_i}[X]) d\rho_i \\ &\leq \int f_i(\cdot - \mathbb{E}_{\rho_i}[X]) d(\nu_i - \rho_i) + L\|\mathbb{E}_{\nu_i}[X] - \mathbb{E}_{\rho_i}[X]\|, \\ \int f_i d(\bar{\nu}_i - \bar{\rho}_i) &\geq \int f_i(\cdot - \mathbb{E}_{\rho_i}[X]) d(\nu_i - \rho_i) - L\|\mathbb{E}_{\nu_i}[X] - \mathbb{E}_{\rho_i}[X]\|.\end{aligned}$$

As the coordinate projections are smooth with uniformly bounded derivatives of all orders, there exists a positive constant C_1 which is independent of $\nu_0, \nu_1, \rho_0, \rho_1$ such that $L\|\mathbb{E}_{\nu_i}[X] - \mathbb{E}_{\rho_i}[X]\| \leq L\sqrt{d_i} \max_{j=1}^{d_i} \left|\int x_j d(\nu_i - \rho_i)(x)\right| \leq C_1\|\nu_i - \rho_i\|_{\infty, \mathcal{F}_i} \leq C_1\|\nu_0 \otimes \nu_1 - \rho_0 \otimes \rho_1\|_{\infty, \mathcal{F}^\oplus}$, where the final inequality follows from the fact that both $\mathcal{F}_0$ and $\mathcal{F}_1$ contain the zero function. As the extended OT potentials are L-Lipschitz (see Lemma 19),

$$\begin{aligned}&|\mathsf{OT}_{\mathbf{A}}(\bar{\nu}_0, \bar{\nu}_1) - \mathsf{OT}_{\mathbf{A}}(\bar{\rho}_0, \bar{\rho}_1)| \\ &\quad\leq \Bigg(\left|\int \varphi^{\mathbf{A}}_{0,\bar{\nu}}(\cdot - \mathbb{E}_{\rho_0}[X]) d(\nu_0 - \rho_0) + \int \varphi^{\mathbf{A}}_{1,\bar{\nu}}(\cdot - \mathbb{E}_{\rho_1}[X]) d(\nu_1 - \rho_1)\right| \\ &\qquad \bigvee \left|\int \varphi^{\mathbf{A}}_{0,\bar{\rho}}(\cdot - \mathbb{E}_{\rho_0}[X]) d(\nu_0 - \rho_0) + \int \varphi^{\mathbf{A}}_{1,\bar{\rho}}(\cdot - \mathbb{E}_{\rho_1}[X]) d(\nu_1 - \rho_1)\right|\Bigg) \\ &\quad + 2C_1\|\nu_0 \otimes \nu_1 - \rho_0 \otimes \rho_1\|_{\infty, \mathcal{F}^\oplus}.\end{aligned}$$

Observe that $\varphi^{\mathbf{A}}_{0,\bar{\nu}}(\cdot - \mathbb{E}_{\rho_0}[X]), \varphi^{\mathbf{A}}_{0,\bar{\rho}}(\cdot - \mathbb{E}_{\rho_0}[X]) \in \mathcal{F}_0$ by construction and that $\varphi^{\mathbf{A}}_{1,\bar{\rho}}(\cdot - \mathbb{E}_{\rho_1}[X]), \varphi^{\mathbf{A}}_{1,\bar{\nu}}(\cdot - \mathbb{E}_{\rho_1}[X])$ are bounded on $\mathcal{X}_1^\circ$ and hence are elements of $\mathcal{F}_1$ up to some normalizing constant. It follows that the the term in brackets on the right hand side of the display can be bounded as $C_2\|\nu_0 \otimes \nu_1 - \rho_0 \otimes \rho_1\|_{\infty, \mathcal{F}^\oplus}$ for some positive constant C_2 which is independent of $\mathbf{A}, \nu_0, \nu_1, \rho_0, \rho_1$. Applying these bounds in (3.46) proves the desired Lipschitz continuity of S_2. ◻

Proof of Theorem 13. By applying the variational form (3.8), the desired Gâteaux differentiability follows from Lemmas 20 and 23 whereas the Lipschitz continuity follows from Lemmas 21 and 24. ◻

3.8.4 Proof of Theorem 14

Given Theorem 13, the first assertion of Theorem 14 follows from the unified approach (Proposition 16).

Specializing to the empirical measures, the result will follow upon showing that the function classes $\mathcal{F}_0$ and $\mathcal{F}_1$ are, respectively, μ_0- and μ_1-Donsker. Indeed, by applying the same argument as in the end of Theorem 10, Donskerness of each class and the independence assumption on the samples imply that $\sqrt{n}(\hat{\mu}_{0,n} - \mu_0) \otimes (\hat{\mu}_{1,n} - \mu_1) \xrightarrow{d} G_{\mu_0 \otimes \mu_1}$ in $\ell^\infty(\mathcal{F}^\oplus)$ where $G_{\mu_0 \otimes \mu_1}(f_0 \oplus f_1) = G_{\mu_0}(f_0) + G_{\mu_1}(f_1)$ and G_{μ_0}, G_{μ_1} are, respectively, tight μ_0- and μ_1-Brownian bridge processes in $\ell^\infty(\mathcal{F}_0)$ and $\ell^\infty(\mathcal{F}_1)$. We begin with $\mathcal{F}_0$.

By Corollary 2.7.2 and the surrounding discussion in [201], the unit ball in $C^k(\mathcal{X}_0)$ is universally Donsker provided that $k = \lfloor d/2 \rfloor + 1$. As unions and pairwise sums of Donsker classes are Donsker (see Example 2.10.7 in [201]), it suffices to show that $\mathcal{G}$ is μ_0-Donsker.

Let $x, \xi, \xi' \in \mathcal{X}_0$, $\zeta, \zeta' \in \mathbb{R}^{d_1}$ with $\|\zeta\|, \|\zeta'\| \le \|\mathcal{X}_1\|_\infty$, $z, z' \in \mathbb{R}^N$ with $\|z\|_\infty, \|z'\|_\infty \le K$, and $\mathbf{A}, \mathbf{A}' \in B_{\mathrm{F}}(M)$ be arbitrary. Observe that

$$
\begin{aligned}
&\left| \min_{1\le i\le N} \left\{ c_{\mathbf{A}}\left(x-\xi, y^{(i)}-\zeta\right) - z_i \right\} - \min_{1\le i\le N} \left\{ c_{\mathbf{A}'}\left(x-\xi', y^{(i)}-\zeta'\right) - z_i' \right\} \right| \\
&\le \max_{1\le i\le N} \Big| -4\|x-\xi\|^2\|y^{(i)}-\zeta\|^2 - 32(x-\xi)^\intercal \mathbf{A}(y^{(i)}-\zeta) - z_i \\
&\qquad\qquad + 4\|x-\xi'\|^2\|y^{(i)}-\zeta'\|^2 + 32(x-\xi')^\intercal \mathbf{A}'(y^{(i)}-\zeta') + z_i' \Big| .
\end{aligned}
$$

By applying the reverse triangle inequality, we see that

$$
\begin{aligned}
&\left| \|x-\xi\|^2\|y^{(i)}-\zeta\|^2 - \|x-\xi'\|^2\|y^{(i)}-\zeta'\|^2 \right| \\
&\le \left| \|x-\xi\|^2 - \|x-\xi'\|^2 \right| \|y^{(i)}-\zeta\|^2 + \left| \|y^{(i)}-\zeta\|^2 - \|y^{(i)}-\zeta'\|^2 \right| \|x-\xi'\|^2 \\
&\le (\|x-\xi\| + \|x-\xi'\|)\|\xi-\xi'\|\|y^{(i)}-\zeta\|^2 + (\|y^{(i)}-\zeta\| + \|y^{(i)}-\zeta'\|)\|\zeta-\zeta'\|\|x-\xi'\|^2 \\
&\lesssim_{\mathcal{X}_0,\mathcal{X}_1} \|\xi-\xi'\| + \|\zeta-\zeta'\|,
\end{aligned}
$$

whereas

$$
\begin{aligned}
&\left|(x-\xi)^\intercal \mathbf{A}(y^{(i)}-\zeta) - (x-\xi')^\intercal \mathbf{A}'(y^{(i)}-\zeta')\right| \\
&\leq \left|(x-\xi)^\intercal (\mathbf{A}-\mathbf{A}')(y^{(i)}-\zeta)\right| + \left|(\xi'-\xi)^\intercal \mathbf{A}'(y^{(i)}-\zeta)\right| + |(x-\xi')^\intercal \mathbf{A}'(\zeta'-\zeta)| \\
&\lesssim_{M,\mathcal{X}_0,\mathcal{X}_1} \|\mathbf{A}-\mathbf{A}'\|_{\mathrm{F}} + \|\xi'-\xi\| + \|\zeta'-\zeta\|\,,
\end{aligned}
$$

so that there exists a constant $C_{M,\mathcal{X}_0,\mathcal{X}_1} < \infty$ satisfying

$$
\begin{aligned}
&\left|\min_{1\leq i\leq N}\left\{c_{\mathbf{A}}\left(x-\xi, y^{(i)}-\zeta\right) - z_i\right\} - \min_{1\leq i\leq N}\left\{c_{\mathbf{A}'}\left(x-\xi', y^{(i)}-\zeta'\right) - z_i'\right\}\right| \\
&\leq C_{M,\mathcal{X}_0,\mathcal{X}_1}\left(\|\mathbf{A}-\mathbf{A}'\|_{\mathrm{F}} + \|\xi'-\xi\| + \|\zeta'-\zeta\| + \|z-z'\|_\infty\right)
\end{aligned}
$$

whereby the function class $\mathcal{G}$ is Lipschitz with respect to the parameter $(\mathbf{A},\xi,\zeta,z) \in B_{\mathrm{F}}(M)\times\mathcal{X}_0\times B_{(\mathbb{R}^{d_1},\|\cdot\|)}(\|\mathcal{X}_1\|_\infty)\times B_{(\mathbb{R}^N,\|\cdot\|_\infty)}(K) =: \mathcal{Q}$, where $B_{(\mathbb{R}^{d_1},\|\cdot\|)}(\|\mathcal{X}_1\|_\infty)$ is the closed ball of radius $\|\mathcal{X}_1\|_\infty$ in $\mathbb{R}^{d_1}$ with respect to $\|\cdot\|$, and $B_{(\mathbb{R}^N,\|\cdot\|_\infty)}(K)$ is defined analogously. Let $\tau : (\mathbf{A},\xi,\zeta,z) \in \mathcal{Q} \mapsto (A_{(\cdot),1}, A_{(\cdot),2},\dots,A_{(\cdot),d_1},\xi,\zeta,z) \in \mathbb{R}^{d_0d_1+d_0+d_1+N}$, where $A_{(\cdot),i}$ is the i-th column of $\mathbf{A}$ for $i\in[d_1]$; we endow this latter Euclidean space with the norm $\|(u,v,w,z)\|_{\tau\mathcal{Q}} = \|u\|+\|v\|+\|w\|+\|z\|_\infty$ where $u\in\mathbb{R}^{d_0d_1}$, $v\in\mathbb{R}^{d_0}$, $w\in\mathbb{R}^{d_1}$, and $z\in\mathbb{R}^N$.

By Theorem 2.7.11 in [201], $N_{[\,]}(2\epsilon C_{M,\mathcal{X}_0,\mathcal{X}_1},\mathcal{G},L^2(\mu_0)) \leq N(\epsilon,\tau\mathcal{Q},\|\cdot\|_{\tau\mathcal{Q}})$. It suffices, therefore, to bound $N(\epsilon,\tau\mathcal{Q},\|\cdot\|_{\tau\mathcal{Q}})$. To this end, observe that $\tau\mathcal{Q}\subset B_{\|\cdot\|_{\tau\mathcal{Q}}}(M+\|\mathcal{X}_0\|_\infty+\|\mathcal{X}_1\|_\infty+K)$, the ball of radius $M+\|\mathcal{X}_0\|_\infty+\|\mathcal{X}_1\|_\infty+K$ in $\mathbb{R}^{d_0d_1+d_0+d_1+N}$ with respect to $\|\cdot\|_{\tau\mathcal{Q}}$. Let S_ϵ be an ϵ-net for this ball with respect to $\|\cdot\|_{\tau\mathcal{Q}}$ and note that $\cup_{x\in S_\varepsilon}(x+\frac{\epsilon}{2}B_{\tau\mathcal{Q}}(1)) \subset (M+\|\mathcal{X}_0\|_\infty+\|\mathcal{X}_1\|_\infty+K+\frac{\epsilon}{2})B_{\tau\mathcal{Q}}(1)$. As the sets on the left hand side of the previous expression are disjoint,

$$
\begin{aligned}
&|S_\epsilon|\left(\frac{\epsilon}{2}\right)^{d_0d_1+d_0+d_1+N}\mathrm{vol}\left(B_{\tau\mathcal{Q}}(1)\right) \\
&\qquad\leq \left(M+\|\mathcal{X}_0\|_\infty+\|\mathcal{X}_1\|_\infty+K+\frac{\epsilon}{2}\right)^{d_0d_1+d_0+d_1+N}\mathrm{vol}\left(B_{\tau\mathcal{Q}}(1)\right)
\end{aligned}
$$

whereby $|S_\epsilon| \leq \left(\frac{2(M+\|\mathcal{X}_0\|_\infty+\|\mathcal{X}_1\|_\infty+K)}{\epsilon}+1\right)^{d_0d_1+d_0+d_1+N}$.

In sum, $N(\epsilon,\tau\mathcal{Q},\|\cdot\|_{\tau\mathcal{Q}}) \leq \left(\frac{2(M+\|\mathcal{X}_0\|_\infty+\|\mathcal{X}_1\|_\infty+K)}{\epsilon}\bigvee 2+1\right)^{d_0d_1+d_0+d_1+N}$ so that $\mathcal{G}$ is μ_0-Donsker provided that it admits a cover which is square integrable with respect to μ_0 (see

Theorem 3.7.38 in [96]). This last condition is evidently satisfied, as $c_{\mathbf{A}}(x-\xi, y^{(i)}-\zeta)-z_i$ can be bounded by a constant independently of the choice of $(\mathbf{A}, \xi, \zeta, z) \in Q$, $x \in \mathcal{X}_0$, and $i \in [N]$.

As for $\mathcal{F}_1$, note that if $f, g \in \mathcal{F}_1$, then

$$\|f-g\|_{L^2(\mu_1)}^2 = \sum_{i=1}^N (f(y^{(i)}) - g(y^{(i)}))^2 \mu_1(\{y^{(i)}\}) \leq \|z_f - z_g\|_\infty^2$$

where $z_f = (f(y^{(1)}), \ldots, f(y^{(N)}))$ and z_g is defined analogously. Thus, $N(\epsilon, \mathcal{F}_1, L^2(\mu_1)) \leq N(\epsilon, \{z \in \mathbb{R}^N : \|z\|_\infty \leq K\}, \|\cdot\|_\infty) \leq \left(\frac{2K}{\epsilon} \bigvee 2 + 1\right)^N$, where the second inequality follows from the first part of the proof. Conclude from Theorem 2.5.2 in [201] that $\mathcal{F}_1$ is μ_1-Donsker. $\square$

3.8.5 Proofs for Section 3.5

Proof of Theorem 9

We first show that the centered measures are 4-sub-Weibull with a particular constant and establish a bound on the second moment of a 4-sub-Weibull distribution.

Lemma 25. Fix $\nu \in \mathcal{P}_{4,\sigma}(\mathbb{R}^d)$. Then, $\bar{\nu} \in \mathcal{P}_{4,\bar{\sigma}}(\mathbb{R}^d)$ for $\bar{\sigma}^2 = \frac{2+\log 2}{\log 2} 8\sigma^2$ and it holds that $M_2(\nu), M_2(\bar{\nu}) \leq 2\sigma$.

Proof. As $\nu \in \mathcal{P}_{4,\sigma}(\mathbb{R}^d)$,

$$\int e^{\frac{\|\cdot\|^4}{2\bar{\sigma}^2}} d\bar{\nu} = \int e^{\frac{\|\cdot - \mathbb{E}_\nu[X]\|^4}{2\bar{\sigma}^2}} d\nu \leq e^{\frac{8\|\mathbb{E}_\nu[X]\|^4}{2\bar{\sigma}^2}} \int e^{\frac{8\|\cdot\|^4}{2\bar{\sigma}^2}} d\nu \leq e^{\frac{8\|\mathbb{E}_\nu[X]\|^4}{2\bar{\sigma}^2}} \left(\int e^{\frac{\|\cdot\|^4}{2\sigma^2}} d\nu\right)^{\frac{8\sigma^2}{\bar{\sigma}^2}}$$
$$\leq e^{\frac{32\sigma^2}{2\bar{\sigma}^2}} 2^{\frac{8\sigma^2}{\bar{\sigma}^2}} = 2,$$

where the first inequality follows from the triangle inequality and the inequality $(|a| + |b|)^p \leq 2^{p-1}(|a|^p + |b|^p)$ for $a, b \in \mathbb{R}$ and $p \geq 1$, the second is due to Jensen's inequality, noting that $8\sigma^2 \leq \bar{\sigma}^2$, and the final inequality follows from the 4-sub-Weibull assumption and the fact that

$$2 \geq \int e^{\frac{\|\cdot\|^4}{2\sigma^2}} d\nu \geq \int \frac{\|\cdot\|^4}{2\sigma^2} d\nu = M_4(\nu)/(2\sigma^2)$$

so that $\|\mathbb{E}_\nu[X]\|^4 \leq \mathbb{E}_\nu[\|X\|^4] \leq 4\sigma^2$ by Jensen's inequality. This proves the first claim. For the second claim, we have by Jensen's inequality that $M_2^2(\nu) \leq M_4(\nu) \leq 4\sigma^2$ as above. Conclude by noting that $M_2(\bar{\nu}) = M_2(\nu) - \|\mathbb{E}_\nu[X]\|^2 \leq M_2(\nu)$. $\square$

Consequently, we set $M = \sigma$ and analyze the regularity properties of EOT potentials for $\mathsf{OT}_{\mathbf{A},\varepsilon}(\nu_0, \nu_1)$ for 4-sub-Weibull distributions with $\mathbf{A} \in B_{\mathrm{F}}(M)$.

Lemma 26 (Regularity of EOT potentials [223])**.** Fix $\sigma > 0$. Then, for any choice of $(\nu_0, \nu_1) \in \mathcal{P}_{4,\sigma}(\mathbb{R}^{d_0}) \times \mathcal{P}_{4,\sigma}(\mathbb{R}^{d_1})$ and $\mathbf{A} \in B_{\mathrm{F}}(M)$, there exists a pair of EOT potentials $(\varphi_0^{\mathbf{A},\varepsilon}, \varphi_1^{\mathbf{A},\varepsilon})$ for $\mathsf{OT}_{\mathbf{A},\varepsilon}(\nu_0, \nu_1)$ satisfying the Schrödinger system (3.6) on $\mathbb{R}^{d_0} \times \mathbb{R}^{d_1}$. Furthermore, this pair is unique up to additive constants on $\mathbb{R}^{d_0} \times \mathbb{R}^{d_1}$ and, for every $k \in \mathbb{N}$, there exists a constant K depending only on k, d_0, d_1, σ, and ε for which

$$\varphi_i^{\mathbf{A},\varepsilon} \leq K(1 + \|\cdot\|^2), \quad -\varphi_i^{\mathbf{A},\varepsilon} \leq K(1 + \|\cdot\|^4), \quad |D^\alpha \varphi_i^{\mathbf{A},\varepsilon}| \leq K(1 + \|\cdot\|^{3k}),$$

for every nonzero multi-index $\alpha \in \mathbb{N}_0^{d_i}$ with $|\alpha| \leq k$ and $i \in \{0, 1\}$.

Proof. Lemma 4 in [223] asserts that there exists a pair of EOT potentials $(\varphi_0^{\mathbf{A},1}, \varphi_1^{\mathbf{A},1})$ for $\mathsf{OT}_{\mathbf{A},1}(\nu_0, \nu_1)$ satisfying the Schrödinger system on $\mathbb{R}^{d_0} \times \mathbb{R}^{d_1}$ and, for any fixed $k \in \mathbb{N}$, there exists a constant K' depending only on k, d_0, d_1, σ for which

$$\varphi_i^{\mathbf{A},1} \leq K'(1 + \|\cdot\|^2), \quad -\varphi_i^{\mathbf{A},1} \leq K'(1 + \|\cdot\|^4), \quad |D^\alpha \varphi_i^{\mathbf{A},1}| \leq K'(1 + \|\cdot\|^{3k}) \qquad (3.47)$$

for every nonzero multi-index $\alpha \in \mathbb{N}_0^{d_i}$ with $|\alpha| \leq k$ and $i \in \{0, 1\}$. To extend this result to the general case $\varepsilon > 0$, observe that if $(\eta_0, \eta_1) \in \mathcal{P}(\mathbb{R}^{d_0}) \times \mathcal{P}(\mathbb{R}^{d_1})$ have finite fourth moments and $\pi \in \Pi(\eta_0, \eta_1)$ with $\pi \ll \eta_0 \otimes \eta_1$, then

$$\int c_{\mathbf{A}} d\pi + \varepsilon \mathsf{D}_{\mathrm{KL}}(\pi \| \eta_0 \otimes \eta_1) = \varepsilon \left(\int c_{\mathbf{A}} d\pi^\varepsilon + \mathsf{D}_{\mathrm{KL}}(\pi^\varepsilon \| \eta_0^\varepsilon \otimes \eta_1^\varepsilon) \right),$$

where $\pi^\varepsilon = (\varepsilon^{-1/4}\,\mathrm{Id}, \varepsilon^{-1/4}\,\mathrm{Id})_\sharp\pi$ and similarly for η_0^ε and η_1^ε. The above display implies that $\mathsf{OT}_{\mathbf{A},\varepsilon}(\eta_0,\eta_1) = \varepsilon\mathsf{OT}_{\mathbf{A},1}(\eta_0^\varepsilon,\eta_1^\varepsilon)$, that if $\pi_\star$ is optimal for $\mathsf{OT}_{\mathbf{A},\varepsilon}(\eta_0,\eta_1)$, then $\pi_\star^\varepsilon$ is optimal for $\mathsf{OT}_{\mathbf{A},1}(\eta_0^\varepsilon,\eta_1^\varepsilon)$, and $\frac{d\pi_\star}{d\eta_0\otimes\eta_1} = e^{\frac{\varphi_0\oplus\varphi_1-c_{\mathbf{A}}}{\varepsilon}}$ where (φ_0,φ_1) are the associated EOT potentials whereas $\frac{d\pi_\star^\varepsilon}{d\eta_0^\varepsilon\otimes\eta_1^\varepsilon} = \frac{d\pi_\star}{d\eta_0\otimes\eta_1}\circ(\varepsilon^{1/4}\,\mathrm{Id}) = e^{\varepsilon^{-1}(\varphi_0(\varepsilon^{1/4}\cdot)\oplus\varphi_1(\varepsilon^{1/4}\cdot))-c_{\mathbf{A}}}$ so that $\varepsilon^{-1}(\varphi_0(\varepsilon^{1/4}\cdot),\varphi_1(\varepsilon^{1/4}\cdot))$ is optimal for $\mathsf{OT}_{\mathbf{A},1}(\eta_0^\varepsilon,\eta_1^\varepsilon)$. We also point out that if η_0,η_1 are 4-sub-Weibull with parameter σ^2, then $\eta_0^\varepsilon,\eta_1^\varepsilon$ are 4-sub-Weibull with parameter σ^2/ε. The claimed estimates then follow readily from (3.47).

Finally, assume that (φ_0,φ_1) is another pair of EOT potentials for $\mathsf{OT}_{\mathbf{A},\varepsilon}(\nu_0,\nu_1)$, then (φ_0,φ_1) coincides with $(\varphi_0^{\mathbf{A},\varepsilon},\varphi_1^{\mathbf{A},\varepsilon})$ up to additive constants $\nu_0\otimes\nu_1$-a.s. Thus, there exists some $a\in\mathbb{R}$ for which

$$e^{-\frac{\varphi_0(x)}{\varepsilon}} = \int e^{\frac{\varphi_1-c_{\mathbf{A}}(x,\cdot)}{\varepsilon}}d\nu_1 = \int e^{\frac{\varphi_1^{\mathbf{A},\varepsilon}-a-c_{\mathbf{A}}(x,\cdot)}{\varepsilon}}d\nu_1 = e^{-\frac{\varphi_0^{\mathbf{A},\varepsilon}(x)+a}{\varepsilon}},$$

for every $x\in\mathbb{R}^{d_0}$ so that φ_0 coincides with $\varphi_0^{\mathbf{A},\varepsilon}+a$ on $\mathbb{R}^{d_0}$. The same argument applies to φ_1. □

To prove Theorem 15, we adopt a similar strategy to the semi-discrete case (Theorem 13). Namely, for $i\in\{0,1\}$, we set $\mu_{i,t} = \mu_i + t(\nu_i-\mu_i)$ for $t\in[0,1]$ and some pair $(\nu_0,\nu_1)\in\mathcal{P}_{4,\sigma}(\mathbb{R}^{d_0})\times\mathcal{P}_{4,\sigma}(\mathbb{R}^{d_1})$ and utilize the variational form $\mathsf{D}_\varepsilon(\bar\mu_0,\bar\mu_1)^2 = \mathsf{S}_1(\bar\mu_1,\bar\mu_1)+\mathsf{S}_{2,\varepsilon}(\bar\mu_0,\bar\mu_1)$ from (3.8). We make precise that, since μ_0,μ_1,ν_0,ν_1 are all 4-sub-Weibull with the same parameter σ, so too are $\mu_{0,t}$ and $\mu_{1,t}$ with the same parameter.

Lemma 27. As $t\downarrow 0$, we have that

$$\begin{aligned}\frac{\mathsf{S}_1(\bar\mu_{0,t},\bar\mu_{1,t})-\mathsf{S}_1(\bar\mu_0,\bar\mu_1)}{t} \to\ & 2\iint \|x-x'\|^4 d\mu_0(x)d(\nu_0-\mu_0)(x') \\ &+2\iint \|y-y'\|^4 d\mu_1(y)d(\nu_1-\mu_1)(y') \\ &-4\iint \|x-\mathbb{E}_{\mu_0}[X]\|^2\|y-\mathbb{E}_{\mu_1}[X]\|^2 d(\nu_0-\mu_0)(x)d\mu_1(y) \\ &-4\iint \|x-\mathbb{E}_{\mu_0}[X]\|^2\|y-\mathbb{E}_{\mu_1}[X]\|^2 d\mu_0(x)d(\nu_1-\mu_1)(y).\end{aligned}$$

The proof of Lemma 27 follows the same lines as the proof of Lemma 20. The only distinction is that boundedness of all relevant integrals is justified using the 4-sub-Weibull assumption rather than compact support. The proof is thus ommitted for brevity.

Lemma 28. For any pairs $(\nu_0, \nu_1), (\rho_0, \rho_1) \in \mathcal{P}_{\mu_0} \times \mathcal{P}_{\mu_1}$, $|\mathsf{S}_1(\bar{\nu}_0, \bar{\nu}_1) - \mathsf{S}_1(\bar{\rho}_0, \bar{\rho}_1)| \leq C\|\nu_0 \otimes \nu_1 - \rho_0 \otimes \rho_1\|_{\infty, \mathcal{F}^{\oplus}}$ for some constant $C > 0$ which is independent of $(\nu_0, \nu_1), (\rho_0, \rho_1)$.

Proof. The same arguments from the proof of Lemma 21 apply in this setting with the distinction that the relevant integrals can be bounded (uniformly in the choice of $\nu_0, \nu_1, \rho_0, \rho_1$) by appealing to the 4-sub-Weibull assumption.

For completeness, we underscore that the functions

$$\begin{aligned}
\xi_0 &: x \in \mathbb{R}^{d_0} \mapsto \int \|x - x'\|^4 d\nu_0(x') + \int \|x - x'\|^4 d\rho_0(x'), \\
\xi_1 &: x \in \mathbb{R}^{d_1} \mapsto \int \|x - x'\|^4 d\nu_1(x') + \int \|x - x'\|^4 d\rho_1(x'), \\
\zeta_0 &: x \in \mathbb{R}^{d_0} \mapsto \|x - \mathbb{E}_{\mu_0}[X]\|^2, \\
\zeta_1 &: x \in \mathbb{R}^{d_1} \mapsto \|x - \mathbb{E}_{\nu_1}[X]\|^2,
\end{aligned}$$

are smooth and have derivatives (of all orders) which can be bounded as $1 + \|\cdot\|^4$ up to a multiplicative constant which is independent of $\nu_0, \nu_1, \rho_0, \rho_1$. Conclude that $\xi_0, \zeta_0 \in \mathcal{F}_0$ and $\xi_1, \zeta_1 \in \mathcal{F}_1$ so that the remainder of the proof follows analogously to the proof of Lemma 21. $\square$

Lemma 29. If $(\mu_0, \mu_1) \in \mathcal{P}_{4,\sigma}(\mathbb{R}^{d_0}) \times \mathcal{P}_{4,\sigma}(\mathbb{R}^{d_1})$ for some $\sigma > 0$, then, for any $\mathbf{A} \in B_{\mathrm{F}}(M)$,

$$\begin{aligned}
\lim_{t \downarrow 0} \frac{\mathsf{OT}_{\mathbf{A},\varepsilon}(\bar{\mu}_{0,t}, \bar{\mu}_{1,t}) - \mathsf{OT}_{\mathbf{A},\varepsilon}(\bar{\mu}_0, \bar{\mu}_1)}{t} &= \int \bar{\varphi}_0^{\mathbf{A},\varepsilon} + 8\mathbb{E}_{(X,Y)\sim\pi^{\mathbf{A},\varepsilon}}[\|Y\|^2 X]^{\intercal}(\cdot) d(\nu_0 - \mu_0) \\
&\quad + \int \bar{\varphi}_1^{\mathbf{A},\varepsilon} + 8\mathbb{E}_{(X,Y)\sim\pi^{\mathbf{A},\varepsilon}}[\|X\|^2 Y]^{\intercal}(\cdot) d(\nu_1 - \mu_1).
\end{aligned}$$

Proof. Fix $\mathbf{A} \in B_{\mathrm{F}}(M)$. Proceeding as in the semi-discrete case (Lemma 22), consider

$$\begin{aligned}
\lim_{t \downarrow 0} \frac{\mathsf{OT}_{\mathbf{A},\varepsilon}(\bar{\mu}_{0,t}, \bar{\mu}_{1,t}) - \mathsf{OT}_{\mathbf{A},\varepsilon}(\bar{\mu}_0, \bar{\mu}_1)}{t} = \lim_{t \downarrow 0} \Bigg(&\frac{\mathsf{OT}_{\mathbf{A},\varepsilon}(\bar{\mu}_{0,t}, \bar{\mu}_{1,t}) - \mathsf{OT}_{\mathbf{A},\varepsilon}(\tilde{\mu}_{0,t}, \tilde{\mu}_{1,t})}{t} \\
&+ \frac{\mathsf{OT}_{\mathbf{A},\varepsilon}(\tilde{\mu}_{0,t}, \tilde{\mu}_{1,t}) - \mathsf{OT}_{\mathbf{A},\varepsilon}(\bar{\mu}_0, \bar{\mu}_1)}{t} \Bigg)
\end{aligned}$$

where $(\tilde\mu_{0,t},\tilde\mu_{1,t}) = \Big((\mathrm{Id}-\mathbb{E}_{\mu_0}[X])_\sharp\mu_{0,t},(\mathrm{Id}-\mathbb{E}_{\mu_1}[X])_\sharp\mu_{1,t}\Big)$. We compute each limit on the right hand side separately.

First, let $(\varphi_0^{(\nu_0,\nu_1)},\varphi_1^{(\nu_0,\nu_1)})$ be some choice of EOT potentials for $\mathsf{OT}_{\mathbf{A},\varepsilon}(\nu_0,\nu_1)$ from Lemma 26 for any choice of 4-sub-Weibull distributions (ν_0,ν_1) and observe that

$$\begin{aligned}
&\mathsf{OT}_{\mathbf{A},\varepsilon}(\tilde\mu_{0,t},\tilde\mu_{1,t})-\mathsf{OT}_{\mathbf{A},\varepsilon}(\tilde\mu_{0,t},\bar\mu_1)\\
&\qquad\le \int\varphi_0^{(\tilde\mu_{0,t},\tilde\mu_{1,t})}d\tilde\mu_{0,t}+\int\varphi_1^{(\tilde\mu_{0,t},\tilde\mu_{1,t})}d\tilde\mu_{1,t}-\int\varphi_0^{(\tilde\mu_{0,t},\tilde\mu_{1,t})}d\tilde\mu_{0,t}\\
&\qquad-\int\varphi_1^{(\tilde\mu_{0,t},\tilde\mu_{1,t})}d\bar\mu_1+\varepsilon\int e^{\frac{\varphi_0^{(\tilde\mu_{0,t},\tilde\mu_{1,t})}\oplus\varphi_1^{(\tilde\mu_{0,t},\tilde\mu_{1,t})}-c_{\mathbf{A}}}{\varepsilon}}d\tilde\mu_{0,t}\otimes\bar\mu_1-\varepsilon\\
&\qquad= t\int\bar\varphi_1^{(\tilde\mu_{0,t},\tilde\mu_{1,t})}d(\nu_1-\mu_1)
\end{aligned}\tag{3.48}$$

where we have used the fact that $\int e^{\frac{\varphi_0^{(\tilde\mu_{0,t},\tilde\mu_{1,t})}\oplus\varphi_1^{(\tilde\mu_{0,t},\tilde\mu_{1,t})}-c_{\mathbf{A}}}{\varepsilon}}d\tilde\mu_{0,t}\equiv 1$ on $\mathbb{R}^{d_1}$ (as the EOT potentials solve the relevant Shrödinger systems) and we recall the notation $(\bar g_0,\bar g_1)=(g_0(\cdot-\mathbb{E}_{\mu_0}[X]),g_1(\cdot-\mathbb{E}_{\mu_1}[X]))$. Similarly,

$$\begin{aligned}
\mathsf{OT}_{\mathbf{A},\varepsilon}(\tilde\mu_{0,t},\tilde\mu_{1,t})-\mathsf{OT}_{\mathbf{A},\varepsilon}(\tilde\mu_{0,t},\bar\mu_1)&\ge t\int\bar\varphi_1^{(\tilde\mu_{0,t},\bar\mu_1)}d(\nu_1-\mu_1),\\
\mathsf{OT}_{\mathbf{A},\varepsilon}(\tilde\mu_{0,t},\bar\mu_1)-\mathsf{OT}_{\mathbf{A},\varepsilon}(\bar\mu_0,\bar\mu_1)&\le t\int\bar\varphi_0^{(\tilde\mu_{0,t},\bar\mu_1)}d(\nu_0-\mu_0),\\
\mathsf{OT}_{\mathbf{A},\varepsilon}(\tilde\mu_{0,t},\bar\mu_1)-\mathsf{OT}_{\mathbf{A},\varepsilon}(\bar\mu_0,\bar\mu_1)&\ge t\int\bar\varphi_0^{(\bar\mu_0,\bar\mu_1)}d(\nu_0-\mu_0),
\end{aligned}\tag{3.49}$$

whereby

$$\begin{aligned}
&\left|\mathsf{OT}_{\mathbf{A},\varepsilon}(\tilde\mu_{0,t},\tilde\mu_{1,t})-\mathsf{OT}_{\mathbf{A},\varepsilon}(\bar\mu_0,\bar\mu_1)\right|\\
&\qquad\le t\Bigg(\left|\int\bar\varphi_0^{(\tilde\mu_{0,t},\bar\mu_1)}d(\nu_0-\mu_0)+\int\bar\varphi_1^{(\tilde\mu_{0,t},\tilde\mu_{1,t})}d(\nu_1-\mu_1)\right|\\
&\qquad\qquad\bigvee\left|\int\bar\varphi_0^{(\bar\mu_0,\bar\mu_1)}d(\nu_0-\mu_0)+\int\bar\varphi_1^{(\tilde\mu_{0,t},\bar\mu_1)}d(\nu_1-\mu_1)\right|\Bigg)
\end{aligned}\tag{3.50}$$

It suffices, therefore to show that both terms in absolute values on the right hand side converge to $\int\bar\varphi_0^{(\bar\mu_0,\bar\mu_1)}d(\nu_0-\mu_0)+\int\bar\varphi_1^{(\bar\mu_0,\bar\mu_1)}d(\nu_1-\mu_1)$ as $t\downarrow 0$.

To this end, fix an arbitrary sequence $t_n\downarrow 0$ and a subsequence $t_{n'}\downarrow 0$. By the Arzèla-Ascoli theorem, there is a subsequence $t_{n''}\downarrow 0$ along which $(\varphi_0^{(\tilde\mu_{0,t_{n''}},\tilde\mu_{1,t_{n''}})},\varphi_1^{(\tilde\mu_{0,t_{n''}},\tilde\mu_{1,t_{n''}})})$

converges to (ψ_0, ψ_1) uniformly on compact sets (in particular pointwise) where (ψ_0, ψ_1) is a pair of continuous functions. Given that the EOT potentials satisfy the estimates from Lemma 26 and that $\mu_0, \mu_1, \nu_0, \nu_1$ are 4-sub-Weibull, we may apply the dominated convergence theorem to obtain that, for every $(x, y) \in \mathbb{R}^{d_0} \times \mathbb{R}^{d_1}$,

$$
\begin{aligned}
e^{-\frac{\varphi_0^{(\tilde{\mu}_{0,t_{n''}},\tilde{\mu}_{1,t_{n''}})}(x)}{\varepsilon}} &= \int e^{\frac{\varphi_1^{(\tilde{\mu}_{0,t_{n''}},\tilde{\mu}_{1,t_{n''}})} - c_{\mathbf{A}}(x,\cdot)}{\varepsilon}} d\tilde{\mu}_{1,t_{n''}} \to \int e^{\frac{\psi_1 - c_{\mathbf{A}}(x,\cdot)}{\varepsilon}} d\bar{\mu}_1 = e^{-\frac{\psi_0(x)}{\varepsilon}}, \\
e^{-\frac{\varphi_1^{(\tilde{\mu}_{0,t_{n''}},\tilde{\mu}_{1,t_{n''}})}(y)}{\varepsilon}} &= \int e^{\frac{\varphi_0^{(\tilde{\mu}_{0,t_{n''}},\tilde{\mu}_{1,t_{n''}})} - c_{\mathbf{A}}(\cdot,y)}{\varepsilon}} d\tilde{\mu}_{0,t_{n''}} \to \int e^{\frac{\psi_0 - c_{\mathbf{A}}(\cdot,y)}{\varepsilon}} d\bar{\mu}_0 = e^{-\frac{\psi_1(y)}{\varepsilon}},
\end{aligned}
$$

so that (ψ_0, ψ_1) solve the Schrödinger system for $\mathsf{OT}_{\mathbf{A},\varepsilon}(\bar{\mu}_0, \bar{\mu}_1)$ on $\mathbb{R}^{d_0} \times \mathbb{R}^{d_1}$. Conclude from Lemma 26 that (ψ_0, ψ_1) coincides with $(\varphi_0^{(\bar{\mu}_0,\bar{\mu}_1)}, \varphi_1^{(\bar{\mu}_0,\bar{\mu}_1)})$ up to additive constants whence $\int \bar{\varphi}_1^{(\tilde{\mu}_{0,t_{n''}},\tilde{\mu}_{1,t_{n''}})} d(\nu_1 - \mu_1) \to \int \bar{\varphi}_1^{(\bar{\mu}_0,\bar{\mu}_1)} d(\nu_1 - \mu_1)$. By an analogous argument it holds that $\int \bar{\varphi}_0^{(\bar{\mu}_{0,t_{n'''}},\bar{\mu}_1)} d(\nu_0 - \mu_0) \to \int \bar{\varphi}_0^{(\bar{\mu}_0,\bar{\mu}_1)} d(\nu_0 - \mu_0)$, $\int \bar{\varphi}_1^{(\tilde{\mu}_{0,t_{n'''}},\bar{\mu}_1)} d(\nu_1 - \mu_1) \to \int \bar{\varphi}_1^{(\bar{\mu}_0,\bar{\mu}_1)} d(\nu_1 - \mu_1)$ up to passing to a further subsequence $t_{n'''}$. Conclude from equations (3.48) and (3.49) that

$$
\begin{aligned}
\lim_{n''' \to \infty} & \frac{\mathsf{OT}_{\mathbf{A},\varepsilon}(\tilde{\mu}_{0,t_{n'''}}, \tilde{\mu}_{1,t_{n'''}}) - \mathsf{OT}_{\mathbf{A},\varepsilon}(\bar{\mu}_0, \bar{\mu}_1)}{t_{n'''}} \\
&= \int \bar{\varphi}_0^{(\bar{\mu}_0,\bar{\mu}_1)} d(\nu_0 - \mu_0) + \int \bar{\varphi}_1^{(\bar{\mu}_0,\bar{\mu}_1)} d(\nu_1 - \mu_1).
\end{aligned}
$$

As the original choice of subsequence was arbitrary and the limit is independent of this choice of subsequence, conclude that

$$
\lim_{t \downarrow 0} \frac{\mathsf{OT}_{\mathbf{A},\varepsilon}(\tilde{\mu}_{0,t}, \tilde{\mu}_{1,t}) - \mathsf{OT}_{\mathbf{A},\varepsilon}(\bar{\mu}_0, \bar{\mu}_1)}{t} = \int \bar{\varphi}_0^{(\bar{\mu}_0,\bar{\mu}_1)} d(\nu_0 - \mu_0) + \int \bar{\varphi}_1^{(\bar{\mu}_0,\bar{\mu}_1)} d(\nu_1 - \mu_1). \quad (3.51)
$$

On the other hand, adopting the notation τ^t from the proof of Lemma 22 and letting $\bar{\pi}_t, \tilde{\pi}_t$ be optimal for $\mathsf{OT}_{\mathbf{A},\varepsilon}(\bar{\mu}_{0,t}, \bar{\mu}_{1,t})$ and $\mathsf{OT}_{\mathbf{A},\varepsilon}(\tilde{\mu}_{0,t}, \tilde{\mu}_{1,t})$ respectively, we have that

$$
\begin{aligned}
\mathsf{OT}_{\mathbf{A},\varepsilon}(\bar{\mu}_{0,t}, \bar{\mu}_{1,t}) - \mathsf{OT}_{\mathbf{A},\varepsilon}(\tilde{\mu}_{0,t}, \tilde{\mu}_{1,t}) &\leq \int c_{\mathbf{A}} \circ \tau^{-t} - c_{\mathbf{A}} d\tilde{\pi}_t, \\
\mathsf{OT}_{\mathbf{A},\varepsilon}(\bar{\mu}_{0,t}, \bar{\mu}_{1,t}) - \mathsf{OT}_{\mathbf{A},\varepsilon}(\tilde{\mu}_{0,t}, \tilde{\mu}_{1,t}) &\geq \int c_{\mathbf{A}} - c_{\mathbf{A}} \circ \tau^{t} d\bar{\pi}_t,
\end{aligned} \quad (3.52)
$$

as D_{KL} is translation invariant. Using the characterization of the Radon-Nikodym derivative of the EOT plan (see Section 3.2.3), we have

$$
\begin{aligned}
&\int c_{\mathbf{A}} \circ \tau^{-t} - c_{\mathbf{A}} d\tilde{\pi}_t \\
&= \int (c_{\mathbf{A}} \circ \tau^{-t} - c_{\mathbf{A}}) \frac{d\tilde{\pi}_t}{d\tilde{\mu}_{0,t} \otimes \tilde{\mu}_{1,t}} d\tilde{\mu}_{0,t} \otimes \tilde{\mu}_{1,t} \\
&= \int \left((c_{\mathbf{A}} \circ \tau^{-t} - c_{\mathbf{A}}) e^{\frac{\tilde{\varphi}_{0,t} \oplus \tilde{\varphi}_{1,t} - c_{\mathbf{A}}}{\varepsilon}} \right) (\cdot - \mathbb{E}_{\mu_0}[X], \cdot - \mathbb{E}_{\mu_1}[X]) d\mu_{0,t} \otimes \mu_{1,t},
\end{aligned}
$$

where $(\tilde{\varphi}_{0,t}, \tilde{\varphi}_{1,t})$ is a pair of EOT potentials for $\mathsf{OT}_{\mathbf{A},\varepsilon}(\tilde{\mu}_{0,t}, \tilde{\mu}_{1,t})$. Arguing as in the first part of the proof, if we fix a sequence $t_n \downarrow 0$ and a subsequence $t_{n'} \downarrow 0$, there exists a further subsequence $t_{n''} \downarrow 0$ along which $\tilde{\varphi}_{0,t_{n''}} \oplus \tilde{\varphi}_{1,t_{n''}} \to \varphi_0 \oplus \varphi_1$ pointwise on $\mathbb{R}^{d_0} \times \mathbb{R}^{d_1}$ where (φ_0, φ_1) is a pair of EOT potentials for $\mathsf{OT}_{\mathbf{A},\varepsilon}(\bar{\mu}_0, \bar{\mu}_1)$ so that $e^{\frac{\tilde{\varphi}_{0,t_{n''}} \oplus \tilde{\varphi}_{1,t_{n''}} - c_{\mathbf{A}}}{\varepsilon}} \to e^{\frac{\varphi_0 \oplus \varphi_1 - c_{\mathbf{A}}}{\varepsilon}}$ pointwise and we highlight that $e^{\frac{\varphi_0 \oplus \varphi_1 - c_{\mathbf{A}}}{\varepsilon}} = \frac{d\pi^{\mathbf{A}}}{d\bar{\mu}_0 \otimes \bar{\mu}_1}$ $\bar{\mu}_0 \otimes \bar{\mu}_1$-a.e. where $\pi^{\mathbf{A}}$ is the EOT plan for $\mathsf{OT}_{\mathbf{A},\varepsilon}(\bar{\mu}_0, \bar{\mu}_1)$. From (3.44), we have that

$$
\begin{aligned}
&t^{-1} \left(c_{\mathbf{A}} \circ \tau^{-t}(x,y) - c_{\mathbf{A}}(x,y) \right) \\
&\qquad\qquad \to 8\langle x, \mathbb{E}_{\nu_0}[X] - \mathbb{E}_{\mu_0}[X] \rangle \|y\|^2 + 8\langle y, \mathbb{E}_{\nu_1}[X] - \mathbb{E}_{\mu_1}[X] \rangle \|x\|^2 \\
&\qquad\qquad + 32 \left(\mathbb{E}_{\nu_0}[X] - \mathbb{E}_{\mu_0}[X] \right)^{\intercal} \mathbf{A} y + 32 x^{\intercal} \mathbf{A} \left(\mathbb{E}_{\nu_1}[X] - \mathbb{E}_{\mu_1}[X] \right)
\end{aligned}
$$

pointwise and, from the potential estimates, $t^{-1} \left(c_{\mathbf{A}} \circ \tau^{-t}(x,y) - c_{\mathbf{A}}(x,y) \right) e^{\frac{\tilde{\varphi}_{0,t} \oplus \tilde{\varphi}_{1,t} - c_{\mathbf{A}}}{\varepsilon}}$ is equiintegrable (in $t \in [0,1]$, $\mathbf{A} \in B_{\mathrm{F}}(M)$) with respect to any product of 4-sub-Weibull distributions. Letting h denote the limit in the above display, the dominated convergence theorem yields

$$
\begin{aligned}
t_{n''}^{-1} \int (c_{\mathbf{A}} \circ \tau^{-t_{n''}} - c_{\mathbf{A}}) d\tilde{\pi}_{t_{n''}} &\to \int \left(h \frac{d\pi^{\mathbf{A}}}{d\bar{\mu}_0 \otimes \bar{\mu}_1} \right) (\cdot - \mathbb{E}_{\mu_0}[X], \cdot - \mathbb{E}_{\mu_1}[X]) d\mu_0 \otimes \mu_1 \\
&= \int h d\pi^{\mathbf{A}} \\
&= \int 8\langle x, \mathbb{E}_{\nu_0}[X] - \mathbb{E}_{\mu_0}[X] \rangle \|y\|^2 \\
&\qquad\qquad + 8\langle y, \mathbb{E}_{\nu_1}[X] - \mathbb{E}_{\mu_1}[X] \rangle \|x\|^2 d\pi^{\mathbf{A}}(x,y),
\end{aligned}
$$

as the marginals of $\pi^{\mathbf{A}}$ are mean-zero. Since the limit is independent of the choice of original subsequence, $t^{-1} \int (c_{\mathbf{A}} \circ \tau^{-t} - c_{\mathbf{A}}) d\tilde{\pi}_t$ converges to the same limit. The same arguments apply to $\int c_{\mathbf{A}} - c_{\mathbf{A}} \circ \tau^t d\bar{\pi}_t$ so that

$$\frac{\mathsf{OT}_{\mathbf{A},1}(\bar{\mu}_{0,t},\bar{\mu}_{1,t}) - \mathsf{OT}_{\mathbf{A},1}(\tilde{\mu}_{0,t},\tilde{\mu}_{1,t})}{t} \to \int 8\langle x, \mathbb{E}_{\nu_0}[X] - \mathbb{E}_{\mu_0}[X]\rangle \|y\|^2 + 8\langle y, \mathbb{E}_{\nu_1}[X] - \mathbb{E}_{\mu_1}[X]\rangle \|x\|^2 d\pi^{\mathbf{A}}(x,y). \tag{3.53}$$

Combining (3.51) and (3.53) proves the claim. □

Lemma 30. If $(\mu_0,\mu_1) \in \mathcal{P}_{4,\sigma}(\mathbb{R}^{d_0}) \times \mathcal{P}_{4,\sigma}(\mathbb{R}^{d_1})$ for some $\sigma > 0$ and we set $\mathcal{A}=$ $\operatorname{argmin}_{B_{\mathrm{F}}(M)} \Phi_{(\bar{\mu}_0,\bar{\mu}_1)}$, then

$$\lim_{t\downarrow 0} \frac{\mathsf{S}_{2,\varepsilon}(\bar{\mu}_{0,t},\bar{\mu}_{1,t}) - \mathsf{S}_{2,\varepsilon}(\bar{\mu}_0,\bar{\mu}_1)}{t} = \inf_{\mathbf{A}\in\mathcal{A}} \left\{ \int g_{0,\pi^{\mathbf{A}}} + \bar{\varphi}_0^{\mathbf{A}} d(\nu_0 - \mu_0) + \int g_{1,\pi^{\mathbf{A}}} + \bar{\varphi}_1^{\mathbf{A}} d(\nu_1 - \mu_1) \right\},$$

where

$$g_{0,\pi} = 8\mathbb{E}_{(X,Y)\sim\pi}[\|Y\|^2 X]^\intercal(\cdot),$$

$$g_{1,\pi} = 8\mathbb{E}_{(X,Y)\sim\pi}[\|X\|^2 Y]^\intercal(\cdot),$$

$\bar{\varphi}_i = \varphi_i(\cdot - \mathbb{E}_{\mu_i}[X])$ for $i = 0,1$, and, for $\mathbf{A} \in \mathcal{A}$, $\pi^{\mathbf{A}}$ is the unique OT plan for $\mathsf{OT}_{\mathbf{A}}(\bar{\mu}_0,\bar{\mu}_1)$, and $(\varphi_0^{\mathbf{A}},\varphi_1^{\mathbf{A}})$ is any choice of EOT potentials from Lemma 26.

Proof. Let $\mathbf{A}^\star \in \mathcal{A}$ be arbitrary. By Lemma 29,

$$\begin{aligned}
&\mathsf{S}_{2,\varepsilon}(\bar{\mu}_{0,t},\bar{\mu}_{1,t}) - \mathsf{S}_{2,\varepsilon}(\bar{\mu}_0,\bar{\mu}_1) \\
&\qquad \le \mathsf{OT}_{\mathbf{A}^\star,\varepsilon}(\bar{\mu}_{0,t},\bar{\mu}_{1,t}) - \mathsf{OT}_{\mathbf{A}^\star,\varepsilon}(\bar{\mu}_0,\bar{\mu}_1) \\
&\qquad = \int g_{0,\pi^{\mathbf{A}^\star}} + \bar{\varphi}_0^{\mathbf{A}^\star} d(\nu_0 - \mu_0) + \int g_{1,\pi^{\mathbf{A}^\star}} + \bar{\varphi}_1^{\mathbf{A}^\star} d(\nu_1 - \mu_1) + o(t).
\end{aligned}$$

It follows that

$$\limsup_{t\downarrow 0} \frac{\mathsf{S}_{2,\varepsilon}(\bar{\mu}_{0,t},\bar{\mu}_{1,t}) - \mathsf{S}_{2,\varepsilon}(\bar{\mu}_0,\bar{\mu}_1)}{t} \le \inf_{\mathbf{A}\in\mathcal{A}} \left\{ \int g_{0,\pi^{\mathbf{A}}} + \bar{\varphi}_0^{\mathbf{A}} d(\nu_0 - \mu_0) + \int g_{1,\pi^{\mathbf{A}}} + \bar{\varphi}_1^{\mathbf{A}} (\nu_1 - \mu_1) \right\}. \tag{3.54}$$

To prove the complementary inequality, let $\mathbf{A}_t$ be optimal for $\mathsf{S}_{2,\varepsilon}(\bar{\mu}_{0,t},\bar{\mu}_{1,t})$ for every $t\in[0,1]$. By the estimates for the EOT potentials, equations (3.50) and (3.52), and the 4-sub-Weibull assumption, $\mathbf{A}\in B_{\mathrm{F}}(M)\mapsto 32\|\mathbf{A}\|_{\mathrm{F}}^2+\mathsf{OT}_{\mathbf{A},\varepsilon}(\bar{\mu}_{0,t},\bar{\mu}_{1,t})$ converges uniformly to $\mathbf{A}\in B_{\mathrm{F}}(M)\mapsto 32\|\mathbf{A}\|_{\mathrm{F}}^2+\mathsf{OT}_{\mathbf{A},\varepsilon}(\bar{\mu}_0,\bar{\mu}_1)$ as $t\downarrow 0$. Proceeding as in the proof of Lemma 23 we obtain that any cluster point of $(\mathbf{A}_t)_{t\downarrow 0}$ is an element of $\mathscr{A}$.

Now, fix a sequence $t_n\downarrow 0$ and an arbitrary subsequence $t_{n'}\downarrow 0$. By the Bolzano-Weierstrass theorem, $(\mathbf{A}_{t_{n'}})_{n'\in\mathbb{N}}$ admits a cluster point $\mathbf{A}^\star\in\mathscr{A}$. Observe that

$$\begin{aligned}
&\frac{\mathsf{S}_{2,\varepsilon}(\bar{\mu}_{0,t_{n'}},\bar{\mu}_{1,t_{n'}})-\mathsf{S}_{2,\varepsilon}(\bar{\mu}_0,\bar{\mu}_1)}{t_{n'}}\\
&\geq\frac{\mathsf{OT}_{\mathbf{A}_{t_{n'}},\varepsilon}(\bar{\mu}_{0,t_{n'}},\bar{\mu}_{1,t_{n'}})-\mathsf{OT}_{\mathbf{A}_{t_{n'}},\varepsilon}(\bar{\mu}_0,\bar{\mu}_1)}{t_{n'}}\\
&\geq\int\bar{\varphi}_0^{\mathbf{A}_{t_{n'}},(\bar{\mu}_0,\bar{\mu}_1)}d(\nu_0-\mu_0)+\int\bar{\varphi}_1^{\mathbf{A}_{t_{n'}},(\bar{\mu}_{0,t_{n'}},\bar{\mu}_1)}d(\nu_1-\mu_1)+t_{n'}^{-1}\int c_{\mathbf{A}_{t_{n'}}}-c_{\mathbf{A}_{t_{n'}}}\circ\tau^{t_{n'}}\,d\bar{\pi}^{\mathbf{A}_{t_{n'}}}
\end{aligned}$$

due to (3.49) and (3.52). It remains to show that the term above converges to $\int g_{0,\pi^{\mathbf{A}^\star}}d(\nu_0-\mu_0)+\int g_{1,\pi^{\mathbf{A}^\star}}d(\nu_1-\mu_1)$ and that $(\bar{\varphi}_0^{\mathbf{A}_{t_{n'}},(\bar{\mu}_0,\bar{\mu}_1)},\bar{\varphi}_1^{\mathbf{A}_{t_{n'}},(\bar{\mu}_{0,t_{n'}},\bar{\mu}_1)})$ converges pointwise to $(\bar{\varphi}_0^{\mathbf{A}^\star,(\bar{\mu}_0,\bar{\mu}_1)},\bar{\varphi}_1^{\mathbf{A}^\star,(\bar{\mu}_0,\bar{\mu}_1)})$ (up to additive constants).

Given Lemma 26, the Arzelà-Ascoli theorem implies that there exists a subsequence $t_{n''}$ along which $(\varphi_0^{\mathbf{A}_{t_{n''}},(\bar{\mu}_{0,t_{n''}},\bar{\mu}_{1,t_{n''}})},\varphi_1^{\mathbf{A}_{t_{n''}},(\bar{\mu}_{0,t_{n''}},\bar{\mu}_{1,t_{n''}})})$ converges to a pair (φ_0,φ_1) of continuous functions uniformly on compact sets (hence pointwise); similar implications hold for the other pairs of EOT potentials. Given that $c_{\mathbf{A}_{n''}}$ converges pointwise to $c_{\mathbf{A}^\star}$ and that μ_0,μ_1,ν_0,ν_1 are 4-sub-Weibull it follows from the dominated convergence theorem that

$$\begin{aligned}
e^{-\frac{\varphi_0^{\mathbf{A}_{t_{n''}},(\bar{\mu}_{0,t_{n''}},\bar{\mu}_{1,t_{n''}})}(x)}{\varepsilon}}&=\int e^{\frac{\varphi_1^{\mathbf{A}_{t_{n''}},(\bar{\mu}_{0,t_{n''}},\bar{\mu}_{1,t_{n''}})}-c_{\mathbf{A}_{t_{n''}}}(x,\cdot)}{\varepsilon}}d\bar{\mu}_{1,t_{n''}}\\
&\to\int e^{\frac{\varphi_1-c_{\mathbf{A}^\star}(x,\cdot)}{\varepsilon}}d\bar{\mu}_1=e^{-\frac{\varphi_0(x)}{\varepsilon}},\\
e^{-\frac{\varphi_1^{\mathbf{A}_{t_{n''}},(\bar{\mu}_{0,t_{n''}},\bar{\mu}_{1,t_{n''}})}(y)}{\varepsilon}}&=\int e^{\frac{\varphi_0^{\mathbf{A}_{t_{n''}},(\bar{\mu}_{0,t_{n''}},\bar{\mu}_{1,t_{n''}})}-c_{\mathbf{A}_{t_{n''}}}(\cdot,y)}{\varepsilon}}d\bar{\mu}_{0,t_{n''}}\\
&\to\int e^{\frac{\varphi_0-c_{\mathbf{A}^\star}(\cdot,y)}{\varepsilon}}d\bar{\mu}_0=e^{-\frac{\varphi_1(y)}{\varepsilon}},
\end{aligned}$$

so that (φ_0,φ_1) is an optimal pair for $\mathsf{OT}_{\mathbf{A}^\star_\varepsilon}(\bar{\mu}_0,\bar{\mu}_1)$ and similarly for the other EOT potentials. Conclude that $e^{\frac{\varphi_0^{\mathbf{A}_{t_{n''}},(\bar{\mu}_{0,t_{n''}},\bar{\mu}_{1,t_{n''}})}\oplus\varphi_1^{\mathbf{A}_{t_{n''}},(\bar{\mu}_{0,t_{n''}},\bar{\mu}_{1,t_{n''}})}-c_{\mathbf{A}_{t_{n''}}}}{\varepsilon}} \to e^{\frac{\varphi_0\oplus\varphi_1-c_{\mathbf{A}_\star}}{\varepsilon}}$ pointwise on $\mathbb{R}^{d_0}\times\mathbb{R}^{d_1}$ and we recall that $e^{\frac{\varphi_0\oplus\varphi_1-c_{\mathbf{A}_\star}}{\varepsilon}} = \frac{d\pi^{\mathbf{A}}}{d\bar{\mu}_0\otimes\bar{\mu}_1}$ $\bar{\mu}_0\otimes\bar{\mu}_1$-a.e. so that, arguing as in the proof of Lemma 29,

$$t_{n'}^{-1}\int c_{\mathbf{A}_{t_{n'}}} - c_{\mathbf{A}_{t_{n'}}}\circ\tau^{t_{n'}}\,d\bar{\pi}^{\mathbf{A}_{t_{n'}}} \to \int\Big(8\langle x,\mathbb{E}_{\nu_0}[X]-\mathbb{E}_{\mu_0}[X]\rangle\|y\|^2$$
$$+8\langle y,\mathbb{E}_{\nu_1}[X]-\mathbb{E}_{\mu_1}[X]\rangle\|x\|^2\Big)\,d\pi^{\mathbf{A}^\star}(x,y).$$

Conclude that

$$\begin{aligned}
&\frac{\mathsf{S}_{2,\varepsilon}(\bar{\mu}_{0,t_{n''}},\bar{\mu}_{1,t_{n''}})-\mathsf{S}_{2,\varepsilon}(\bar{\mu}_0,\bar{\mu}_1)}{t_{n''}}\\
&\quad\to\int\bar{\varphi}_0^{\mathbf{A}^\star,(\bar{\mu}_0,\bar{\mu}_1)}d(\nu_0-\mu_0)+\int\bar{\varphi}_1^{\mathbf{A}^\star,(\bar{\mu}_0,\bar{\mu}_1)}d(\nu_1-\mu_1)\\
&\quad+\int 8\langle x,\mathbb{E}_{\nu_0}[X]-\mathbb{E}_{\mu_0}[X]\rangle\|y\|^2+8\langle y,\mathbb{E}_{\nu_1}[X]-\mathbb{E}_{\mu_1}[X]\rangle\|x\|^2 d\pi^{\mathbf{A}^\star}(x,y),\\
&\quad\geq\inf_{\mathbf{A}\in\mathcal{A}}\left\{\int g_{0,\pi^{\mathbf{A}}}+\bar{\varphi}_0^{\mathbf{A}}d(\nu_0-\mu_0)+\int g_{1,\pi^{\mathbf{A}}}+\bar{\varphi}_1^{\mathbf{A}}(\nu_1-\mu_1)\right\}.
\end{aligned}$$

Given that this final lower bound is independent of the original choice of subsequence,

$$\begin{aligned}
&\liminf_{t\downarrow 0}\frac{\mathsf{S}_{2,\varepsilon}(\bar{\mu}_{0,t},\bar{\mu}_{1,t})-\mathsf{S}_{2,\varepsilon}(\bar{\mu}_0,\bar{\mu}_1)}{t}\\
&\qquad\geq\inf_{\mathbf{A}\in\mathcal{A}}\left\{\int g_{0,\pi^{\mathbf{A}}}+\bar{\varphi}_0^{\mathbf{A}}d(\nu_0-\mu_0)+\int g_{1,\pi^{\mathbf{A}}}+\bar{\varphi}_1^{\mathbf{A}}(\nu_1-\mu_1)\right\}
\end{aligned}$$

which, in combination with (3.54) proves the claim. □

Lemma 31. For any pairs $(\nu_0,\nu_1),(\rho_0,\rho_1)\in\mathcal{P}_{4,\sigma}(\mathbb{R}^{d_0})\times\mathcal{P}_{4,\sigma}(\mathbb{R}^{d_1})$, we have that $\left|\mathsf{S}_{2,\varepsilon}(\bar{\nu}_0,\bar{\nu}_1)-\mathsf{S}_{2,\varepsilon}(\bar{\rho}_0,\bar{\rho}_1)\right|\leq C\|\nu_0\otimes\nu_1-\rho_0\otimes\rho_1\|_{\infty,\mathcal{F}^\oplus}$ for some constant $C>0$ which is independent of $(\nu_0,\nu_1),(\rho_0,\rho_1)$.

Proof. Observe that

$$|\mathsf{S}_{2,\varepsilon}(\bar{\nu}_0,\bar{\nu}_1)-\mathsf{S}_{2,\varepsilon}(\bar{\rho}_0,\bar{\rho}_1)|\leq\sup_{\mathbf{A}\in B_{\mathrm{F}}(M)}\left|\mathsf{OT}_{\mathbf{A},\varepsilon}(\bar{\nu}_0,\bar{\nu}_1)-\mathsf{OT}_{\mathbf{A},\varepsilon}(\bar{\rho}_0,\bar{\rho}_1)\right|$$

and

$$\begin{aligned}\mathsf{OT}_{\mathbf{A},\varepsilon}(\bar\nu_0,\bar\nu_1)-\mathsf{OT}_{\mathbf{A},\varepsilon}(\bar\rho_0,\bar\rho_1) &= \mathsf{OT}_{\mathbf{A},\varepsilon}(\bar\nu_0,\bar\nu_1)-\mathsf{OT}_{\mathbf{A},\varepsilon}(\bar\nu_0,\bar\rho_1)\\ &\quad+\mathsf{OT}_{\mathbf{A},\varepsilon}(\bar\nu_0,\bar\rho_1)-\mathsf{OT}_{\mathbf{A},\varepsilon}(\bar\rho_0,\bar\rho_1)\\ &\le \int \varphi_1^{\mathbf{A},(\bar\nu_0,\bar\nu_1)}d(\bar\nu_1-\bar\rho_1)+\int \varphi_0^{\mathbf{A},(\bar\nu_0,\bar\rho_1)}d(\bar\nu_0-\bar\rho_0)\\ \mathsf{OT}_{\mathbf{A},\varepsilon}(\bar\nu_0,\bar\nu_1)-\mathsf{OT}_{\mathbf{A},\varepsilon}(\bar\rho_0,\bar\rho_1) &\ge \int \varphi_0^{\mathbf{A},(\bar\nu_0,\bar\rho_1)}d(\bar\nu_0-\bar\rho_0)+\int \varphi_1^{\mathbf{A},(\bar\rho_0,\bar\rho_1)}d(\bar\nu_1-\bar\rho_1).\end{aligned} \tag{3.55}$$

By a Taylor expansion, for any $x\in\mathbb{R}^{d_1}$,

$$\begin{aligned}\varphi_1^{\mathbf{A},(\bar\nu_0,\bar\nu_1)}(x-\mathbb{E}_{\nu_1}[X]) &= \varphi_1^{\mathbf{A},(\bar\nu_0,\bar\nu_1)}(x-\mathbb{E}_{\rho_1}[X])\\ &\quad+\sum_{i=1}^{d_1}(\mathbb{E}_{\rho_1}[X]-\mathbb{E}_{\nu_1}[X])_i\partial_i\varphi_1^{\mathbf{A},(\bar\nu_0,\bar\nu_1)}(x-(1-c)\mathbb{E}_{\rho_1}[X]-c\mathbb{E}_{\nu_1}[X]),\end{aligned}$$

for some $c\in(0,1)$. It follows from the uniform bounds on the 4-th moments of 4-sub-Weibull distributions established in Lemma 25 and the estimates on the derivatives on the EOT potentials (see Lemma 26) that

$$\begin{aligned}&\left|\varphi_1^{\mathbf{A},(\bar\nu_0,\bar\nu_1)}(x-\mathbb{E}_{\nu_1}[X])-\varphi_1^{\mathbf{A},(\bar\nu_0,\bar\nu_1)}(x-\mathbb{E}_{\rho_1}[X])\right|\\ &\quad=\left|\sum_{i=1}^{d_1}(\mathbb{E}_{\rho_1}[X]-\mathbb{E}_{\nu_1}[X])_i\partial_i\varphi_1^{\mathbf{A},(\bar\nu_0,\bar\nu_1)}(x-(1-c)\mathbb{E}_{\rho_1}[X]-c\mathbb{E}_{\nu_1}[X])\right|\\ &\quad\le\sum_{i=1}^{d_1}\left|(\mathbb{E}_{\rho_1}[X]-\mathbb{E}_{\nu_1}[X])_i\right|K_1(1+\|x-(1-c)\mathbb{E}_{\rho_1}[X]-c\mathbb{E}_{\nu_1}[X]\|^3)\\ &\quad\le C_1(1+\|x\|^3)\max_{i=1}^{d_1}\left|(\mathbb{E}_{\rho_1}[X]-\mathbb{E}_{\nu_1}[X])_i\right|,\end{aligned}$$

where K_1 is the constant from Lemma 26 corresponding to $k=1$ and C_1 is a constant depending only on σ, d_0, d_1, and ε. It is easy to see that, for $i\in[N]$, $(\mathbb{E}_{\rho_1}[X]-\mathbb{E}_{\nu_1}[X])_i = \int x_i d(\rho_1-\nu_1)(x)\le\|\nu_0\otimes\nu_1-\rho_0\otimes\rho_1\|_{\infty,\mathcal{F}^\oplus}$ given that $x_i\in\mathcal{F}_1$ and that $0\in\mathcal{F}_0$.

In sum,

$$\begin{aligned}\int\varphi_1^{\mathbf{A},(\bar\nu_0,\bar\nu_1)}d(\bar\nu_1-\bar\rho_1) &= \underbrace{\int\varphi_1^{\mathbf{A},(\bar\nu_0,\bar\nu_1)}(\cdot-\mathbb{E}_{\nu_1}[X])-\varphi_1^{\mathbf{A},(\bar\nu_0,\bar\nu_1)}(\cdot-\mathbb{E}_{\rho_1}[X])d\nu_1}_{\le C_1(1+M_3(\nu_1))\|\nu_0\otimes\nu_1-\rho_0\otimes\rho_1\|_{\infty,\mathcal{F}^\oplus}}\\ &\quad+\int\varphi_1^{\mathbf{A},(\bar\nu_0,\bar\nu_1)}(\cdot-\mathbb{E}_{\rho_1}[X])d(\nu_1-\rho_1).\end{aligned}$$

Analogous bounds hold for each of the integrals in (3.55) whereby

$$
\begin{aligned}
&\left|\mathsf{OT}_{\mathbf{A},\varepsilon}(\bar{\nu}_0,\bar{\nu}_1)-\mathsf{OT}_{\mathbf{A},\varepsilon}(\bar{\rho}_0,\bar{\rho}_1)\right| \\
&\leq \left(\left|\int \varphi_1^{\mathbf{A},(\bar{\nu}_0,\bar{\nu}_1)}(\cdot-\mathbb{E}_{\rho_1}[X])d(\nu_1-\rho_1)+\int \varphi_0^{\mathbf{A},(\bar{\nu}_0,\bar{\rho}_1)}(\cdot-\mathbb{E}_{\rho_0}[X])d(\nu_0-\rho_0)\right|\right. \\
&\left.\bigvee \left|\int \varphi_0^{\mathbf{A},(\bar{\nu}_0,\bar{\rho}_1)}(\cdot-\mathbb{E}_{\rho_0}[X])d(\nu_0-\rho_0)+\int \varphi_0^{\mathbf{A},(\bar{\rho}_0,\bar{\rho}_1)}(\cdot-\mathbb{E}_{\rho_1}[X])d(\nu_1-\rho_1)\right|\right) \\
&+C_2\|\nu_0\otimes\nu_1-\rho_0\otimes\rho_1\|_{\infty,\mathcal{F}^{\oplus}},
\end{aligned}
$$

where C_2 depends only on $\sigma, d_0, d_1, \varepsilon$. As $\varphi_0^{\mathbf{A},(\bar{\nu}_0,\bar{\rho}_1)}(\cdot-\mathbb{E}_{\rho_0}[X])\oplus\varphi_1^{\mathbf{A},(\bar{\nu}_0,\bar{\nu}_1)}(\cdot-\mathbb{E}_{\rho_1}[X])$ and $\varphi_0^{\mathbf{A},(\bar{\nu}_0,\bar{\rho}_1)}(\cdot-\mathbb{E}_{\rho_0}[X])\oplus\varphi_0^{\mathbf{A},(\bar{\rho}_0,\bar{\rho}_1)}(\cdot-\mathbb{E}_{\rho_1}[X])$ are elements of $\mathcal{F}^{\oplus}$ up to scaling by a constant that does not depend on $\nu_0,\nu_1,\rho_0,\rho_1$, it follows that $\left|\mathsf{OT}_{\mathbf{A},\varepsilon}(\bar{\nu}_0,\bar{\nu}_1)-\mathsf{OT}_{\mathbf{A},\varepsilon}(\bar{\rho}_0,\bar{\rho}_1)\right| \leq C_3\|\nu_0\otimes\nu_1-\rho_0\otimes\rho_1\|_{\infty,\mathcal{F}^{\oplus}}$ for some C_3 which is independent of $\nu_0,\nu_1,\rho_0,\rho_1$. □

Proof of Theorem 15. The claimed Gâteaux differentiability follows from Lemmas 27 and 30 whereas Lipschitz continuity follows from Lemmas 28 and 31 by applying the variational form (3.8). □

Proof of Theorem 16

Given Theorem 15, the proof of the first statement in Theorem 16 follows directly from Proposition 16. Under the assumption that $\varepsilon > 16\sqrt{M_4(\bar{\mu}_0)M_4(\bar{\mu}_1)}$, Theorem 6 in [164] asserts that $\Phi_{(\bar{\mu}_0,\bar{\mu}_1)}$ is strictly convex so that it admits at most one minimizer, $\mathbf{A}^{\star}$. Recall from the discussion following (3.9) that $\mathcal{A}$ is always non-empty so that $\mathcal{A}=\{\mathbf{A}^{\star}\}$ and the limit distribution simplifies to $\chi_{\mu_0\otimes\mu_1}((f_0+g_{0,\pi}+\bar{\varphi}_0^{\mathbf{A}^{\star}})\oplus(f_1+g_{1,\pi}+\bar{\varphi}_1^{\mathbf{A}^{\star}}))$

As for the statement regarding the empirical measures, it will follow upon showing that $\mathcal{F}_0$ and $\mathcal{F}_1$ are, respectively, μ_0- and μ_1-Donsker and that the empirical measures are 4-sub-Weibull with some parameter with probability approaching one. As noted in the proof of Theorem 10, Donskerness of the classes and the independence assumption imply that

$\sqrt{n}(\hat{\mu}_{0,n} - \mu_0) \otimes (\hat{\mu}_{1,n} - \mu_1) \xrightarrow{d} G_{\mu_0 \otimes \mu_1}$ in $\ell^\infty(\mathcal{F}^\oplus)$ where $G_{\mu_0 \otimes \mu_1}(f_0 \oplus f_1) = G_{\mu_0}(f_0) + G_{\mu_1}(f_1)$ and G_{μ_0}, G_{μ_1} are, respectively, tight μ_0- and μ_1-Brownian bridge processes in $\ell^\infty(\mathcal{F}_0)$ and $\ell^\infty(\mathcal{F}_1)$.

Donskerness of $\mathcal{F}_0$ and $\mathcal{F}_1$ with respect to any 4-sub-Weibull distribution follows from a direct adaptation of Lemma E.24 in [106] (see also Lemma 8 in [104]) and is thus omitted.

Fixing $i \in \{0, 1\}$ and some $\tilde{\sigma} > \sigma$, the law of large numbers yields

$$\int e^{\frac{\|\cdot\|^4}{2\tilde{\sigma}^2}} d\hat{\mu}_{i,n} \to \int e^{\frac{\|\cdot\|^4}{2\tilde{\sigma}^2}} d\mu_i \leq 2^{\frac{\sigma^2}{\tilde{\sigma}^2}} < 2$$

a.s. so that $\hat{\mu}_{i,n}$ is 4-sub-Weibull with parameter $\tilde{\sigma}^2$ with probability approaching 1.

The proof of bootstrap consistency follows the same lines as the proof of Theorem 7 in [106] and is thus omitted for brevity. □

3.8.6 Proofs for Section 3.6

Proof of Proposition 8

If $\mathsf{D}(\mu_0, \mu_1) = 0$, there exists an isometry $T : \mathrm{spt}(\mu_0) \to \mathrm{spt}(\mu_1)$ for which $T_\sharp \mu_0 = \mu_1$. For any $i \in [N]$, let $T(-e_i) = y$ and observe that, by construction,

$$\frac{1}{(N-1)N(N+1)} \sum_{\substack{i,j=1 \\ i \neq j}}^{N} \eta_{\rho_{0,ij}}(-e_i) + \frac{1}{N+1} = \mu_0(T^{-1}(\{y\})) = \mu_1(y),$$

where

$$\mu_1(y) = \frac{1}{(N-1)N(N+1)} \sum_{\substack{i,j=1 \\ i \neq j}}^{N} \eta_{\rho_{1,ij}}(y) + \frac{1}{N+1} \sum_{i=1}^{N} \delta_{-e_i}(y).$$

Note that $\sum_{\substack{i,j=1\\ i\neq j}}^{N} \eta_{\rho_{1,ij}}(y) \leq N(N-1)$ with equality if and only if $\eta_{\rho_{1,ij}}(y) = 1$ for every $i, j \in [N]$ with $i < j$. Thus, if $y \notin (-e_i)_{i=1}^N$, it holds that $\mu_1(y) \leq \frac{1}{N+1}$ with equality if and only if

$$1 = \eta_{\rho_{1,ij}}(y) = \int \mathbb{1}_{\{y\}}(-e_i + te_j)d\rho_{ij}(t)$$

for every $i, j \in [N], i < j$ i.e. $y = -e_i + se_j$ for some $s \in [0, \infty)$ and $\rho_{ij}(\{s\}) = 1$, evidently this equality cannot hold simultaneously for all such i, j. Consequently, $y \in (-e_i)_{i=1}^N$ so that, for every $i \in [N]$, $T(-e_i) = -e_{\sigma'(i)}$ for some permutation σ' on N elements.

By extending the isometry to an isometric self-map on $\mathbb{R}^N$, we have by the Mazur-Ulam theorem that $T(x) = Ux + b$ for some orthogonal matrix $U \in \mathbb{R}^{N\times N}$ and $b \in \mathbb{R}^N$. For any $i \in [N]$,

$$\begin{aligned} 1 = \|T(-e_i)\|^2 = \|-Ue_i\|^2 + 2b^\intercal(U(-e_i)) + \|b\|^2 = \|e_i\|^2 + 2b^\intercal(T(-e_i) - b) + \|b\|^2 \\ = 1 - \|b\|^2 - 2b^\intercal e_{\sigma'(i)}. \end{aligned}$$

As the above equality holds for every $i \in [N]$ and σ' is a permutation, it follows that $\|b\|^2 = -2b_i$ for every $i \in [N]$ i.e. $b = -\frac{2}{N}(1, \dots, 1)$ or $b = (0, \dots, 0)$. Let $t \in [0, \infty)$ and $i \neq j \in [N]$ be arbitrary and observe that

$$T(-e_i + te_j) = U(-e_i + te_j) + b = -e_{\sigma'(i)} - (U(-e_j) + b) + b = -e_{\sigma'(i)} + te_{\sigma'(j)} + b.$$

Since $T_\sharp\mu_0 = \mu_1$, for any Borel set $I \subset [0, \infty)$,

$$\begin{aligned} \mu_1(\cup_{t\in I}\{-e_{\sigma'(i)} + te_{\sigma'(j)} + b\}) &= \mu_0(\cup_{t\in I}T^{-1}(\{-e_{\sigma'(i)} + te_{\sigma'(j)} + b\})) \\ &= \mu_0(\cup_{t\in I}\{-e_i + te_j\}). \end{aligned}$$

If $b = -\frac{2}{N}(1, \dots, 1)$, $\cup_{t\in I}\left\{-e_{\sigma'(i)} + te_{\sigma'(j)} + b\right\}$ has measure 0 with respect to μ_1 whereas $\mu_0(\cup_{t\in I}\left\{-e_i + t_{e_j}\right\}) \neq 0$ for some set I (e.g. $0 \in I$). Conclude that $b = 0$ and, for every measurable subset, I, of $[0, \infty)$ and every $i, j \in [N]$ with $i < j$,

$$\begin{aligned} \frac{\rho_{0,ij}(I)}{N(N+1)(N-1)} + \frac{1}{N+1}\delta_0(I) &= \mu_1(\cup_{t\in I}\{-e_{\sigma'(i)} + te_{\sigma'(j)}\}) \\ &= \frac{\rho_{1,\sigma'(i)\sigma'(j)}(I)}{N(N+1)(N-1)} + \frac{1}{N+1}\delta_0(I). \end{aligned}$$

It follows that $\rho_{0,ij} = \rho_{1,\sigma'(i)\sigma'(j)}$ as measures on $\mathbb{R}$ or, equivalently, $\rho_{1,ij} = \rho_{0,\sigma(i)\sigma(j)}$ where σ is the inverse of σ' for every $i, j \in [N]$ with $i \neq j$

Conversely, if there exists a permutation σ for which $\rho_{1,ij} = \rho_{0,\sigma(i)\sigma(j)}$ for every $i, j \in [N]$ with $i \neq j$, consider the linear isometry $T : (x_1, \dots, x_N) \in \mathbb{R}^N \mapsto (x_{\sigma'(1)}, \dots, x_{\sigma'(N)}) \in \mathbb{R}^N$, where σ' is the inverse of σ. For any Borel set $A \subset \mathbb{R}^N$,

$$T_\sharp \mu_0(A) = \frac{1}{N(N+1)(N-1)} \sum_{\substack{i,j=1 \\ i\neq j}}^N \eta_{\rho_{0,ij}}(T^{-1}(A)) + \frac{1}{N+1} \sum_{i=1}^N \delta_{-e_i}(T^{-1}(A)).$$

It is clear that $\sum_{i=1}^N \delta_{-e_i}(T^{-1}(A)) = \sum_{i=1}^N \delta_{-e_{\sigma'(i)}}(A)$, whereas

$$\begin{aligned}
\eta_{\rho_{0,ij}}(T^{-1}(A)) = \int \mathbb{1}_{T^{-1}(A)}(-e_i + te_j) d\rho_{0,ij}(t) &= \int \mathbb{1}_A(-e_{\sigma'(i)} + te_{\sigma'(j)}) d\rho_{0,ij}(t) \\
&= \int \mathbb{1}_A(-e_{\sigma'(i)} + te_{\sigma'(j)}) d\rho_{1,\sigma'(i)\sigma'(j)}(t) \\
&= \eta_{\rho_{1,\sigma'(i)\sigma'(j)}}(A),
\end{aligned}$$

the previous two displays imply that $T_\sharp \mu_0 = \mu_1$ whereby $\mathsf{D}(\mu_0, \mu_1) = 0$. ☐

Proof of Theorem 17

For a set $A \subset \mathbb{R}^N$, let $\mathcal{F}(A, L)$ denote the set of convex L-Lipschitz functions on A. For an arbitrary $f \in \mathcal{F}([-1, K]^N, 1)$

$$\begin{aligned}
\mu_{0,n}(f) &= \frac{1}{(N-1)N(N+1)} \sum_{\substack{i,j=1 \\ i\neq j}}^N \eta_{\hat{\rho}_{0,ij,n}}(f) + \frac{1}{N+1} \sum_{i=1}^N f(-e_i), \\
&= \frac{1}{(N-1)N(N+1)} \sum_{\substack{i,j=1 \\ i\neq j}}^N \frac{1}{n} \sum_{k=1}^n f(-e_i + \mathbf{M}_{ij}^{G_{0,k}} e_j) + \frac{1}{N+1} \sum_{i=1}^N f(-e_i),
\end{aligned}$$

whereas

$$\mu_0(f) = \frac{1}{(N-1)N(N+1)} \sum_{\substack{i,j=1 \\ i\neq j}}^N \int f(-e_i + te_j) d\rho_{0,ij}(t) + \frac{1}{N+1} \sum_{i=1}^N f(-e_i),$$

thus

$$
\begin{aligned}
&(\mu_{0,n} - \mu_0)(f) \\
&= \frac{1}{(N-1)N(N+1)} \sum_{\substack{i,j=1 \\ i \neq j}}^{N} \left(\frac{1}{n} \sum_{k=1}^{n} f(-e_i + \mathbf{M}_{ij}^{G_{0,k}} e_j) - \int f(-e_i + te_j) d\rho_{0,ij}(t) \right) \\
&= \frac{1}{(N-1)N(N+1)} \sum_{\substack{i,j=1 \\ i \neq j}}^{N} (\hat{\rho}_{0,ij,n} - \rho_{0,ij}) \left(f(-e_i + (\cdot)e_j) \right).
\end{aligned}
\tag{3.56}
$$

In sum,

$$
\begin{aligned}
&\mathbb{E}\left[\sup_{f \in \mathcal{F}([-1,K]^N,1)} |(\mu_{0,n} - \mu_0)(f)| \right] \\
&\leq \frac{1}{(N-1)N(N+1)} \sum_{\substack{i,j=1 \\ i \neq j}}^{N} \mathbb{E}\left[\sup_{f \in \mathcal{F}([-1,K]^N,1)} |(\hat{\rho}_{0,ij,n} - \rho_{0,ij}) \left(f(-e_i + (\cdot)e_j) \right)| \right]
\end{aligned}
$$

Since $t \in [0, K] \mapsto f(-e_i + te_j)$ is 1-Lipschitz and convex[9] the supremum on the right hand side can be replaced by a supremum over $\mathcal{F}([0, K], 1)$, which admits an L^∞ metric entropy scaling as $N(\mathcal{F}([0, K], 1), \epsilon, L^\infty) \lesssim \epsilon^{-\frac{1}{2}}$ (see Theorem 1 in [112]). Evidently, the same result if $\mu_{0,n} - \mu_0$ is replaced by $\mu_{1,n} - \mu_1$. The remainder of the proof for the expected rate of convergence follows by the same argument as the proof of Theorem 3 in [223] with only minor modifications.

The limit distribution results follow from the proofs of Theorem 9 and Corollary 5 upon characterizing the weak limit of $\sqrt{n}(\mu_{0,n} - \mu_0)$; the arguments evidently also apply to $\sqrt{n}(\mu_{1,n} - \mu_1)$. We provide a general argument which applies to compactly supported weight distributions and specialize to the finitely discrete case at the end, as the stability results derived herein can only be applied in this case.

To this end, it was previously noted that $N(\mathcal{F}([0, K], 1), \epsilon, L^\infty) \lesssim \epsilon^{-\frac{1}{2}}$ so that, by Theorem 2.5.2 in [199], $\mathcal{F}([0, K], 1)$ is $\rho_{0,ij}$-Donsker for every $i, j \in [N], i < j$. As

[9]Indeed, for any $t, s \in [0, K]$, $|f(-e_i + te_j) - f(-e_i + se_j)| \leq \|te_j - se_j\| = L|t - s|$ and, for $\lambda \in [0, 1]$, $f(-e_i + (t\lambda + (1 - \lambda)s)e_j) = f(\lambda(-e_i + te_j) + (1 - \lambda)(-e_i + se_j)) \leq \lambda f(-e_i + te_j) + (1 - \lambda)f(-e_i + se_j)$.

the samples from different edges are independent, Example 1.4.6 in [201] and Lemma 3.2.4 [73] imply that

$$\sqrt{n}((\hat{\rho}_{0,12,n},\ldots,\hat{\rho}_{0,N(N-1),n})-(\rho_{0,12},\ldots,\rho_{0,N(N-1)}))\xrightarrow{d}(G_{\rho_{0,12}},\ldots,G_{\rho_{0,N(N-1)}}) \tag{3.57}$$

in $\prod_{\substack{i,j=1\\i<j}}^{N}\ell^\infty(\mathcal{F}([0,K],1))$, where, for $i,j\in[N], i<j$, $G_{\rho_{0,ij}}$ is a $\rho_{0,ij}$-Brownian bridge process in $\ell^\infty(\mathcal{F}([0,K],1))$ and all processes are independent. Observe that the map $\Upsilon:\prod_{\substack{i,j=1\\i<j}}^{N}\ell^\infty(\mathcal{F}([0,K],1))\to\ell^\infty(\mathcal{F}([-1,K]^N,1/2))$, defined via

$$\begin{aligned}\Upsilon:&(\eta_{12},\ldots,\eta_{N(N-1)})\\ &\mapsto\left(f\in\mathcal{F}([-1,K]^N,1/2)\mapsto\sum_{\substack{i,j=1\\i<j}}^{N}\eta_{ij}\left(f(-e_i+(\cdot)e_j)+f(-e_j+(\cdot)e_i)\right)\right),\end{aligned}$$

is continuous, so that, by (3.57), the representation of $(\mu_{0,n}-\mu_0)$ furnished in (3.56), and the continuous mapping theorem, we have that

$$\sqrt{n}(\mu_{0,n}-\mu_0)\xrightarrow{d}\frac{1}{(N-1)N(N+1)}\Upsilon\left(G_{\rho_{0,12}},\ldots,G_{\rho_{0,N(N-1)}}\right)\text{ in }\ell^\infty(\mathcal{F}([-1,K]^N,1/2)).$$

We highlight that, for any $f\in\ell^\infty(\mathcal{F}([-1,K]^N,1/2))$,

$$\Upsilon\left(G_{\rho_{0,12}},\ldots,G_{\rho_{0,N(N-1)}}\right)(f)=N\left(0,\frac{\sum_{\substack{i,j=1\\i<j}}^{N}\mathrm{Var}_{\rho_{0,ij}}(f(-e_i+(\cdot)e_j)+f(-e_j+(\cdot)e_i))}{(N-1)^2N^2(N+1)^2}\right) \tag{3.58}$$

in distribution since, for any $f\in\mathcal{F}([0,K],1)$, $G_{\rho_{0,ij}}(f)=N(0,\mathrm{Var}_{\rho_{0,ij}}(f))$ in distribution ($i,j\in[N], i<j$).

In the case that each of the $\rho_{0,ij}$ are finitely discrete, we let $(w_{ij}^{(k)})_{k=1}^{K_{ij}}$ denote the set of support points and let $(p_{ij}^{(k)})_{k=1}^{K_{ij}}$ be the corresponding probabilities $p_{ij}^{(k)}=\rho_{0,ij}(w_{ij}^{(k)})$ for $k\in[K_{ij}]$ and any $i,j\in[N]$ with $i<j$. Then, for any bounded measurable function g on

$\mathbb{R}$, $\mathrm{Var}_{\rho_{0,ij}}(g)$ is given by

$$\begin{pmatrix} g(w_{ij}^{(1)}) \\ \vdots \\ g(w_{ij}^{(K_{ij})}) \end{pmatrix}^\top \begin{pmatrix} p_{ij}^{(1)}(1-p_{ij}^{(1)}) & -p_{ij}^{(1)}p_{ij}^{(2)} & \dots & -p_{ij}^{(1)}p_{ij}^{(K_{ij})} \\ -p_{ij}^{(2)}p_{ij}^{(1)} & p_{ij}^{(2)}(1-p_{ij}^{(2)}) & \dots & -p_{ij}^{(2)}p_{ij}^{(K_{ij})} \\ \vdots & \vdots & \ddots & \vdots \\ -p_{ij}^{(K_{ij})}p_{ij}^{(1)} & -p_{ij}^{(K_{ij})}p_{ij}^{(2)} & \dots & p_{ij}^{(K_{ij})}(1-p_{ij}^{(K_{ij})}) \end{pmatrix} \begin{pmatrix} g(w_{ij}^{(1)}) \\ \vdots \\ g(w_{ij}^{(K_{ij})}) \end{pmatrix}$$

so that, if B_{ij} denotes the above multinomial covariance matrix, the variance in (3.58) can be written in terms of a block diagonal matrix

$$\check{f}^\top \mathbf{\Sigma}_Z \check{f} := \frac{1}{(N-1)^2N^2(N+1)^2}\check{f}^\top \begin{pmatrix} \boxed{B_{12}} & & & \\ & \boxed{B_{13}} & & \\ & & \ddots & \\ & & & \boxed{B_{N(N-1)}} \end{pmatrix} \check{f}, \tag{3.59}$$

where $\check{f} = (f_{12}, \dots, f_{N(N-1)})$, and

$$f_{ij} = \left(f(-e_i + w_{ij}^{(1)}e_j) + f(-e_j + w_{ij}^{(1)}e_i), \dots, f(-e_i + w_{ij}^{(K_{ij})}e_j) + f(-e_j + w_{ij}^{(K_{ij})}e_i)\right),$$

for $i, j \in [N]$ with $i < j$. The limit theorem then follows from Theorem 10.

It remains to show that Assumption 2 holds in this setting. The expression for $\chi_{\mu_0}(f_0)$ is given in (3.58) for $f_0 \in \{([-1, K]^N, 1/2)$, an analogous expression holds for $\chi_{\mu_1}(f_1)$ for f_1 in the same space. Now, assume that $g_1(\cdot - \mathbb{E}_{\mu_1}[X]) \in \{([-1, K]^N, 1/2)$ and let T be a Gromov-Monge map between $\bar{\mu}_0$ and $\bar{\mu}_1$. In particular, $T = S(\cdot + \mathbb{E}_{\mu_0}[X]) - \mathbb{E}_{\mu_1}[X]$ where S corresponds to a permutation of axes according to some permutation σ' as follows from Proposition 8. Then, letting σ denote the inverse of σ,

$$\begin{aligned} &\chi_{\mu_1}(g_1(\cdot - \mathbb{E}_{\mu_1}[X])) \\ &= N\left(0, \frac{\sum_{\substack{i,j=1\\ i<j}}^N \mathrm{Var}_{\rho_{1,ij}}(g_1(-e_i + (\cdot)e_j - \mathbb{E}_{\mu_1}[X]) + g_1(-e_j + (\cdot)e_i - \mathbb{E}_{\mu_1}[X]))}{(N-1)^2N^2(N+1)^2}\right), \\ &= N\left(0, \frac{\sum_{\substack{i,j=1\\ i<j}}^N \mathrm{Var}_{\rho_{0,\sigma(i)\sigma(j)}}(g_1(-e_i + (\cdot)e_j - \mathbb{E}_{\mu_1}[X]) + g_1(-e_j + (\cdot)e_i - \mathbb{E}_{\mu_1}[X]))}{(N-1)^2N^2(N+1)^2}\right). \end{aligned}$$

For any $t \in \mathbb{R}$, $-e_i + te_j - \mathbb{E}_{\mu_1}[X] = T(-e_{\sigma(i)} + t\sigma(j) - \mathbb{E}_{\mu_0}[X])$ so that

$$
\begin{aligned}
&\mathrm{Var}_{\rho_{0,\sigma(i)\sigma(j)}}(g_1(-e_i+(\cdot)e_j-\mathbb{E}_{\mu_1}[X])+g_1(-e_j+(\cdot)e_i-\mathbb{E}_{\mu_1}[X])) \\
&= \mathrm{Var}_{\rho_{0,\sigma(i)\sigma(j)}}(g_1(T(-e_{\sigma(i)} + (\cdot)e_{\sigma(j)} - \mathbb{E}_{\mu_0}[X])) + g_1(T(-e_{\sigma(j)} + (\cdot)e_{\sigma(i)} - \mathbb{E}_{\mu_0}[X]))),
\end{aligned}
$$

and

$$
\begin{aligned}
&\chi_{\mu_1}(g_1(\cdot - \mathbb{E}_{\mu_1}[X])) \\
&= N\left(0, \frac{\sum_{\substack{i,j=1\\ i<j}}^{N} \mathrm{Var}_{\rho_{0,ij}}(g_1(T(-e_i+(\cdot)e_j-\mathbb{E}_{\mu_0}[X]))+g_1(T(-e_j+(\cdot)e_i-\mathbb{E}_{\mu_0}[X])))}{(N-1)^2N^2(N+1)^2}\right), \\
&= \chi_{\mu_0}((g_1 \circ T)(\cdot - \mathbb{E}_{\mu_0}[X])),
\end{aligned}
$$

as desired.

Letting Σ_Z be the covariance matrix in (3.59), and Σ_n be the same covariance matrix with the edge distributions $\rho_{0,ij}$ replaced by empirical estimates thereof, it is easy to see that, conditionally on $G_{0,1}, \dots, G_{0,n}$ and $G_{1,1}, \dots, G_{1,n}$ $\mathbb{E}_{\mu_{0,n}}[X] \to \mathbb{E}_{\mu_0}[X], \Sigma_{\mu_{0,n}} \to \Sigma_{\mu_0}[X]$, and $\Sigma_n \to \Sigma_Z$ given almost every realization of the data. Moreover, it is easy to see that $\Sigma_Z \mathbb{1}_N = \Sigma_n \mathbb{1}_N = 0$ with probability 1 so that the required orthogonality condition holds. The fact that the proposed test is of asymptotic level α then follows form Theorem 18. Under the alternative, it is easy to see that the test statistic diverges to $+\infty$ so that the test is asymptotically consistent. $\square$

3.9 Concluding Remarks

This work initiated the study of distributional limits for the GW problem, covering both discrete and semi-discrete cases, as well as general 4-sub-Weibull distributions for the entropically regularized GW distance. To this end, we characterized the stability properties of the GW and entropic GW distances along perturbations of the marginal distributions, which may be of independent interest and offer utility beyond our statistical exploration. As the derived limits are generally non-Gaussian, we prove the consistency of a direct estimator of the limiting variable in the case of discrete distributions. Remarkably, we can efficiently sample from the direct estimator using linear programming, in spite of the fact that computing the empirical GW value is NP-complete in general. The developed limit distribution theory was leveraged for an application to graph isomorphism testing. To facilitate that, we proposed a framework for embedding distributions of random graphs into $\mathcal{P}(\mathbb{R}^N)$, i.e., distributions over a Euclidean space. This allows casting the graph isomorphism testing question as a test for equality of the embedded distributions under the GW distance. We then leverage the derived limit laws to furnish an asymptotically consistent test, and propose efficient procedures for (approximate) evaluation of the GW test statistic. Numerical results validating the test performance on several random graph models were provided, along with simulations of the derived limiting laws.

Future research directions stemming from this work are abundant. Firstly, although we did not provide limit distributions for the empirical GW distance between compactly supported distributions in low dimensions, the numerical experiments in Section 3.7.2 suggest that such a result may hold, at least under the null. Extending our current proof technique to the compactly supported case would require, among other results, proving that the OT potentials $(\varphi_{0,t}^{\mathbf{A}}, \varphi_{1,t}^{\mathbf{A}})$ for $\mathsf{OT}_{\mathbf{A}}((\cdot - \mathbb{E}_{\mu_0}[X])_\sharp \mu_{0,t}, (\cdot - \mathbb{E}_{\mu_1}[X])_\sharp \mu_{1,t})$ where $\mathbf{A} \in \operatorname{argmin}_{\mathbf{A} \in \mathbb{R}^{d_0 \times d_1}} \Phi_{(\bar{\mu}_0, \bar{\mu}_1)}$ converge pointwise as $t \downarrow 0$ to a pair of OT potentials for $\mathsf{OT}_{\mathbf{A}}(\bar{\mu}_0, \bar{\mu}_1)$ which maximizes $\Upsilon : (\psi_0, \psi_1) \mapsto \int \bar{\psi}_0 d(\nu_0 - \mu_0) + \int \bar{\psi}_1 d(\nu_1 - \mu_1)$ over all such OR potentials. While it is straightforward to show that the sequence of OT potentials converges along a subsequence to a pair of OT potentials for $\mathsf{OT}_{\mathbf{A}}(\bar{\mu}_0, \bar{\mu}_1)$, the main challenge lies in proving that this pair is maximal for Υ.

Another promising direction pertains to isometry testing of heterogeneous datasets using the GW distance. Our approach to testing for graph isomorphisms, presented in Section 3.6, relies on a particular embedding of random graph distributions into the space of distributions on a Euclidean space. The key property of this construction is that the embeddings are identified under the GW distance if and only if the distributions on graphs are isomorphic. This embedding, provided in (3.15), consists of two terms: the first compares the individual edge distributions, while the second ensures that the embeddings can only be related by specific isometries. This hints at a broader strategy for testing whether two datasets are generated from distributions that are related by particular isometric transformations, but not others. We plan to explore this direction further, aiming a general isometry testing methodology between heterogeneous datasets via the GW distance.

CHAPTER 4

ENTROPIC GROMOV-WASSERSTEIN DISTANCES: STABILITY AND ALGORITHMS

This chapter is based on the paper "Entropic Gromov-Wasserstein Distances: Stability and Algorithms" [164], published in the *Journal of Machine Learning Research*. This is joint work with Ziv Goldfeld and Kengo Kato.

4.1 Introduction

The Gromov-Wasserstein (GW) distance compares probability distributions that are supported on possibly distinct metric spaces by aligning them with one another. Given two metric measure (mm) spaces $(\mathcal{X}_0, \mathsf{d}_0, \mu_0)$ and $(\mathcal{X}_1, \mathsf{d}_1, \mu_1)$, the (p, q)-GW distance between them is

$$\mathsf{D}_{p,q}(\mu_0, \mu_1) := \inf_{\pi \in \Pi(\mu_0,\mu_1)} \left(\int_{\mathcal{X}_0 \times \mathcal{X}_1} \int_{\mathcal{X}_0 \times \mathcal{X}_1} \left| \mathsf{d}_0^q(x, x') - \mathsf{d}_1^q(y, y') \right|^p d\pi \otimes \pi(x, y, x', y') \right)^{\frac{1}{p}}, \quad (4.1)$$

where $\Pi(\mu_0, \mu_1)$ is the set of couplings between μ_0 and μ_1. This approach, proposed in [141], is an optimal transport (OT) based L^p-relaxation of the classical Gromov-Hausdorff distance between metric spaces. The GW distance defines a metric on the space of all mm spaces modulo measure preserving isometries.[1] From an applied standpoint, a solution to the GW problem between two heterogeneous datasets yields not only a quantification of discrepancy, but also an optimal alignment $\pi^\star$ between them. As such, alignment methods inspired by the GW problem have been proposed for many applications, encompassing single-cell genomics [23, 67], alignment of language models [5, 108], shape matching [140, 125], graph matching [217, 216], heterogeneous

[1] Two mm spaces $(\mathcal{X}_0, \mathsf{d}_0, \mu_0)$ and $(\mathcal{X}_1, \mathsf{d}_1, \mu_1)$ are isomorphic if there exists an isometry $T : \mathrm{spt}(\mu_0) \to \mathrm{spt}(\mu_1)$ for which $\mu_0 \circ T^{-1} = \mu_1$ as measures. The quotient space is then the one induced by this equivalence relation.

domain adaptation [219], and generative modeling [32].

Exact computation of the GW distance is a quadratic assignment problem, which is known to be NP-complete [51]. To remedy this, various computationally tractable reformulations of the distance have been proposed. We postpone full discussion of such methods to Section 4.1.2 and focus here on the entropic GW (EGW) distance [160, 186]

$$\mathsf{S}_{p,q}^{\varepsilon}(\mu_0,\mu_1) := \inf_{\pi\in\Pi(\mu_0,\mu_1)} \iint \left|\mathsf{d}_0^q(x,x') - \mathsf{d}_1^q(y,y')\right|^p d\pi\otimes\pi(x,y,x',y') + \varepsilon \mathsf{D}_{\mathsf{KL}}(\pi\|\mu_0\otimes\mu_1),$$

which is at the center of this work. Entropic regularization by means of the Kullback-Leibler (KL) divergence above transforms the linear Kantorovich OT problem [120] into a strictly convex one and enables directly solving it using Sinkhorn's algorithm [53]. Although the EGW problem is, in general, not convex, [186] propose to solve it via an iterative approach with Sinkhorn iterations. This method is known to converge to a stationary point of a tight (albeit non-convex) relaxation of the EGW problem, but this is an asymptotic statement and the overall computational complexity is unknown. Similar limitations apply for the popular mirror descent-based approach from [160]. To the best of our knowledge, there is currently no known algorithm for computing the EGW distance subject to non-asymptotic convergence rate bounds, let alone global optimality guarantees. Further, these methods do not account for the error incurred by using Sinkhorn iterations, nor do they address the behaviour of the solutions they obtain in the limit of vanishing regularization.

The goal of this work is to close the aforementioned computational gaps, targeting algorithms with non-asymptotic guarantees, accounting for inexactness in Sinkhorn's algorithm, characterizing convexity regimes of the EGW problem, and establishing convergence of stationary points to the EGW problem to stationary points of the GW problem as $\varepsilon \downarrow 0$. All of these will be achieved as a consequence of a new stability analysis of the EGW variational representation from [223].

4.1.1 Contributions

Theorem 1 in [223] shows that the EGW distance with quadratic cost between the Euclidean mm spaces $(\mathbb{R}^{d_0}, \|\cdot\|, \mu_0)$ and $(\mathbb{R}^{d_1}, \|\cdot\|, \mu_1)$ can be written as:

$$\mathsf{S}_{2,2}^{\varepsilon}(\mu_0, \mu_1) = C_{\mu_0,\mu_1} + \inf_{\boldsymbol{A}\in\mathcal{D}_M} \{32\|\boldsymbol{A}\|_F^2 + \mathsf{OT}_{\boldsymbol{A},\varepsilon}(\mu_0, \mu_1)\}, \tag{4.2}$$

where C_{μ_0,μ_1} is a constant that depends only on the moments of the marginals, $\mathcal{D}_M \subset \mathbb{R}^{d_0\times d_1}$ is a compact rectangle, and $\mathsf{OT}_{\boldsymbol{A},\varepsilon}(\mu_0, \mu_1)$ is an EOT problem with a particular cost function that depends on the auxiliary variable $\boldsymbol{A}$. This representation connects EGW to the well-understood EOT problem, thereby unlocking powerful tools for analysis. To exploit this connection for new computational advancements, we begin by analyzing the stability of the objective in (4.2) in $\boldsymbol{A}$ and derive its first and second-order Fréchet derivatives. This, in turn, enables us to derive its convexity and smoothness properties and devise new algorithms for solving the EGW problem, subject to formal convergence guarantees.

The Fréchet derivatives of the objective function from (4.2) in $\boldsymbol{A}$ reveal that this problem falls under the paradigm of smooth constrained optimization. Indeed, the derivatives imply, respectively, upper and lower bounds on the top and bottom eigenvalues of the Hessian matrix of the objective; L-smoothness then follows from the mean value inequality. By further requiring positive semidefiniteness of the Hessian we obtain a sharp and primitive sufficient condition on ε under which (4.2) becomes convex. These regularity properties are used to lift the accelerated first-order methods for smooth (non-convex) optimization [89] and convex programming with an inexact oracle [55] to the EGW problem. This yields the first algorithms with non-asymptotic convergence guarantees toward global or local EGW solutions, depending on whether the problem is convex or not (e.g., under the aforementioned condition). Our algorithms compute not only the EGW cost, but also provide an approximate optimizer—namely a coupling, which serves as the alignment scheme that achieves the said cost.

Specifically, our method iteratively solves the optimization problem in the space of auxiliary matrices $\boldsymbol{A} \in \mathbb{R}^{d_0 \times d_1}$, with each iterate calling the Sinkhorn algorithm to obtain an approximate solution (viz. the inexact oracle) to the corresponding EOT problem. The time complexity of the Sinkhorn algorithm governs the overall runtime, which is therefore $O(N^2)$, for μ_0 and μ_1 as distributions on N points. This presents a significant speedup to the $O(N^3)$ runtime of popular iterative algorithms from [160, 186]. Under certain low-rank assumptions on the cost matrix, [177] recently showed the mirror descent approach from [160] can be sped up to run in $O(N^2)$ time, which is comparable to our method. Nevertheless, our algorithms are coupled with formal convergence guarantees, non-asymptotic error bounds, and global optimality claims under the said convexity condition, while no such assurances are available for other methods.

As the derivative of the objective from (4.2) depends on the optimal coupling for a particular EOT problem, we also account for the error incurred by solving this problem numerically (e.g. via Sinkhorn iterations). In particular, we characterize how well the output of a standard implementation of Sinkhorn's algorithm approximates the true EOT coupling. This effect was not analyzed before, as existing literature focused on approximating the unregularized OT cost, while treating the KL divergence term as a bias.

The EGW distance serves as a proxy to unregularized GW, which renders the vanishing regularization parameter regime, namely $\varepsilon \downarrow 0$, of central importance. We show that stationary points of the variational EGW problem converge (possibly along a subsequence) to a stationary point of the variational GW problem as $\varepsilon \downarrow 0$. Only convergence to a stationary point can be guaranteed in the limit since the underlying variational problem may fail to be convex when ε is small. Nonetheless, this is the first result that provides stationarity guarantees for limiting solutions, which clarifies how the local solutions obtained using our algorithm approximate local solutions to the unregularized problem.

4.1.2 Literature review

The computational intractability of the GW problem in (4.1) has inspired several reformulations that aim to alleviate this issue. The sliced GW distance [206] attempts to reduce the computational burden by considering the average of GW distances between one-dimensional projections of the marginals. However, unlike OT in one dimension, the GW problem does not have a known simple solution even in one dimension [14]. Another approach is to relax the strict marginal constraints by optimizing over the weights of one of the marginals as in semi-relaxed GW [209] or by using f-divergence penalties; this leads to the unbalanced GW distance [180], which lends itself well for convex/conic relaxations. A variant that directly optimizes over bi-directional Monge maps between the mm spaces was considered in [224]. The fused GW distance [205] enables comparing both feature and structural properties of structured data. Although most of these relaxations can be shown to converge to the GW distance (in terms of function values) under certain regimes, they involve solving non-convex problems, which limits their utility for numerical resolution of the GW problem. The recent work of [41] proposes a semidefinite relaxation of the GW problem along with a certificate of optimality which, upon obtaining a solution to the relaxed problem, establishes if it is optimal for the original problem.

While these methods offer certain advantages, it is the approach based on entropic regularization [160, 186] that is most frequently used in application. A low-rank variant of the EGW problem was proposed in [177], where the distance distortion cost is only optimized over coupling admitting a certain low-rank structure. They arrive at a linear time algorithm for this problem by adapting the mirror descent method of [160]. As an intermediate step of their analysis, they show that if μ_0 and μ_1 are supported on N distinct points, then the $O(N^3)$ complexity of mirror descent (see, e.g., Remark 1 in [160]) can be reduced to $O(N^2)$ by assuming that the matrices of pairwise costs admit a low-rank decomposition (without imposing any structure on the couplings). This decomposition holds, for instance, when the cost is the squared Euclidean distance and the sample size dominates the ambient dimension. Although mirror descent seems to solve the EGW problem quite well in practice, formal guarantees concerning convergence rates or local optimality are lacking.

Other related work explores structural properties of GW distances, on which some of our findings also reflect. The existence of Monge maps for the GW problem was studied in [75] and they show that optimal couplings are induced by a bimap (viz. two-way map) under general conditions. [66] focused on the GW distance between Gaussian distributions, deriving upper and lower bounds on the optimal cost. A closed-form expression under the Gaussian setting was derived in [127] for the EGW distance with inner product cost.

As we establish stability results for the EGW problem by utilizing its connection to the EOT problem with a parametrized cost, we mention that different notions of stability for EOT have been studied. For instance, [90] concerns stability of the EOT cost for varying marginals, cost function, and regularization parameter. The related works [38, 78, 157] concern stability of the EOT cost and related objects for weakly convergent marginal distributions.

4.1.3 Organization

This paper is organized as follows. In Section 4.2 we compile background material on EOT and the EGW problems. In Section 4.3, we describe the smoothness and convexity of the variational EGW problem. In Section 4.4, we analyze and test two algorithms for solving the EGW problem. We compile the proofs for Sections 4.3 and 4.4 in Section 4.5. Section 4.6 contains some concluding remarks.

4.1.4 Notation

Denote by $\mathcal{P}(\mathbb{R}^d)$ the collection of all Borel probability measures on $\mathbb{R}^d$, and by $\mathcal{P}_p(\mathbb{R}^d)$ the set of all $\mu \in \mathcal{P}(\mathbb{R}^d)$ with finite p-th moment ($p > 0$). The pushforward of $\mu \in \mathcal{P}(\mathbb{R}^{d_0})$ through a measurable map $T : \mathbb{R}^{d_0} \to \mathbb{R}^{d_1}$ is denoted by $T_\sharp\mu := \mu \circ T^{-1}$. The Frobenius inner product on $\mathbb{R}^{d_0 \times d_1}$ is defined by $\langle \boldsymbol{A}, \boldsymbol{B} \rangle_F = \mathrm{tr}\,(\boldsymbol{A}^\intercal \boldsymbol{B})$; the associated norm is denoted by $\|\cdot\|_F$. For a nonemtpy set $S \subset \mathbb{R}^d$, $C(S)$ is the set of continuous functions on S. We adopt the shorthands $a \wedge b = \min\{a, b\}$ and $a \vee b = \max\{a, b\}$.

4.2 Background and Preliminaries

We first establish notation and review standard definitions and results underpinning our analysis of the EGW distance.

4.2.1 Entropic Optimal Transport

Entropic regularization transforms the linear OT problem into a strictly convex one. Given distributions $\mu_i \in \mathcal{P}(\mathbb{R}^{d_i})$, $i = 0, 1$, and a Borel cost function $c : \mathbb{R}^{d_0} \times \mathbb{R}^{d_1} \to \mathbb{R}$ that is bounded from below on $\mathrm{spt}(\mu_0) \times \mathrm{spt}(\mu_1)$, the primal EOT problem is obtained by regularizing the standard OT problem via the Kullback-Leibler (KL) divergence,

$$\mathsf{OT}_\varepsilon(\mu_0, \mu_1) = \inf_{\pi \in \Pi(\mu_0,\mu_1)} \int c\, d\pi + \varepsilon \mathsf{D}_{\mathsf{KL}}(\pi \| \mu_0 \otimes \mu_1),$$

where $\varepsilon > 0$ is a regularization parameter and

$$\mathsf{D}_{\mathsf{KL}}(\mu_0 \| \mu_1) = \begin{cases} \int \log\left(\frac{d\mu_0}{d\mu_1}\right) d\mu_0, & \text{if } \mu_0 \ll \mu_1, \\ +\infty, & \text{otherwise.} \end{cases}$$

Classical OT is obtained from the above by setting $\varepsilon = 0$. When $c \in L^1(\mu_0 \otimes \mu_1)$, EOT admits the following dual formulation,

$$\mathsf{OT}_\varepsilon(\mu_0, \mu_1) = \sup_{(\varphi_0,\varphi_1) \in L^1(\mu_0) \times L^1(\mu_1)} \int \varphi_0 d\mu_0 + \int \varphi_1 d\mu_1 - \varepsilon \int e^{\frac{\varphi_0 \oplus \varphi_1 - c}{\varepsilon}} d\mu_0 \otimes \mu_1 + \varepsilon,$$

where $\varphi_0 \oplus \varphi_1(x, y) = \varphi_0(x) + \varphi_1(y)$. For $\varepsilon > 0$, the set of solutions to the dual problem coincides with the set of solutions to the so-called Schrödinger system,

$$\begin{aligned} \int e^{\frac{\varphi_0(x)+\varphi_1(\cdot)-c(x,\cdot)}{\varepsilon}} d\mu_1 &= 1, \quad \mu_0\text{-a.e. } x \in \mathbb{R}^{d_0}, \\ \int e^{\frac{\varphi_0(\cdot)+\varphi_1(y)-c(\cdot,y)}{\varepsilon}} d\mu_0 &= 1, \quad \mu_1\text{-a.e. } y \in \mathbb{R}^{d_1}, \end{aligned} \tag{4.3}$$

for $(\varphi_0, \varphi_1) \in L^1(\mu_0) \times L^1(\mu_1)$. A pair $(\varphi_0, \varphi_1) \in L^1(\mu_0) \times L^1(\mu_1)$ solving (4.3) is known to be a.s. unique up to additive constants in the sense that any other solution $(\bar\varphi_0, \bar\varphi_1)$ satisfies $\bar\varphi_0 = \varphi_0 + a$ μ_0-a.s. and $\bar\varphi_1 = \varphi_1 - a$ μ_1-a.s. for some $a \in \mathbb{R}$. Moreover, the unique EOT coupling π_ε is characterized by

$$\frac{d\pi_\varepsilon}{d\mu_0 \otimes \mu_1}(x, y) = e^{\frac{\varphi_0(x)+\varphi_1(y)-c(x,y)}{\varepsilon}}, \tag{4.4}$$

and, under some additional conditions on the cost and marginals which hold throughout this paper, (4.3) admits a pair of continuous solutions which is unique up to additive constants and satisfies the system at all points $(x, y) \in \mathbb{R}^{d_0} \times \mathbb{R}^{d_1}$. We call such continuous solutions EOT potentials. The reader is referred to [155] for a comprehensive overview of EOT.

4.2.2 Entropic Gromov-Wasserstein Distance

This work studies stability and computational aspects of the entropically regularized GW distance under the quadratic cost. By analogy to OT, EGW serves as a proxy of the standard (p, q)-GW distance, which quantifies discrepancy between complete and separable mm spaces $(\mathcal{X}_0, \mathsf{d}_0, \mu_0)$ and $(\mathcal{X}_1, \mathsf{d}_1, \mu_1)$ as [141, 189]

$$\mathsf{D}_{p,q}(\mu_0, \mu_1) := \inf_{\pi \in \Pi(\mu_0, \mu_1)} \|\Gamma_q\|_{L^p(\pi \otimes \pi)},$$

where $\Gamma_q(x, y, x', y') = \left|\mathsf{d}_0^q(x, x') - \mathsf{d}_1^q(y, y')\right|$ is the distance distortion cost. This definition is the L^p-relaxation of the Gromov-Hausdorff distance between metric spaces,[2] and gives rise to a metric on the collection of all isomorphism classes of mm spaces with finite pq-size, i.e., such that $\int \mathsf{d}(x, x')^{pq} \, d\mu \otimes \mu(x, x') < \infty$.

From here on out, we instantiate the mm spaces as the Euclidean spaces $(\mathbb{R}^{d_i}, \|\cdot\|, \mu_i)$, for $i = 0, 1$, and focus on the EGW distance for the quadratic cost.

[2]The Gromov-Hausdorff distance between $(\mathcal{X}_0, \mathsf{d}_0)$ and $(\mathcal{X}_1, \mathsf{d}_1)$ is given by $\frac{1}{2} \inf_{R \in \mathcal{R}(\mathcal{X}_0, \mathcal{X}_1)} \|\Gamma_1\|_{L^\infty(R)}$, where $\mathcal{R}(\mathcal{X}_0, \mathcal{X}_1)$ is the collection of all correspondence sets of $\mathcal{X}_0$ and $\mathcal{X}_1$, i.e., subsets $R \subset \mathcal{X}_0 \times \mathcal{X}_1$ such that the coordinate projection maps are surjective when restricted to R. The correspondence set can be thought of as $\mathrm{spt}(\pi)$ in the GW formulation.

Quadratic Cost The quadratic EGW distance, which corresponds to the $p = q = 2$ case, is defined as

$$\mathsf{S}_\varepsilon(\mu_0, \mu_1) = \inf_{\pi\in\Pi(\mu_0,\mu_1)} \int \left| \|x - x'\|^2 - \|y - y'\|^2 \right|^2 d\pi\otimes\pi(x, y, x', y') + \varepsilon \mathsf{D}_{\mathsf{KL}}(\pi\|\mu_0\otimes\mu_1). \quad (4.5)$$

One readily verifies that, like the standard GW distance, EGW is invariant to isometric actions on the marginal spaces such as orthogonal rotations and translations. In addition, note that $\mathsf{S}_\varepsilon(\mu_0, \mu_1) = \varepsilon \mathsf{S}_1(\mu_0^\varepsilon, \mu_1^\varepsilon)$, where $\mu_i^\varepsilon = (\varepsilon^{-1/4}\,\mathrm{Id})_\sharp \mu_i$. In general, (4.5) is a non-convex quadratic program. Non-convexity can easily be discerned from the representation (4.6).

When μ_0, μ_1 are centered, which we may assume without loss of generality, the EGW distance decomposes[3] as (cf. Section 5.3 in [223])

$$\mathsf{S}_\varepsilon(\mu_0, \mu_1) = \mathsf{S}^1(\mu_0, \mu_1) + \mathsf{S}^2_\varepsilon(\mu_0, \mu_1), \quad (4.6)$$

where

$$\mathsf{S}^1(\mu_0, \mu_1) = \int \|x - x'\|^4 d\mu_0\otimes\mu_0(x, x') + \int \|y - y'\|^4 d\mu_1\otimes\mu_1(y, y') - 4\int \|x\|^2\|y\|^2 d\mu_0\otimes\mu_1(x, y),$$

$$\mathsf{S}^2_\varepsilon(\mu_0, \mu_1) = \inf_{\pi\in\Pi(\mu_0,\mu_1)} \int -4\|x\|^2\|y\|^2 d\pi(x, y) - 8 \sum_{\substack{1\le i\le d_0\\ 1\le j\le d_1}} \left(\int x_i y_j \, d\pi(x, y) \right)^2 + \varepsilon \mathsf{D}_{\mathsf{KL}}(\pi\|\mu_0\otimes\mu_1).$$

Evidently, S^1 depends only on the moments of the marginal distributions μ_0, μ_1, while S^2_ε captures the dependence on the coupling. A key observation in [223] is that S^2_ε admits a variational form that ties it to the well understood EOT problem.

Lemma 32 (EGW duality; Theorem 1 in [223])**.** Fix $\varepsilon > 0$, $(\mu_0, \mu_1) \in \mathcal{P}_4(\mathbb{R}^{d_0}) \times \mathcal{P}_4(\mathbb{R}^{d_1})$, and let $M_{\mu_0,\mu_1} := \sqrt{M_2(\mu_0)M_2(\mu_1)}$. Then, for any $M \ge M_{\mu_0,\mu_1}$,

$$\mathsf{S}^2_\varepsilon(\mu_0, \mu_1) = \inf_{\boldsymbol{A}\in\mathcal{D}_M} 32\|\boldsymbol{A}\|_F^2 + \mathsf{OT}_{\boldsymbol{A},\varepsilon}(\mu_0, \mu_1), \quad (4.7)$$

[3]A similar decomposition holds for the inner product cost, obtained by replacing the squared Euclidean distances in Equation (4.5) by inner products. As such, all results derived in this manuscript apply to the inner product cost with minor modifications. See Section C.5 for additional details.

where $\mathcal{D}_M := \{\boldsymbol{A} \in \mathbb{R}^{d_0 \times d_1} : \|\boldsymbol{A}\|_F \le M/2\}$ and $\mathsf{OT}_{\boldsymbol{A},\varepsilon}(\mu_0, \mu_1)$ is the EOT problem with the cost function $c_{\boldsymbol{A}} : (x, y) \in \mathbb{R}^{d_0} \times \mathbb{R}^{d_1} \mapsto -4\|x\|^2\|y\|^2 - 32x^\intercal \boldsymbol{A} y$ and regularization parameter ε. Moreover, the infimum is achieved at some $\boldsymbol{A}^\star \in \mathcal{D}_{M_{\mu_0,\mu_1}}$.

An analogous result holds in the unregularized ($\varepsilon = 0$) case, see Corollary 1 in [223]. The proof of Theorem 1 in [223] demonstrates that if μ_0 and μ_1 are centered and $\pi_\star$ is optimal for the original EGW formulation, then $\boldsymbol{A}^\star = \frac{1}{2}\int xy^\intercal \, d\pi_\star(x, y)$ is optimal for (4.7) and $\pi_\star = \pi_{\boldsymbol{A}^\star}$, where $\pi_{\boldsymbol{A}^\star}$ is the unique EOT coupling for $\mathsf{OT}_{\boldsymbol{A}^\star,\varepsilon}(\mu_0, \mu_1)$. It follows from Jensen's inequality and the Cauchy-Schwarz inequality that $\boldsymbol{A}^\star \in \mathcal{D}_M$. Corollary 6 ahead expands on this connection by establishing a one-to-one correspondence between solutions of S_ε and S^2_ε and shows that all solutions of (4.7) lie in $\mathcal{D}_M$.

Although (4.7) illustrates a connection between the EGW and EOT problems, the outer minimization over $\mathcal{D}_M$ necessitates studying EOT with a parametrized cost function $c_{\boldsymbol{A}}$.

4.3 Stability of Entropic Gromov-Wasserstein Distances

We analyze the stability of the EGW problems with respect to (w.r.t.) the matrix $\boldsymbol{A}$ appearing in the dual formulation (4.7). Specifically, we characterize the first and second derivatives of the objective function whose optimization defines S^2_ε. These, in turn, elucidates its smoothness and convexity properties. Our stability analysis is later used to (i) gain insight into the structure of optimal couplings for the EGW problem; and (ii) devise novel approaches for computing the EGW distance with formal convergence guarantees.

Throughout this section, we restrict to compactly supported distributions, as some of the technical details do not directly translate to the unbounded setting (e.g., the proof of Lemma 34). For a Fréchet differentiable map $F : U \to V$ between normed vector spaces U and V,[4] we denote the derivative of F at the point $u \in U$ evaluated at $v \in V$ by $DF_{[u]}(v)$.

Fix compactly supported distributions $(\mu_0, \mu_1) \in \mathcal{P}(\mathbb{R}^{d_0}) \times \mathcal{P}(\mathbb{R}^{d_1})$ and some $\varepsilon > 0$. Let

$$\Phi : \boldsymbol{A} \in \mathbb{R}^{d_0 \times d_1} \mapsto 32\|\boldsymbol{A}\|_F^2 + \mathsf{OT}_{\boldsymbol{A},\varepsilon}(\mu_0, \mu_1)$$

denote the objective in (4.7). We first characterize the derivatives of Φ and then prove that this map is weakly convex and L-smooth.[5]

Proposition 10 (First and second derivatives)**.** The map $\Phi : \boldsymbol{A} \in \mathbb{R}^{d_0 \times d_1} \mapsto 32\|\boldsymbol{A}\|_F^2 + \mathsf{OT}_{\boldsymbol{A},\varepsilon}(\mu_0, \mu_1)$ is smooth, coercive, and has first and second-order Fréchet derivatives at $\boldsymbol{A} \in \mathbb{R}^{d_0 \times d_1}$ given by

$$\begin{aligned} D\Phi_{[\boldsymbol{A}]}(\boldsymbol{B}) &= 64 \operatorname{tr}(\boldsymbol{A}^\intercal \boldsymbol{B}) - 32 \int x^\intercal \boldsymbol{B} y \, d\pi_{\boldsymbol{A}}(x, y), \\ D^2\Phi_{[\boldsymbol{A}]}(\boldsymbol{B}, \boldsymbol{C}) &= 64 \operatorname{tr}(\boldsymbol{B}^\intercal \boldsymbol{C}) + 32\varepsilon^{-1} \int x^\intercal \boldsymbol{B} y \left(h_0^{\boldsymbol{A},\boldsymbol{C}}(x) + h_1^{\boldsymbol{A},\boldsymbol{C}}(y) - 32 x^\intercal \boldsymbol{C} y \right) d\pi_{\boldsymbol{A}}(x, y), \end{aligned}$$

where $\boldsymbol{B}, \boldsymbol{C} \in \mathbb{R}^{d_0 \times d_1}$, $\pi_{\boldsymbol{A}}$ is the unique EOT coupling for $\mathsf{OT}_{\boldsymbol{A},\varepsilon}(\mu_0, \mu_1)$, and $(h_0^{\boldsymbol{A},\boldsymbol{C}}, h_1^{\boldsymbol{A},\boldsymbol{C}})$ is the unique (up to additive constants) pair of functions in $C(\operatorname{spt}(\mu_0)) \times C(\operatorname{spt}(\mu_1))$ satisfying

$$\begin{aligned} \int \left(h_0^{\boldsymbol{A},\boldsymbol{C}}(x) + h_1^{\boldsymbol{A},\boldsymbol{C}}(y) - 32 x^\intercal \boldsymbol{C} y \right) e^{\frac{\varphi_0^{\boldsymbol{A}}(x) + \varphi_1^{\boldsymbol{A}}(y) - c_{\boldsymbol{A}}(x,y)}{\varepsilon}} d\mu_1(y) = 0, \quad \forall\, x \in \operatorname{spt}(\mu_0), \\ \int \left(h_0^{\boldsymbol{A},\boldsymbol{C}}(x) + h_1^{\boldsymbol{A},\boldsymbol{C}}(y) - 32 x^\intercal \boldsymbol{C} y \right) e^{\frac{\varphi_0^{\boldsymbol{A}}(x) + \varphi_1^{\boldsymbol{A}}(y) - c_{\boldsymbol{A}}(x,y)}{\varepsilon}} d\mu_0(x) = 0, \quad \forall\, y \in \operatorname{spt}(\mu_1). \end{aligned} \tag{4.8}$$

Here, $(\varphi_0^{\boldsymbol{A}}, \varphi_1^{\boldsymbol{A}})$ is any pair of EOT potentials for $\mathsf{OT}_{\boldsymbol{A},\varepsilon}(\mu_0, \mu_1)$.

Proposition 10 essentially follows from the implicit mapping theorem, which enables us to compute the Fréchet derivative of the EOT potentials for $\mathsf{OT}_{(\cdot),\varepsilon}(\mu_0, \mu_1)$ using the

[4]A map $F : U \to V$ is Fréchet differentiable at $u \in U$ if there exists a bounded linear operator $A : U \to V$ for which $F(u + h) = F(u) + Ah + o(h)$ as $h \to 0$. If such an A exists, it is called the Fréchet derivative to F at u.

[5]A function $f : \mathbb{R}^d \to \mathbb{R}$ is ρ-weakly convex if $f + \frac{\rho}{2}\| \cdot \|^2$ is convex; f is L-smooth if its gradient is L-Lipschitz, i.e., $\|\nabla f(x) - \nabla f(y)\| \le L\|x - y\|$, for all $x, y \in \mathbb{R}^d$.

Schrödinger system (4.3). The derivative of $\mathsf{OT}_{(\cdot),\varepsilon}(\mu_0,\mu_1)$, which is a simple function of the EOT potentials, is then readily obtained. By differentiating the Frobenius norm, this further yields the derivative of Φ. See Section 4.5.1 for details.

The following remark clarifies that the first and second Fréchet derivatives of Φ can be identified with the gradient and Hessian of a function on $\mathbb{R}^{d_0 d_1}$. Recall that the Frobenius inner product, $\langle \boldsymbol{A}, \boldsymbol{B}\rangle_F = \operatorname{tr}(\boldsymbol{A}^\intercal \boldsymbol{B}) = \sum_{\substack{1\le i\le d_0\\ 1\le j\le d_1}} \boldsymbol{A}_{ij}\boldsymbol{B}_{ij}$, is simply the Euclidean inner product between the vectorized matrices $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{d_0\times d_1}$.

Remark 20 (Interpreting derivatives as gradient/Hessian)**.** The first derivative of Φ from Proposition 10 can be written as

$$D\Phi_{[\boldsymbol{A}]}(\boldsymbol{B}) = \left\langle 64\boldsymbol{A} - 32\int xy^\intercal \, d\pi_{\boldsymbol{A}}(x,y), \boldsymbol{B}\right\rangle_F.$$

Recall that if f is a continuously differentiable function f on $\mathbb{R}^d$, its directional derivative at x along the direction y is $Df_{[x]}(y) = \langle \nabla f(x), y\rangle$, for $x, y \in \mathbb{R}^d$. By analogy, we may think of $D\Phi_{[\boldsymbol{A}]}$ as $64\boldsymbol{A} - 32\int xy^\intercal \, d\pi_{\boldsymbol{A}}(x,y)$. This perspective is utilized in Section 4.4 when studying computational guarantees for the EGW distance, as it is simpler to view iterates as matrices rather than abstract linear operators. By the same token, the second derivative of Φ at $\boldsymbol{A} \in \mathbb{R}^{d_0\times d_1}$ is a bilinear form on $\mathbb{R}^{d_0\times d_1}$ and hence can be identified with a $d_0 d_1 \times d_0 d_1$ matrix by analogy with the Hessian.

As a direct corollary to Proposition 10, we provide an (implicit) characterization of the stationary points of Φ and connect its minimizers to solutions of S_ε. Details are provided in Section 4.5.2.

Corollary 6 (Stationary points and correspondence between S_ε and S_ε^2)**.**

(i) A matrix $\boldsymbol{A} \in \mathbb{R}^{d_0\times d_1}$ is a stationary point of Φ if and only if $\boldsymbol{A} = \frac{1}{2}\int xy^\intercal \, d\pi_{\boldsymbol{A}}(x,y)$. As Φ is coercive, all minimizers of Φ are stationary points and hence contained in $\mathcal{D}_{M_{\mu_0,\mu_1}}$.

(ii) If μ_0 and μ_1 are centered, then a given matrix $\boldsymbol{A}$ minimizes Φ if and only if $\pi_{\boldsymbol{A}}$ is optimal for S_ε and satisfies $\frac{1}{2}\int xy^\intercal \, d\pi_{\boldsymbol{A}}(x,y) = \boldsymbol{A}$.

(iii) Suppose μ_0 and μ_1 are centered. If S_ε admits a unique optimal coupling $\pi_\star$, then Φ admits a unique minimizer $\boldsymbol{A}^\star$ and $\pi_\star = \pi_{\boldsymbol{A}^\star}$. Conversely, if Φ admits a unique minimizer $\boldsymbol{A}^\star$, then $\pi_{\boldsymbol{A}^\star}$ is a unique optimal coupling for S_ε.

Although the second derivative of Φ involves the implicitly defined functions $(h_0^{\boldsymbol{A},\boldsymbol{C}}, h_1^{\boldsymbol{A},\boldsymbol{C}})$, its maximal and minimal eigenvalues admit the following explicit bounds.

Corollary 7 (Hessian eigenvalue bounds)**.** The following hold for any $\boldsymbol{A} \in \mathbb{R}^{d_0\times d_1}$:

(i) The minimal eigenvalue of $D^2\Phi_{[\boldsymbol{A}]}$ satisfies

$$\begin{aligned}\lambda_{\min}\left(D^2\Phi_{[\boldsymbol{A}]}\right) &= 64 + \varepsilon^{-1} \inf_{\|\boldsymbol{C}\|_F=1} \int \left[\left(h_0^{\boldsymbol{A},\boldsymbol{C}}(x) + h_1^{\boldsymbol{A},\boldsymbol{C}}(y)\right)^2 - 32^2(x^\intercal \boldsymbol{C} y)^2\right] d\pi_{\boldsymbol{A}}(x,y) \\ &\geq 64 - 32^2\varepsilon^{-1} \sup_{\|\boldsymbol{C}\|_F=1} \operatorname{Var}_{\pi_{\boldsymbol{A}}}(X^\intercal \boldsymbol{C} Y),\end{aligned}$$

where the variance admits the uniform bound $\sup_{\|\boldsymbol{C}\|_F=1} \operatorname{Var}_{\pi_{\boldsymbol{A}}}(X^\intercal \boldsymbol{C} Y) \leq \sqrt{M_4(\mu_0)M_4(\mu_1)}$.

(ii) The maximal eigenvalue of $D^2\Phi_{[\boldsymbol{A}]}$ satisfies $\lambda_{\max}\left(D^2\Phi_{[\boldsymbol{A}]}\right) \leq 64$.

Corollary 7 follows from Proposition 10 by considering the variational form of the maximal and minimal eigenvalues; see Section 4.5.3 for details. We note that, in general, the variance bound in Item (i) is sharp up to constants in arbitrary dimensions. For example, it is attained up to a factor of 2 by $\mu_0 = \frac{1}{2}(\delta_0 + \delta_a)$ and $\mu_1 = \frac{1}{2}(\delta_0 + \delta_b)$ for $a \in \mathbb{R}^{d_0}$ and $b \in \mathbb{R}^{d_1}$; see Section C.1 for the full variance computation.

Armed with the eigenvalue bounds, we now state the main result of this section addressing convexity and smoothness of Φ.

Theorem 19 (Convexity and L-smoothness)**.** The map Φ is weakly convex with parameter at most $32^2\varepsilon^{-1}\sqrt{M_4(\mu_0)M_4(\mu_1)}-64$ and, if $\sqrt{M_4(\mu_0)M_4(\mu_1)} < \frac{\varepsilon}{16}$, then it is strictly convex and admits a unique minimizer. Moreover, for any $M > 0$, Φ is L-smooth on $\mathcal{D}_M$ with

$$L = \max_{\boldsymbol{A}\in\mathcal{D}_M} \lambda_{\max}\left(D^2\Phi_{[\boldsymbol{A}]}\right) \vee \left(- \lambda_{\min}\left(D^2\Phi_{[\boldsymbol{A}]}\right)\right) \le 64 \vee \left(32^2\varepsilon^{-1}\sqrt{M_4(\mu_0)M_4(\mu_1)} - 64\right).$$

Theorem 19 shows that Φ is amenable to optimization by accelerated gradient methods with step size L and establishes sufficient conditions to guarantee convergence of these algorithms to a global minimizer (i.e., convexity of Φ).

In general, optimal EGW couplings may not be unique. Theorem 19 provides sufficient conditions for uniqueness of solutions to both (4.7) and the EGW problem by the connection discussed in Corollary 6 when the marginals are centered. When the optimal EGW coupling is unique, symmetries in the marginal spaces result in certain structural properties for the optimal $\boldsymbol{A}^\star$ in (4.7). The following remark expands on these connections.

Remark 21 (Symmetries and uniqueness of couplings)**.** Fix $\varepsilon > 0$ and a pair of centered distributions $(\mu_0, \mu_1) \in \mathcal{P}_4(\mathbb{R}^{d_0}) \times \mathcal{P}_4(\mathbb{R}^{d_1})$. Assume that Φ admits a unique minimizer $\boldsymbol{A}^\star$ and let $\pi_{\boldsymbol{A}^\star}$ be the associated EOT/EGW coupling (e.g., under the conditions of Theorem 19, given that μ_0, μ_1 are compactly supported). If, for $i = 0, 1$, μ_i is invariant under the action of the orthogonal transformation $U_i : \mathbb{R}^{d_i} \to \mathbb{R}^{d_i}$ in the sense that $(U_i)_\sharp\mu_i = \mu_i$ it follows that $(U_0, U_1)_\sharp\pi_{\boldsymbol{A}^\star}$ is also an optimal EGW coupling, whence $(U_0, U_1)_\sharp\pi_{\boldsymbol{A}^\star} = \pi_{\boldsymbol{A}^\star}$ by uniqueness. Thus, by Corollary 6, $\boldsymbol{U}_0^\intercal\boldsymbol{A}^\star\boldsymbol{U}_1 = \boldsymbol{A}^\star$ where $\boldsymbol{U}_i$ is the matrix associated with U_i, for $i = 0, 1$. The previous equality holds for any pair of orthogonal transformations leaving the marginals invariant, so the rows of $\boldsymbol{A}^\star$ are left eigenvectors of $\boldsymbol{U}_1$ with eigenvalue 1 and its columns are right eigenvectors of $\boldsymbol{U}_0^\intercal$ with eigenvalue 1 for every $\boldsymbol{U}_i$ such that $(U_i)_\sharp\mu_i = \mu_i$. Thus, we see that symmetries of the marginals dictate the structure of $\boldsymbol{A}^\star$. For example, if $\mu_1 = (-\operatorname{Id})_\sharp \mu_1$, we have that $\boldsymbol{A}^\star = -\boldsymbol{A}^\star$, so $\boldsymbol{A}^\star = \mathbf{0}$.

4.4 Computational Guarantees

Building on the stability theory from Section 4.3, we now study computation of the EGW problem. The goal is to compute the distance between two discrete distributions $\mu_0 \in \mathcal{P}(\mathbb{R}^{d_0})$ and $\mu_1 \in \mathcal{P}(\mathbb{R}^{d_1})$ supported on N_0 and N_1 atoms $(x^{(i)})_{i=1}^{N_0}$ and $(y^{(j)})_{j=1}^{N_1}$, respectively. In light of the decomposition (4.6), we focus on S_ε^2, which is given by a smooth optimization problem whose convexity depends on the value of ε (cf. Section 4.3). Throughout, we adopt the perspective of Remark 20 and treat $D\Phi_{[\boldsymbol{A}]}$, for $\boldsymbol{A} \in \mathbb{R}^{d_0 \times d_1}$, as the matrix $64\boldsymbol{A} - 32 \int xy^\intercal \, d\pi_{\boldsymbol{A}}(x, y)$.

4.4.1 Inexact Oracle Methods

As these problems are already $d_0 d_1$-dimensional and computing the second Fréchet derivative of Φ may be infeasible (in particular, it requires solving (4.8)), we focus on first-order methods. Given the regularity of the S_ε^2 optimization problem, standard out-of-the-box numerical routines are likely to yield good results in practice. However, to provide meaningful formal guarantees one must account for the fact that evaluation of Φ and its gradient, for $\boldsymbol{A} \in \mathcal{D}_M$, requires computing the corresponding EOT plan, which often entails an approximation. Indeed, an explicit characterizations of the EOT plan between arbitrary distributions is unknown and algorithms typically rely on a fast numerical proxy of the coupling. We model this under the scope of gradient methods with inexact gradient oracles [55, 70, 76].

For a fixed $\varepsilon > 0$ and μ_0, μ_1 as above, our goal is thus to solve

$$\min_{A \in \mathcal{D}_M} 32\|A\|_F^2 + \mathsf{OT}_{A,\varepsilon}(\mu_0, \mu_1),$$

where $M > M_{\mu_0,\mu_1}$, which guarantees that all the optimizers are within the optimization domain (cf. Corollary 6). As we are in the discrete setting, the EOT coupling π^{A} for $\mathsf{OT}_{A,\varepsilon}(\mu_0, \mu_1)$, $A \in \mathcal{D}_M$, is represented by $\mathbf{\Pi}^{A} \in \mathbb{R}^{N_0 \times N_1}$, where $\mathbf{\Pi}^{A}_{ij} = \pi^{A}(x^{(i)}, y^{(j)})$. The inexact oracle paradigm assumes that, for any $A \in \mathcal{D}_M$, we have access to a δ-oracle approximation $\widetilde{\mathbf{\Pi}}^{A}$ of $\mathbf{\Pi}^{A}$ with $\|\widetilde{\mathbf{\Pi}}^{A} - \mathbf{\Pi}^{A}\|_\infty < \delta$. Such oracles can be obtained, for instance, by numerical resolution of the EOT problem. To this end, Sinkhorn's algorithm [184, 53] serves as the canonical approach.

Proposition 11 (Inexact oracle via Sinkhorn iterations)**.** Fix $\delta > 0$. Then, Sinkhorn's algorithm (Algorithm 7) returns an $(e^{\delta} - 1)$-oracle approximation $\widetilde{\mathbf{\Pi}}^{A}$ of $\mathbf{\Pi}^{A}$ in at most $\tilde{k}$ iterations, where $\tilde{k}$ depends only on μ_0, μ_1, A, δ, and ε, and is given explicitly in (C.4).

The proof of Proposition 11 follows by combining a number of known results. Complete details can be found in Section C.2. To our knowledge, the majority of the literature concerning the use of Sinkhorn's algorithm for EOT focuses on approximating unregularized OT and treats the KL divergence term as a bias. Here we quantify the accuracy of estimating the true EOT plan, which may be of an independent interest.

With these preparations, we first discuss the case where Φ is known to be convex on $\mathcal{D}_M$.

4.4.2 Convex Case

Assume that Φ is convex on $\mathcal{D}_M$, e.g., under the setting of Theorem 19. As convexity implies that the minimal eigenvalue of $D^2\Phi_{[A]}$ is positive for any $A \in \mathcal{D}_M$, Theorem 19 further yields that Φ is 64-smooth. With that, we can the apply inexact oracle first-order method from [55]. To describe the approach, assume that we are given a δ-oracle $\widetilde{\mathbf{\Pi}}^{\boldsymbol{A}}$ for the EOT plan $\mathbf{\Pi}^{\boldsymbol{A}}$ for $\mathsf{OT}_{\boldsymbol{A},\varepsilon}(\mu_0, \mu_1)$, and define the corresponding gradient approximation

$$\widetilde{D}\Phi_{[\boldsymbol{A}]} = 64\boldsymbol{A} - 32 \sum_{\substack{1 \le i \le N_0 \\ 1 \le j \le N_1}} x^{(i)}(y^{(j)})^\intercal \widetilde{\mathbf{\Pi}}^{\boldsymbol{A}}_{ij}. \tag{4.9}$$

We now present the algorithm and follow it with formal convergence guarantees.

Algorithm 5 Fast gradient method with inexact oracle

Fix $L = 64$ and let $\alpha_k = \frac{k+1}{2}$, and $\tau_k = \frac{2}{k+3}$

$k \leftarrow 0$

$\boldsymbol{A}_0 \leftarrow \mathbf{0}$

$\boldsymbol{G}_0 \leftarrow \widetilde{D}\Phi_{[\boldsymbol{A}_0]}$

$\boldsymbol{W}_0 \leftarrow \alpha_0 \boldsymbol{G}_0$

while stopping condition is not met **do**

    $\boldsymbol{B}_k \leftarrow \min\left(1, \frac{M}{2\|\boldsymbol{A}_k - L^{-1}\boldsymbol{G}_k\|_F}\right)(\boldsymbol{A}_k - L^{-1}\boldsymbol{G}_k)$

    $\boldsymbol{C}_k \leftarrow -\min\left(1, \frac{M}{2\|L^{-1}\boldsymbol{W}_k\|_F}\right) L^{-1}\boldsymbol{W}_k$

    $\boldsymbol{A}_{k+1} \leftarrow \tau_k \boldsymbol{C}_k + (1 - \tau_k)\boldsymbol{B}_k$

    $\boldsymbol{G}_{k+1} \leftarrow \widetilde{D}\Phi_{[\boldsymbol{A}_{k+1}]}$

    $\boldsymbol{W}_{k+1} \leftarrow \boldsymbol{W}_k + \alpha_{k+1}\boldsymbol{G}_{k+1}$

    $k \leftarrow k + 1$

return $\boldsymbol{B}_k$

The multiplication operations in Algorithm 5 are applied entrywise and it is understood that $\min(1, M/0) = 1$. Due to inexactness, stopping conditions based on insufficient progress of functions values or setting a threshold on the norm of the gradient require care. A condition based on the number of iterations is discussed in Remark 22.

We now provide formal convergence guarantees for Algorithm 5.

Theorem 20 (Fast convergence rates)**.** Assume that Φ is convex and L-smooth on $\mathcal{D}_M$ and that $\widetilde{\mathbf{\Pi}}^{\boldsymbol{A}}$ is a δ-oracle for $\mathbf{\Pi}^{\boldsymbol{A}}$. Then, the iterates $\boldsymbol{B}_k$ in Algorithm 5 with $\widetilde{D}\Phi_{[\boldsymbol{A}_k]}$ given by (4.9) satisfy

$$\Phi(\boldsymbol{B}_k) - \Phi(\boldsymbol{B}^\star) \le \frac{2L\|\boldsymbol{B}^\star\|_F^2}{(k+1)(k+2)} + 3\delta', \tag{4.10}$$

where $\boldsymbol{B}^\star$ is a global minimizer of Φ and $\delta' = 32M\delta \sum_{\substack{1\le i\le N_0\\ 1\le j\le N_1}} \left\|x^{(i)}\right\| \left\|y^{(j)}\right\|$. Moreover, for any $\eta > 3\delta'$, Algorithm 5 requires at most

$$k = \left\lceil -\frac{3}{2} + \frac{1}{2}\sqrt{1 + \frac{8L\|\boldsymbol{B}^\star\|_F^2}{\eta - 3\delta'}} \right\rceil \le \left\lceil -\frac{3}{2} + \frac{1}{2}\sqrt{1 + \frac{128M^2}{\eta - 3\delta'}} \right\rceil \tag{4.11}$$

iterations to achieve an η-approximate solution.

The proof of Theorem 20, given in Section 4.5.5, follows from Theorem 2.2 in [55] after casting our problem as an instance of their setting. Some implications of Theorem 20 are discussed in the following remark.

Remark 22 (Optimal rates and stopping conditions)**.** First, consider the convergence rate of the function values in (4.10). The first term on the right-hand side exhibits the optimal complexity bound for smooth constrained optimization of $O(1/k^2)$ (cf., e.g., [150]). The second term accounts for the underlying oracle error. Notably, the progress of the optimization procedure and the oracle error are completely decoupled in this bound.

Next, we describe a stopping condition based on the number of iterations. Observe that all terms involved in the upper bound in (4.11) are explicit as soon as a desired precision η is chosen since the oracle error δ can be fixed according to Proposition 11. Consequently, (4.11) can be used as an explicit stopping condition for Algorithm 5.

4.4.3 General Case

We now discuss an optimization procedure which does not require convexity of the objective. This accounts for the fact that outside the sufficient conditions of Theorem 19, convexity of Φ is generally unknown. However, the same result shows that Φ is L-smooth with $L = 64 \vee \left(32^2\varepsilon^{-1}\sqrt{M_4(\mu_0)M_4(\mu_1)} - 64\right)$ and $\mathsf{OT}_{(\cdot),\varepsilon}$ is L'-smooth with $L' =$

$32^2\varepsilon^{-1}\sqrt{M_4(\mu_0)M_4(\mu_1)}$. Hence, we adapt the smooth non-convex optimization routine of [89] to account for our inexact oracle. Notably, their method adapts to the convexity of Φ as described in Theorem 21.

We now present the algorithm and describe its convergence rate.

Algorithm 6 Adaptive gradient method with inexact oracle

Given $\boldsymbol{C}_0 \in \mathcal{D}_M$, fix the step sequences $\beta_k = \frac{1}{2L}$, $\gamma_k = \frac{k}{4L}$, and $\tau_k = \frac{2}{k+2}$.

$k \leftarrow 1$
$\boldsymbol{A}_1 \leftarrow \boldsymbol{C}_0$
$\boldsymbol{G}_1 \leftarrow \widetilde{D}\Phi_{[\boldsymbol{A}_1]}$
while stopping condition is not met **do**
    $\boldsymbol{B}_k \leftarrow \min\left(1, \frac{M}{2\|\boldsymbol{A}_k - \beta_k \boldsymbol{G}_k\|_F}\right)(\boldsymbol{A}_k - \beta_k \boldsymbol{G}_k)$
    $\boldsymbol{C}_k \leftarrow \min\left(1, \frac{M}{2\|\boldsymbol{C}_{k-1} - \gamma_k \boldsymbol{G}_k\|_F}\right)(\boldsymbol{C}_{k-1} - \gamma_k \boldsymbol{G}_k)$
    $\boldsymbol{A}_{k+1} \leftarrow \tau_k \boldsymbol{C}_k + (1-\tau_k)\boldsymbol{B}_k$
    $\boldsymbol{G}_{k+1} \leftarrow \widetilde{D}\Phi_{[\boldsymbol{A}_{k+1}]}$
    $k \leftarrow k+1$
return $\boldsymbol{B}_k$

Unlike Algorithm 5, which can be initialized at any fixed $\boldsymbol{A}_0$, the starting point in Algorithm 6 should be chosen according to some selection rule that avoids initializing at a stationary point (e.g., random initialization). Indeed, if $\boldsymbol{A}_1$ is set to a stationary point of Φ, then $D\Phi_{[\boldsymbol{A}_1]} = \boldsymbol{0}$ and, consequently $\widetilde{D}\Phi_{[\boldsymbol{A}_1]} \approx \boldsymbol{0}$ (given that the approximate gradient is reasonably accurate), which may result in premature and undesirable termination. Clearly, this is not a concern for Algorithm 5 since it assumes convexity of Φ, whereby any stationary point is a global optimum.

The following result follows by adapting the proofs of Theorem 2 and Corollary 2 in [89]. For completeness, we provide a self-contained argument in Section C.3 along with a discussion of how this problem fits in the framework of [89].

Theorem 21 (Adaptive convergence rate)**.** Assume that Φ is L-smooth on $\mathcal{D}_M$ and that $\widetilde{\boldsymbol{\Pi}}^{\boldsymbol{A}}$ is a δ-oracle for $\boldsymbol{\Pi}^{\boldsymbol{A}}$. Then, the iterates $\boldsymbol{A}_k, \boldsymbol{B}_k$ in Algorithm 6 with $\widetilde{D}\Phi_{[\boldsymbol{A}_k]}$ given by

(4.9) satisfy

1. If Φ is non-convex and $\mathsf{OT}_{(\cdot),\varepsilon}(\mu_0, \mu_1)$ is L'-smooth, then

$$\min_{1\le i\le k} \left\|\beta_i^{-1}(\boldsymbol{B}_i - \boldsymbol{A}_i)\right\|_F^2 \le \frac{96L^2}{k(k+1)(k+2)}\|\boldsymbol{C}_0 - \boldsymbol{B}^\star\|_F^2 + \frac{24LL'}{k}\left(\|\boldsymbol{B}^\star\|_F^2 + \frac{5M^2}{16}\right) + 8L\delta',$$

where $\boldsymbol{B}^\star$ is a global minimizer of Φ, and $\delta' = 32M\delta \sum_{\substack{1\le i\le N_0\\ 1\le j\le N_1}} \left\|x^{(i)}\right\| \left\|y^{(j)}\right\|$.

2. If Φ is convex, then

$$\min_{1\le i\le k} \left\|\beta_i^{-1}\left(\boldsymbol{B}_i - \boldsymbol{A}_i\right)\right\|_F^2 \le \frac{96L^2}{k(k+1)(k+2)}\|\boldsymbol{C}_0 - \boldsymbol{B}^\star\|_F^2 + 8L\delta'.$$

To better motivate this result, we show that when $\left\|\beta_k^{-1}\left(\boldsymbol{B}_k - \boldsymbol{A}_k\right)\right\|_F$ is small, $D\Phi_{[\boldsymbol{A}_k]}$ is approximately stationary.

Corollary 8 (Approximate stationarity)**.** Let $\boldsymbol{A}_k, \boldsymbol{B}_k$ be iterates from Algorithm 6 and assume that $\boldsymbol{B}_k \in \operatorname{int}(\mathcal{D}_M)$. Then,

$$\|D\Phi_{[\boldsymbol{A}_k]}\|_F < 32\delta \sum_{\substack{1\le i\le N_0\\ 1\le j\le N_1}} \|x^{(i)}\|\|y^{(j)}\| + \left\|\beta_k^{-1}(\boldsymbol{B}_k - \boldsymbol{A}_k)\right\|_F.$$

The proof of Corollary 8 follows from the δ-oracle assumption and the fact that when $\boldsymbol{B}_k$ is an interior point of $\mathcal{D}_M$, we have $\boldsymbol{B}_k = \boldsymbol{A}_k - \beta_k \boldsymbol{G}_k$. See Section 4.5.6 for the full details. When $\boldsymbol{B}_k$ is not an interior point of $\mathcal{D}_M$, the interpretation of $\|\beta_k^{-1}(\boldsymbol{B}_k - \boldsymbol{A}_k)\|_F$ is less straightforward. However, as all stationary points of Φ are contained in $\mathcal{D}_{M_{\mu_0,\mu_1}}$, it is expected that Algorithm 6 will converge to an interior point when $M > M_{\mu_0,\mu_1}$. By analogy with Remark 22, when all iterates are interior points Algorithm 6 yields a bound on the total number of iterations required to achieve an approximate stationary point.

The following remark addresses the distinctions between the convex and non-convex settings in Theorem 21.

Remark 23 (Adaptivity of Algorithm 6)**.** As in Theorem 20, the convergence rates in Theorem 21 are decoupled into a term related to the progress of the optimization procedure and a term related to the oracle error.

In the case where Φ is non-convex, the dominant term in the optimization error is $O(1/k)$, which coincides with the best known rates for solving general unconstrained nonlinear programs [89]. On the other hand, when Φ is convex, the rate of convergence improves to $O(1/k^3)$ which essentially matches the best known rates for the norm of the gradient in the unconstrained accelerated gradient method applied to a convex L-smooth function (see Theorem 6 in [183] and Theorem 3.1 in [44][6]). This adaptivity is beneficial, as Φ may be convex beyond the conditions derived in Theorem 19.

An empirical comparison of Algorithms 5 and 6 in the convex setting is included in Section 4.4.5. In particular, Algorithm 5 is seen to outperform Algorithm 6 in terms of average runtime despite having the same per iteration complexity when the inexact gradient is computed using standard Sinkhorn iterations.

Remark 24 (Computational complexity of Algorithms 5 and 6)**.** As Sinkhorn's algorithm is known to have a complexity of $O(N_0 N_1)$ (cf. e.g. [177]), the gradient approximation (4.9) can be computed in $O(N_0 N_1)$ time. It follows that Algorithms 5 and 6 admit a computational complexity of $O(N_0 N_1)$.

4.4.4 Approximating Unregularized Gromov-Wasserstein Distances

The EGW distance can approximate unregularized GW to an arbitrary precision, with error $|\mathsf{S}_\varepsilon(\mu_0, \mu_1) - \mathsf{S}_0(\mu_0, \mu_1)| = O\left(\varepsilon \log\left(\frac{1}{\varepsilon}\right)\right)$ as $\varepsilon \downarrow 0$ (see Proposition 1 in [223]). It is

[6] More precisely, [44] show that the iterates $(y_i)_{i=1}^k$ generated by the accelerated gradient method applied to a convex L-smooth function f are such that $\min_{0 \le i \le k} \|\nabla f(y_i)\|^2 = o(1/k^3)$.

therefore natural to ask if the proposed algorithms can be used to approximate the unregularized GW distance between finitely supported marginals. Note, however, that for $\varepsilon > 0$ sufficiently small, Φ may fail to be convex (see Theorem 19), whence Algorithm 6 may only converge to an approximate stationary point of Φ. As such, it is not guaranteed that itself $\mathsf{S}_\varepsilon(\mu_0, \mu_1)$ (i.e., the global minimum of the entropic problem) can be approximated to within a desired accuracy. Nevertheless, we show that these approximate stationary points can be used to capture a notion of local optimality for $\mathsf{S}_0^2(\mu_0, \mu_1)$, keeping in mind that the quadratic GW distance decomposes as $\mathsf{S}_0(\mu_0, \mu_1) = \mathsf{S}^1(\mu_0, \mu_1) + \mathsf{S}_0^2(\mu_0, \mu_1)$ for centered distributions (see Corollary 1 in [223]).

As opposed to EOT, optimal solutions to unregularized OT may fail to be unique, which entails that $\Phi_0 := 32\|\cdot\|_F^2 + \mathsf{OT}_{(\cdot),0}(\mu_0, \mu_1)$ may be non-differentiable. Remarkably, the following analogue of Corollary 6 on stationary points still holds.

Proposition 12 (Optimality conditions for Φ_0)**.** Let $(\mu_0, \mu_1) \in \mathcal{P}(\mathbb{R}^{d_0}) \times \mathcal{P}(\mathbb{R}^{d_1})$ be compactly supported. If $\bar{\boldsymbol{A}}$ is a local minimizer of Φ_0, then there exists a solution $\bar{\pi}$ to $\mathsf{OT}_{\bar{\boldsymbol{A}},0}(\mu_0, \mu_1)$ with $\bar{\boldsymbol{A}} = \frac{1}{2}\int xy^\intercal d\bar{\pi}(x, y) \in \mathcal{D}_M$ for any $M > M_{\mu_0,\mu_1}$. If $\bar{\boldsymbol{A}}$ is globally optimal and μ_0, μ_1 are centered, then $\bar{\pi}$ solves $\mathsf{S}_0(\mu_0, \mu_1)$.

The proof of Proposition 12, which will appear in Section 4.5.7, follows similar lines to the proofs provided in the regularized case with the caveat that Φ_0 is merely locally Lipschitz. To adapt these results, we characterize the Clarke subdifferential of Φ_0 [49].

Proposition 12 shows that any minimizer, $\bar{\boldsymbol{A}}_0 \in \mathcal{D}_M$, of Φ_0 satisfies an analogous criterion to the stationary points, $\bar{\boldsymbol{A}}_\varepsilon \in \mathcal{D}_M$, of $\Phi_\varepsilon := 32\|\cdot\|_F^2 + \mathsf{OT}_{(\cdot),\varepsilon}(\mu_0, \mu_1)$ for $\varepsilon > 0$ (namely, that $\bar{\boldsymbol{A}}_\varepsilon = \frac{1}{2}\int xy^\intercal d\bar{\pi}_\varepsilon(x, y)$ for some $\bar{\pi}_\varepsilon$ solving $\mathsf{OT}_{\bar{\boldsymbol{A}}_\varepsilon,0}(\mu_0, \mu_1)$ for $\varepsilon \geq 0$). As noted previously, when ε is small, we can only guarantee that Algorithm 6 will converge to an approximate stationary point, $\boldsymbol{A}_\varepsilon^\star \in \mathcal{D}_M$, satisfying $\|D(\Phi_\varepsilon)_{[\boldsymbol{A}_\varepsilon^\star]}\|_F =$

$\|64\boldsymbol{A}_\varepsilon^\star - 32\int xy^\intercal d\pi_\varepsilon^\star(x,y)\|_F \le \delta$, for some $\delta > 0$, where $\pi_\varepsilon^\star$ solves $\mathsf{OT}_{\boldsymbol{A}_\varepsilon^\star,\varepsilon}(\mu_0,\mu_1)$. We now show that the limit points of $\boldsymbol{A}_\varepsilon^\star$, as $\varepsilon \downarrow 0$, are approximately stationary for Φ_0. Furthermore, if the matrices along the sequence are globally/locally optimal (in an appropriate sense), we show that global/local optimality is inherited at the limit.

Theorem 22 (Convergence of approximate stationary points)**.** Let $(\mu_0,\mu_1) \in \mathcal{P}(\mathbb{R}^{d_0}) \times \mathcal{P}(\mathbb{R}^{d_1})$ be compactly supported and fix $\delta \ge 0$. For $\varepsilon > 0$, let $\boldsymbol{A}_\varepsilon^\star \in \mathcal{D}_M$ be such that $\|D(\Phi_\varepsilon)_{[\boldsymbol{A}_\varepsilon^\star]}\|_F \le \delta$ and let $\boldsymbol{A}_0^\star$ be a cluster point of $(\boldsymbol{A}_\varepsilon^\star)_{\varepsilon>0}$ (as $\varepsilon \downarrow 0$). Then,

1. $\|64\boldsymbol{A}_0^\star - 32\int xy^\intercal d\pi_{\boldsymbol{A}_0^\star}\|_F \le \delta$, where $\pi_{\boldsymbol{A}_0^\star}$ is a solution of $\mathsf{OT}_{\boldsymbol{A}_0^\star,0}(\mu_0,\mu_1)$.
2. if $\boldsymbol{A}_\varepsilon^\star$ is a global minimizer of Φ_ε for all $\varepsilon > 0$ sufficiently small, then $\boldsymbol{A}_0^\star$ minimizes Φ_0.
3. if, up to a subsequence $\varepsilon_n \downarrow 0$ along which $\boldsymbol{A}_{\varepsilon_n}^\star \to \boldsymbol{A}_0^\star$, $\boldsymbol{A}_{\varepsilon_n}^\star$ minimizes Φ_{ε_n} on a ball of fixed radius $r > 0$ centred at $\boldsymbol{A}_0^\star$, then $\boldsymbol{A}_0^\star$ is a local minimizer of Φ_0.

Recall that $\mathcal{D}_M$ is compact such that a cluster point $\boldsymbol{A}_0^\star$ always exists in Theorem 22. In light of Proposition 12, these limit points can be thought of as approximate stationary points for Φ_0. This enables approximating local solutions to the unregularized variational GW problem by stationary points obtained using Algorithm 6 for small ε. The proof of Theorem 22 uses the notion of Γ-convergence (see e.g. [29]) to show that the solutions of $\mathsf{OT}_{\boldsymbol{A}_\varepsilon^\star,\varepsilon}(\mu_0,\mu_1)$ converge to a solution of $\mathsf{OT}_{\boldsymbol{A}_0^\star,0}(\mu_0,\mu_1)$ up to a subsequence and that the local/global minimizers of Φ_ε admit local/global minimizers of Φ_0 as cluster points, see Section 4.5.8 for details.

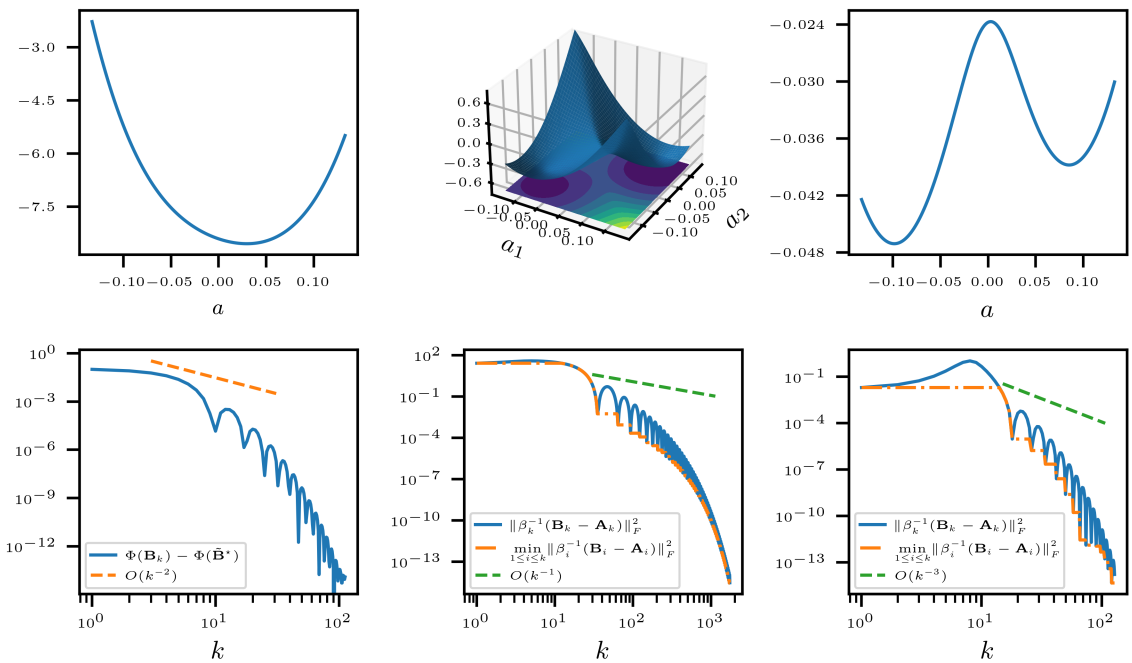
4.4.5 Numerical Experiments

We conclude this section with some experiments that empirically validate the rates obtained in Theorems 20 and 21 and the computational complexity discussed in Remark 24. All experiments were performed on a desktop computer with 16 GB of RAM and an Intel i5-10600k CPU using the Python programming language. Implementations of Algorithms 5 and 6 are available at `https://github.com/gabrielrioux/EntropicGromovWasserstein`. The considered marginal distributions described below, μ_0, μ_1 were randomly generated.

Convergence rates. Figure 4.1 (a) presents an example of applying Algorithm 5 to a convex Φ, where the marginals are $\mu_0 = 0.4\delta_{-1.4} + 0.6\delta_{1.2}$ and $\mu_1 = 0.4\delta_{-1.01} + 0.6\delta_{1.31}$, with ε chosen large enough to guarantee convexity. The theoretical rate of $O(k^{-2})$ from Theorem 20 on the optimality gap $\Phi(\boldsymbol{B}_k) - \Phi(\boldsymbol{B}^\star)$ is seen to hold.[7] Figure 4.1 (b) illustrates the progress of Algorithm 6 applied to a non-convex Φ, for $\mu_0 = \frac{1}{3}(\delta_{0.3} + \delta_{-0.8} + \delta_{-0.5})$ and $\mu_1 = \frac{1}{3}(\delta_{(0.1,0.6)} + \delta_{(-0.5,0.3)} + \delta_{(0.4,-0.3)})$, with $\varepsilon = 0.07$ which makes Φ non-convex. The $O(k^{-1})$ rate for $\min_{1\leq i\leq k} \left\|\beta_i^{-1}(\boldsymbol{B}_i - \boldsymbol{A}_i)\right\|_F^2$ in the non-convex case from Theorem 21 is well reflected in this example. Figure 4.1 (c) shows that Algorithm 6 can match the theoretical rate of $O(k^{-3})$ in the convex regime when initialized in a region of local convexity. In this example, the generated marginals are $\mu_0 = \frac{1}{5}(\delta_{-0.1} + \delta_{-0.2} + \delta_{0.2} + \delta_{-0.3} + \delta_{0.3})$ and $\mu_1 = \frac{1}{5}(\delta_{0.2} + \delta_{-0.3} + \delta_{0.3} + \delta_{-0.4} + \delta_{0.4})$ and $\varepsilon = 0.03$. The stopping condition used in all these example is $\|\boldsymbol{G}_k\|_F < 5 \times 10^{-8}$ and the approximate gradient (4.9) is computed using the standard implementation of Sinkhorn's algorithm from the Python Optimal Transport package [82].

Time complexity. To study the time complexity of Algorithms 5 and 6, we first

[7] The plot shows the approximate gap $\Phi(\boldsymbol{B}_k) - \Phi(\bar{\boldsymbol{B}}^\star)$, where $\bar{\boldsymbol{B}}^\star$ is the iterate attaining the minimal objective value.



(a) (b) (c)

Figure 4.1: The top row compiles plots of Φ for the different examples described in the text. The bottom row consists of plots tracking the progress of the iterates. In (a), the gap to optimality in terms of function value is tracked and is seen to converge to 0 faster than k^{-2}. In (b) and (c), Algorithm 6 is initialized at $\boldsymbol{C}_0 = (1, 1) \times 10^{-5}$ and $\boldsymbol{C}_0 = 1 \times 10^{-5}$, respectively and, in both cases, the progress is tracked in terms of the squared norm of the approximate gradient.

choose the dimension $d \in \{1, 16, 64, 128\}$ and let $\mu_0, \mu_1 \in \mathcal{P}(\mathbb{R}^d)$ be supported on $N \in \{16, 32, 64, 128, 256, 512, 1024, 2048, 4096, 8192, 16384\}$ samples of a mean-zero normal distribution with standard deviation 0.05 for μ_0 and 0.1 for μ_1. The weights are chosen uniformly at random from $[0, 1)$ and normalized so as to sum to 1. This procedure is repeated to generate a collection of pairs of random distributions $\{(\mu_{0,i}, \mu_{1,i})\}_{i=1}^{50}$. In the sequel, a *single experiment* refers to the process of timing the computation of $\mathsf{S}_\varepsilon(\mu_{0,i}, \mu_{1,i})$ for some fixed d, N and all $i = 1, \ldots, 50$. For practical reasons, we choose to abort an experiment before all 50 EGW distances have been computed if the total runtime for this experiment exceeds 1 hour. The average runtime is then computed among all completed calculations in a single experiment.

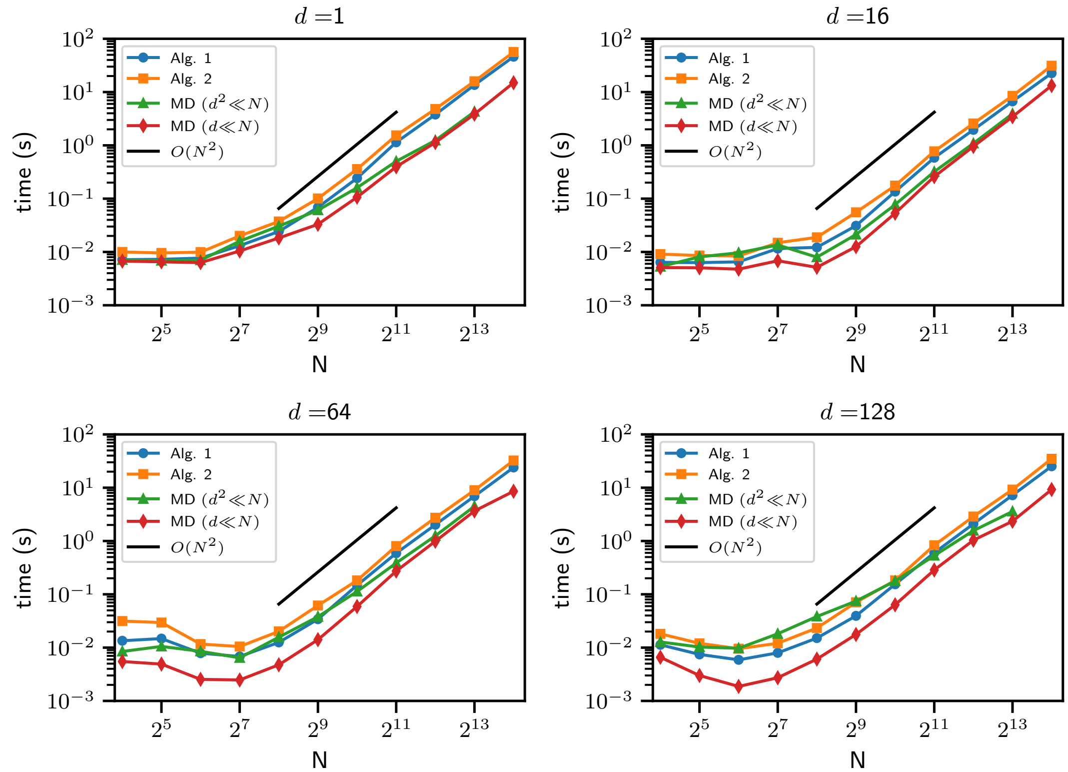


Figure 4.2: The various plots compile the average runtime of Algorithms 5 and 6, and two versions of the mirror descent algorithm in the convex regime for different values of d and N.

The convex case: First, ε is chosen as $1.05 \times 16\sqrt{M_4(\mu_0)M_4(\mu_1)}$ so as to guarantee convexity of Φ for each instance by Theorem 19 and M is set to $\sqrt{M_2(\mu_0)M_2(\mu_1)} + 10^{-5}$. Figure 4.2 presents the average runtime of both algorithms in this setting with the stopping condition $\|\boldsymbol{G}_k\|_F < 10^{-6}$. We compare the performance of our methods with the two implementations of the $O(N^2)$ mirror descent algorithm provided in [177].[8] The first implementation includes certain algorithmic tweaks when $d^2 \ll N$, whereas the second only requires $d \ll N$ to achieve the quadratic complexity. Our implementation of the mirror descent algorithm is based on the code provided in [177] with some small modifications (e.g., EOT couplings are computed using Sinkhorn's algorithm from

[8]We do not compare with the original implementation of mirror descent [160] or the iterative algorithm from [186] due to their much slower $O(N^3)$ runtime.

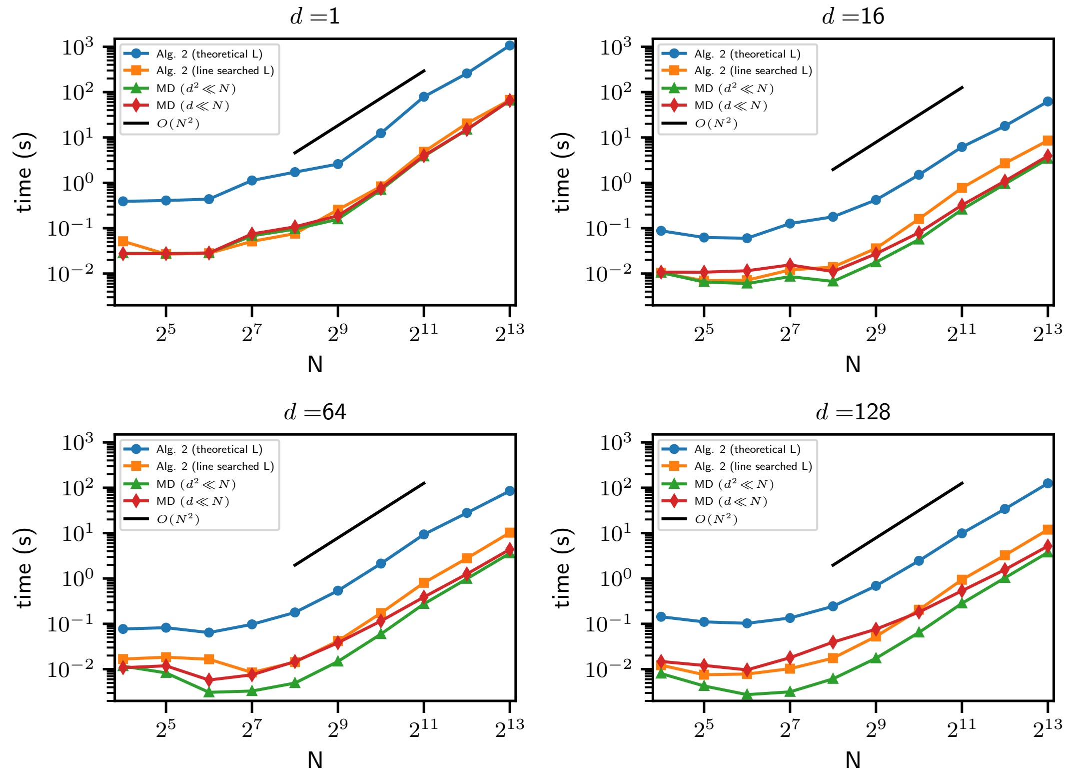


Figure 4.3: The various plots compile the average runtime of Algorithm 6 with the two methods for choosing L, and two versions of the mirror descent algorithm in the non-convex regime for different values of d and N.

the Python Optimal Transport package [82] and some extraneous logging features are removed) the algorithm is run until the generated couplings differ by less than 10^{-6} under the Frobenius norm. We note that the first version of the mirror descent algorithm encounters "out of memory" errors for $N = 16384$.

The plots show that the four algorithms perform similarly on the considered examples, and empirically validate the $O(N^2)$ computational complexity from Remark 24. To verify that the algorithms all converge to solutions with similar objective values, we evaluate the relative error[9] between all pairs of algorithms for each d, N. The largest relative error

[9] Relative error is measured by $\max_{i\in C(d,N)} \left|\mathsf{S}^{A1}_{\varepsilon}(\mu_{0,i},\mu_{1,i}) - \mathsf{S}^{A2}_{\varepsilon}(\mu_{0,i},\mu_{1,i})\right| / (\mathsf{S}^{A1}_{\varepsilon}(\mu_{0,i},\mu_{1,i}) \wedge \mathsf{S}^{A2}_{\varepsilon}(\mu_{0,i},\mu_{1,i}))$, where $\mathsf{S}^{A1}_{\varepsilon}(\mu_{0,i},\mu_{1,i})$ and $\mathsf{S}^{A2}_{\varepsilon}(\mu_{0,i},\mu_{1,i})$ denote the objective values achieved by the first and second algorithm of the pair, and $C(d, N)$ is the collection of completed runs from a given experiment.

we observe is 3.3×10^{-6} for $d = 1$ and, for the other choices of d, is at most 7.9×10^{-13}. We conclude that the values obtained are in good agreement.

The non-convex case: To evaluate the performance of Algorithm 6 when convexity is unknown, we set ε to violate the condition of Theorem 19, but still be large enough so as to avoid numerical errors. If errors in running Algorithm 6 or the mirror descent methods occur, we double ε until all algorithms converge without errors. The initial point $\boldsymbol{C}_0$ for Algorithm 6 is taken as the matrix of all ones scaled by $\min\{M, 1\} \times 10^{-5}$. We consider two ways of choosing the smoothness parameter L, which effectively dictates the rate of convergence. The first is to set L equals to the theoretical upper bound from Theorem 19, i.e., $L = 64 \vee \left(32^2 \varepsilon^{-1} \sqrt{M_4(\mu_0) M_4(\mu_1)} - 64\right)$. As this choice may be too conservative in practice, we also consider setting L via a line search. Namely, we fix a value for L (e.g., the theoretical upper bound or an arbitrary value) and check if the algorithm converges for a given choice of $d, N, \mu_{0,i}, \mu_{1,i}$. If so, we multiply L by 0.99 and repeat this procedure until the algorithm no longer converges. For each d and N, we choose the value of L that attains the fastest convergence, and repeat this procedure for 5 pairs of distributions. For Algorithm 6 with the choice of L that comes from the theoretical bound and the two versions of mirror descent we follow the same methodology as in the convex case, i.e., averaging over 50 pairs and stopping an experiment after 1 hour. The average runtimes of all methods are reported in Figure 4.3. The restriction to 5 runs in the line search case is only out of convenience and we note that all algorithms yield similar results if we restrict to 5 runs throughout.

The plots again validate the $O(N^2)$ time complexity for all four approaches. However, we see that choosing L in Algorithm 6 according to the theoretical upper bound may indeed be too conservative, as it results in a 10× or larger slowdown compared to the other methods. By setting L via the line search, on the other hand, Algorithm 6 and mirror descent exhibit similar performance. This suggests that the longer runtime of Algorithm 6 with the theoretical L value can be attributed to this being an overly conservative choice as opposed to a fundamental limitation of this method. Optimization routines that update L at each iteration have been proposed in [195, 13, 151], but require solving an additional EOT problem at each step for our application. As such, these approaches may reduce the number of iterations required for convergence, at the cost of increasing the per iteration complexity.

Real-world data. We next assess the performance of our algorithms on real-world data from the Fashion-MNIST dataset [215]. Since the ground truth EGW value is unknown, we test the performance of the algorithm in capturing the invariance of the EGW distance to isometries. As the EGW distance generally does not nullify between isomorphic mm spaces,[10] we consider a centered/debiased version, inspired by debiased EOT (also known as Sinkhorn divergence, [81, 88]). Define the debiased quadratic EGW distance between $(\mu_0, \mu_1) \in \mathcal{P}(\mathbb{R}^{d_0}) \times \mathcal{P}(\mathbb{R}^{d_1})$ as

$$\bar{\mathsf{S}}_\varepsilon(\mu_0, \mu_1) := \mathsf{S}_\varepsilon(\mu_0, \mu_1) - \frac{1}{2}(\mathsf{S}_\varepsilon(\mu_0, \mu_0) + \mathsf{S}_\varepsilon(\mu_1, \mu_1)).$$

The recentering guarantees that $\bar{\mathsf{S}}_\varepsilon$ nullifies whenever $(\mathbb{R}^{d_0}, \|\cdot\|, \mu_0)$ and $(\mathbb{R}^{d_1}, \|\cdot\|, \mu_1)$ are isomorphic, as desired.

Having that, our experiment consists of comparing a fixed image from the Fashion MNIST dataset to rotated versions of itself and other images from the dataset, by computing the debiased EGW distance between them. By doing so, we empirically validate that computation of $\bar{\mathsf{S}}_\varepsilon$ using Algorithm 6 is invariant to isometric transformations on a real-world dataset. Precisely, we pad the 28×28 pixel images from the dataset with

[10] Indeed, $\mathsf{D}_{\mathsf{KL}}(\pi \| \mu_0 \otimes \mu_1) \geq 0$ with equality if and only if $\pi = \mu_0 \otimes \mu_1$ whereas the integral of the distance distortion cost vanishes if and only if the coupling is induced by some isometry $T : \mathbb{R}^{d_0} \to \mathbb{R}^{d_1}$. By Corollary 6, an optimal coupling $\pi^\star$ for $\mathsf{S}_\varepsilon(\mu_0, \mu_1)$ is equivalent to $\mu_0 \otimes \mu_1$ (in the sense that $\pi^\star \ll \mu_0 \otimes \mu_1$ and $\mu_0 \otimes \mu_1 \ll \pi^\star$), so $\pi^\star$ cannot be induced by an isometry unless μ_0, μ_1 are supported on one point.

zeros on all sides such that the effective image size is 40×40 pixels (this guarantees that no nonzero pixels are lost upon rotation) and treat these padded images as probability distributions on a 40×40 grid of points in $\mathbb{R}^2$ with weights proportional to the pixel intensity value. We then compare these distributions using $\bar{\mathsf{S}}_\varepsilon$ upon removing the points with zero mass from each distribution, the values thus obtained are included in Figure 4.4.

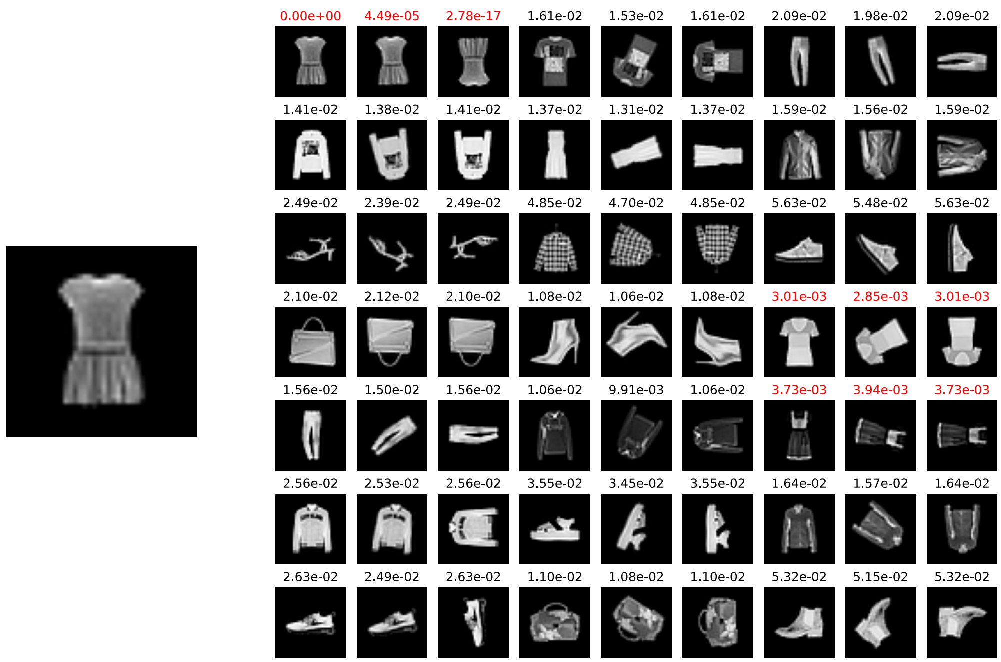


Figure 4.4: We compare the image on the left to the images on the right using debiased EGW with $\varepsilon = 0.1$ as a figure of merit. The corresponding value of $\bar{\mathsf{S}}_\varepsilon$ is included on top of the corresponding images. The images are presented in groups of three, where the leftmost image in the group is the original image, the middle image is obtained via a random rotation of the original image, and the rightmost image is a rotation by a multiple of 90°. Distance values smaller than 5×10^{-3} are in red.

We note that rotating an image by an angle which is not a multiple of 90° does not correspond to an isometric action on $\mathbb{R}^2$, as it requires interpolating the pixel back onto the 40×40 grid of points. Nevertheless, we see from Figure 4.4, that the images subject to random rotations and the unrotated images achieve similar values of $\bar{\mathsf{S}}_\varepsilon$ relative to the fixed reference image. The discrepancy between these two values can be thought of as a quantification of the distortion to the image structure caused by the interpolation procedure. When the rotation is a multiple of 90°, we see that the values obtained are identical, as no interpolation is performed on the image.

Figure 4.4 marks in red values corresponding to images that are closest to the reference image in the debiased EGW distance. These images have a notable structural similarity to the reference in their overall shape and/or features (e.g., pleats on the dress). We also observe that quite naturally the images with largest discrepancy are those of shoes, which have a distinct structure, and the plaid shirt which has a different pattern. Interestingly, between these two extremes, there are images which have comparable values, but are structurally dissimilar (for instance the images in the center column except for the plaid shirt and the sandal). This behaviour demonstrates that the debiased EGW distance does not simply compare the shapes of the images, but rather takes into account the intricate interplay between the intensity values of the images.

4.5 Proofs

4.5.1 Proof of Proposition 10

We first fix some notation. Let $S_i = \mathrm{spt}(\mu_i)$ for $i = 0, 1$ and define the Banach spaces

$$\mathfrak{E} = \left\{(f_0, f_1) \in C(S_0) \times C(S_1) : \int f_0 d\mu_0 = 0\right\},$$
$$\mathfrak{F} = \left\{(f_0, f_1) \in C(S_0) \times C(S_1) : \int f_0 d\mu_0 = \int f_1 d\mu_1\right\}.$$

Consider the map $\Upsilon : \mathbb{R}^{d_0 \times d_1} \times \mathfrak{E} \to C(S_0) \times C(S_1)$ given by

$$\Upsilon : (\boldsymbol{A}, \varphi_0, \varphi_1) \mapsto \left(\int e^{\frac{\varphi_0(\cdot)+\varphi_1(y)-c_{\boldsymbol{A}}(\cdot,y)}{\varepsilon}} d\mu_1(y) - 1, \int e^{\frac{\varphi_0(x)+\varphi_1(\cdot)-c_{\boldsymbol{A}}(x,\cdot)}{\varepsilon}} d\mu_0(x) - 1\right).$$

For fixed $\boldsymbol{A} \in \mathbb{R}^{d_0 \times d_1}$, the solution to the equation $\Upsilon(\boldsymbol{A}, \cdot, \cdot) = 0$ is the unique pair of EOT potentials $(\varphi_0^{\boldsymbol{A}}, \varphi_1^{\boldsymbol{A}})$ for μ_0, μ_1 with the cost $c_{\boldsymbol{A}}$ satisfying the normalization from $\mathfrak{E}$. Observe that, by compactness of S_0 and S_1, the potentials are bounded.

The following lemmas verify the conditions to apply the implicit mapping theorem to Υ in order to obtain the Fréchet derivative of the map $\boldsymbol{A} \in \mathbb{R}^{d_0 \times d_1} \mapsto (\varphi_0^{\boldsymbol{A}}, \varphi_1^{\boldsymbol{A}})$. Given that $\mathsf{OT}_{\boldsymbol{A},\varepsilon}(\mu_0, \mu_1) = \int \varphi_0^{\boldsymbol{A}} d\mu_0 + \int \varphi_1^{\boldsymbol{A}} d\mu_1$, the derivative of the map $\boldsymbol{A} \mapsto \mathsf{OT}_{\boldsymbol{A},\varepsilon}(\mu_0, \mu_1)$ and that of Φ itself will readily follow.

Lemma 33. The map Υ is smooth with first derivative at $(\boldsymbol{A}, \varphi_0, \varphi_1) \in \mathbb{R}^{d_0 \times d_1} \times \mathfrak{E}$ given by,

$$
\begin{aligned}
D\Upsilon_{[\boldsymbol{A},\varphi_0,\varphi_1]}(\boldsymbol{B}, h_0, h_1) = \varepsilon^{-1}\Bigg(&\int (h_0(\cdot) + h_1(y) + 32(\cdot)^{\intercal}\boldsymbol{B}y) e^{\frac{\varphi_0(\cdot)+\varphi_1(y)-c_{\boldsymbol{A}}(\cdot,y)}{\varepsilon}} d\mu_1(y), \\
&\int (h_0(x) + h_1(\cdot) + 32x^{\intercal}\boldsymbol{B}(\cdot)) e^{\frac{\varphi_0(x)+\varphi_1(\cdot)-c_{\boldsymbol{A}}(x,\cdot)}{\varepsilon}} d\mu_0(x)\Bigg),
\end{aligned}
$$

where $(\boldsymbol{B}, h_0, h_1) \in \mathbb{R}^{d_0 \times d_1} \times \mathfrak{E}$.

The proof of this result is straightforward, but included in Section C.4.1 for completeness. Now, define $\xi_{\boldsymbol{A}} := \varepsilon D\Upsilon_{[\boldsymbol{A},\varphi_0^{\boldsymbol{A}},\varphi_1^{\boldsymbol{A}}]}(0, \cdot, \cdot)$ and let $\pi_{\boldsymbol{A}}$ be the EOT coupling for $\mathsf{OT}_{\boldsymbol{A},\varepsilon}(\mu_0, \mu_1)$. Note that for any $(h_0, h_1) \in \mathfrak{E}$, we have $\xi_{\boldsymbol{A}}(h_0, h_1) \in \mathfrak{F}$, which follows by recalling that $\frac{d\pi_{\boldsymbol{A}}}{d\mu_0 \otimes \mu_1}(x, y) = e^{\frac{\varphi_0^{\boldsymbol{A}}(x)+\varphi_1^{\boldsymbol{A}}(y)-c_{\boldsymbol{A}}(x,y)}{\varepsilon}}$ and observing

$$
\begin{aligned}
\int (\xi_{\boldsymbol{A}}(h_0, h_1))_1 d\mu_0 &= \int h_0 d\mu_0 + \int h_1 d\pi_{\boldsymbol{A}} = \int h_0 d\mu_0 + \int h_1 d\mu_1 \\
\int (\xi_{\boldsymbol{A}}(h_0, h_1))_2 d\mu_1 &= \int h_0 d\pi_{\boldsymbol{A}} + \int h_1 d\mu_1 = \int h_0 d\mu_0 + \int h_1 d\mu_1.
\end{aligned}
$$

We next prove that $\xi_{\boldsymbol{A}}$ is an isomorphism between $\mathfrak{E}$ and $\mathfrak{F}$ by following the proof of Proposition 3.1 in [38].

Lemma 34. The map $\xi_{\boldsymbol{A}}$ is an isomorphism between $\mathfrak{E}$ and $\mathfrak{F}$.

Proof. Observe that $\xi_{\boldsymbol{A}}$ extends naturally to a map on $L^2(\mu_0) \times L^2(\mu_1)$ and admits the representation

$$\xi_{\boldsymbol{A}} : (f_0, f_1) \in L^2(\mu_0) \times L^2(\mu_1) \mapsto (f_0, f_1) + \mathcal{L}(f_0, f_1) \in L^2(\mu_0) \times L^2(\mu_1),$$

where

$$\mathcal{L}(f_0, f_1) = \left(\int f_1(y) e^{\frac{\varphi_0^{\boldsymbol{A}}(\cdot)+\varphi_1^{\boldsymbol{A}}(y)-c_{\boldsymbol{A}}(\cdot,y)}{\varepsilon}} d\mu_1(y), \int f_0(x) e^{\frac{\varphi_0^{\boldsymbol{A}}(x)+\varphi_1^{\boldsymbol{A}}(\cdot)-c_{\boldsymbol{A}}(x,\cdot)}{\varepsilon}} d\mu_0(x) \right).$$

Lemma 73 in Appendix C.4.2 demonstrates that $\mathcal{L}$ is a compact linear self-map of $L^2(\mu_0) \times L^2(\mu_1)$.

We first show that $\xi_{\boldsymbol{A}}$ is injective on $E := \{(f_0, f_1) \in L^2(\mu_0) \times L^2(\mu_1) : \int f_0 d\mu_0 = 0\}$. Suppose that $(\bar{f}_0, \bar{f}_1)$ satisfies $\xi_{\boldsymbol{A}}(\bar{f}_0, \bar{f}_1) = 0$. Multiplying $(\xi_{\boldsymbol{A}}(\bar{f}_0, \bar{f}_1))_1$ by $\bar{f}_0$ and $(\xi_{\boldsymbol{A}}(\bar{f}_0, \bar{f}_1))_2$ by $\bar{f}_1$, we have

$$\begin{aligned}
\int \left(\bar{f}_0^2(\cdot) + \bar{f}_0(\cdot)\bar{f}_1(y) \right) e^{\frac{\varphi_0^{\boldsymbol{A}}(\cdot)+\varphi_1^{\boldsymbol{A}}(y)-c_{\boldsymbol{A}}(\cdot,y)}{\varepsilon}} d\mu_1(y) = 0, \\
\int \left(\bar{f}_0(x) f_1(\cdot) + \bar{f}_1^2(\cdot) \right) e^{\frac{\varphi_0^{\boldsymbol{A}}(x)+\varphi_1^{\boldsymbol{A}}(\cdot)-c_{\boldsymbol{A}}(x,\cdot)}{\varepsilon}} d\mu_0(x) = 0,
\end{aligned}$$

and summing these equations gives $\int (\bar{f}_0 + \bar{f}_1)^2 d\pi_{\boldsymbol{A}} = 0$. As $\pi_{\boldsymbol{A}}$ is equivalent to $\mu_0 \otimes \mu_1$, we have $\bar{f}_0 + \bar{f}_1 = 0$ $\mu_0 \otimes \mu_1$-a.e., which further implies that $(\bar{f}_0, \bar{f}_1) = (a, -a)$ $\mu_0 \otimes \mu_1$-a.e. for some $a \in \mathbb{R}$. Consequently, $\ker(\xi_{\boldsymbol{A}})$ is 1-dimensional and $\xi_{\boldsymbol{A}}$ is injective on E.

Next, we show that $\xi_{\boldsymbol{A}}$ is onto $F := \{(f_0, f_1) \in L^2(\mu_0) \times L^2(\mu_1) : \int f_0 d\mu_0 = \int f_1 d\mu_1\}$. As in the lead-up to this lemma, $\xi_{\boldsymbol{A}}(E) \subset F$. By the Fredholm alternative (cf. Theorem 6.6 in [30]), $(\mathrm{Id} + \mathcal{L})(L^2(\mu_0) \times L^2(\mu_1))$ has codimension 1 and, as F has codimension 1, we must have $\xi_{\boldsymbol{A}}(E) = F$.

As such, for any $(g_0, g_1) \in \mathfrak{F} \subset F$, there exists $(h_0, h_1) \in E$ for which

$$\xi_{\boldsymbol{A}}(h_0, h_1) = (h_0, h_1) + \mathcal{L}(h_0, h_1) = (g_0, g_1).$$

As $\mathcal{L}(h_0, h_1) \in C(S_0) \times C(S_1)$, $(h_0, h_1) = (g_0, g_1) - \mathcal{L}(h_0, h_1) \in C(S_0) \times C(S_1)$ with $\int h_0 d\mu_0 = 0$, and thus $(h_0, h_1) \in \mathfrak{E}$. This implies that $\xi_{\boldsymbol{A}}(\mathfrak{E}) \supset \mathfrak{F}$ and from before we have $\xi_{\boldsymbol{A}}(\mathfrak{E}) \subset \mathfrak{F}$, yielding $\xi_{\boldsymbol{A}}(\mathfrak{E}) = \mathfrak{F}$. We have shown that $\xi_A : \mathfrak{E} \to \mathfrak{F}$ is a continuous linear bijection and hence an isomorphism by the open mapping theorem (cf. Corollary 2.7 in [30]). $\square$

We now apply the implicit mapping theorem to obtain the Fréchet derivative of $(\varphi_0^{(\cdot)}, \varphi_1^{(\cdot)})$.

Lemma 35. The map $\boldsymbol{A} \in \mathbb{R}^{d_0 \times d_1} \mapsto (\varphi_0^{\boldsymbol{A}}, \varphi_1^{\boldsymbol{A}}) \in \mathfrak{E}$ is smooth with Fréchet derivative

$$D\left(\varphi_0^{(\cdot)}, \varphi_1^{(\cdot)}\right)_{[\boldsymbol{A}]}(\boldsymbol{B}) = -\left(h_0^{\boldsymbol{A},\boldsymbol{B}}, h_1^{\boldsymbol{A},\boldsymbol{B}}\right),$$

where $\left(h_0^{\boldsymbol{A},\boldsymbol{B}}, h_1^{\boldsymbol{A},\boldsymbol{B}}\right) \in \mathfrak{E}$ satisfies

$$\begin{aligned}
&\int \left(h_0^{\boldsymbol{A},\boldsymbol{B}}(x) + h_1^{\boldsymbol{A},\boldsymbol{B}}(y) - 32x^\intercal \boldsymbol{B} y\right) e^{\frac{\varphi_0^{\boldsymbol{A}}(x)+\varphi_1^{\boldsymbol{A}}(y)-c_{\boldsymbol{A}}(x,y)}{\varepsilon}} d\mu_1(y) = 0, \quad \forall\, x \in \mathrm{spt}(\mu_0),\\
&\int \left(h_0^{\boldsymbol{A},\boldsymbol{B}}(x) + h_1^{\boldsymbol{A},\boldsymbol{B}}(y) - 32x^\intercal \boldsymbol{B} y\right) e^{\frac{\varphi_0^{\boldsymbol{A}}(x)+\varphi_1^{\boldsymbol{A}}(y)-c_{\boldsymbol{A}}(x,y)}{\varepsilon}} d\mu_0(x) = 0, \quad \forall\, y \in \mathrm{spt}(\mu_1),
\end{aligned} \tag{4.12}$$

with $(\varphi_0^{\boldsymbol{A}}, \varphi_1^{\boldsymbol{A}})$ being any pair of EOT potentials for (μ_0, μ_1) with the cost $c_{\boldsymbol{A}}$.

Proof. Fix $\boldsymbol{A} \in \mathbb{R}^{d_0 \times d_1}$ with corresponding EOT potentials $\left(\varphi_0^{\boldsymbol{A}}, \varphi_1^{\boldsymbol{A}}\right)$. For notational convenience, define the shorthands $D_1\Upsilon_{\boldsymbol{A}} = D\Upsilon_{[\boldsymbol{A},\varphi_0^{\boldsymbol{A}},\varphi_1^{\boldsymbol{A}}]}(\cdot, 0, 0)$ and $D_2\Upsilon_{\boldsymbol{A}} = D\Upsilon_{[\boldsymbol{A},\varphi_0^{\boldsymbol{A}},\varphi_1^{\boldsymbol{A}}]}(0, \cdot, \cdot)$ (cf. Lemma 33). By Lemma 34, $D_2\Upsilon_{\boldsymbol{A}}$ is an isomorphism and we may invoke the implicit mapping theorem (cf. Theorem 5.14 in [27]). This implies that there exists an open neighborhood $U \subset \mathbb{R}^{d_0 \times d_1}$ of $\boldsymbol{A}$ and a smooth map $g : U \to \mathfrak{E}$ for which $\Upsilon(\boldsymbol{B}, g(\boldsymbol{B})) = 0$ for every $\boldsymbol{B} \in U$ and

$$Dg_{[A]}(\boldsymbol{B}) = -(D_2\Upsilon_{\boldsymbol{A}})^{-1}\left(D_1\Upsilon_{\boldsymbol{A}}(\boldsymbol{B})\right),$$

i.e., $-Dg_{[A]}(\boldsymbol{B})$ solves (4.12). By construction, $g(\boldsymbol{B}) = (\varphi_0^{\boldsymbol{B}}, \varphi_1^{\boldsymbol{B}})$ and by repeating this process at any $\boldsymbol{A} \in \mathbb{R}^{d_0 \times d_1}$, differentiability of the potentials is extended to the entire space $\mathbb{R}^{d_0 \times d_1}$. $\square$

Given the dual form of the EOT cost, Lemma 35 suffices to prove Proposition 10.

Proof of Proposition 10. As $\mathsf{OT}_{\boldsymbol{A},\varepsilon}(\mu_0,\mu_1) = \int \varphi_0^{\boldsymbol{A}} d\mu_0 + \int \varphi_1^{\boldsymbol{A}} d\mu_1$, Lemma 35 implies that $\mathsf{OT}_{(\cdot),\varepsilon}(\mu_0,\mu_1)$ is smooth with first derivative at $\boldsymbol{A} \in \mathbb{R}^{d_0\times d_1}$ given by

$$D(\mathsf{OT}_{(\cdot),\varepsilon}(\mu_0,\mu_1))_{[\boldsymbol{A}]}(\boldsymbol{B}) = -\int h_0^{\boldsymbol{A},\boldsymbol{B}} d\mu_0 - \int h_1^{\boldsymbol{A},\boldsymbol{B}} d\mu_1,$$

where $\boldsymbol{B} \in \mathbb{R}^{d_0\times d_1}$. Integrating the first equation in (4.12) w.r.t. μ_0 while using $\frac{d\pi_{\boldsymbol{A}}}{\mu_0\otimes\mu_1}(x,y) = e^{\frac{\varphi_0^{\boldsymbol{A}}(x)+\varphi_1^{\boldsymbol{A}}(y)-c_{\boldsymbol{A}}(x,y)}{\varepsilon}}$, yields

$$\int \left(h_0^{\boldsymbol{A},\boldsymbol{B}}(x) + h_1^{\boldsymbol{A},\boldsymbol{B}}(y)\right) d\pi_{\boldsymbol{A}}(x,y) = \int h_0^{\boldsymbol{A},\boldsymbol{B}} d\mu_0 + \int h_1^{\boldsymbol{A},\boldsymbol{B}} d\mu_1 = 32 \int x^\intercal \boldsymbol{B} y \, d\pi_{\boldsymbol{A}}(x,y), \tag{4.13}$$

whence

$$D(\mathsf{OT}_{(\cdot),\varepsilon}(\mu_0,\mu_1))_{[\boldsymbol{A}]}(\boldsymbol{B}) = -32 \int x^\intercal \boldsymbol{B} y \, d\pi_{\boldsymbol{A}}(x,y).$$

As $\|\boldsymbol{A}\|_F^2 = \mathrm{tr}(\boldsymbol{A}^\intercal \boldsymbol{A})$, we have $D(32\|\cdot\|_F^2)_{[\boldsymbol{A}]}(\boldsymbol{B}) = 64\mathrm{tr}(\boldsymbol{A}^\intercal \boldsymbol{B})$, which together with the display above yields

$$D\Phi_{[\boldsymbol{A}]}(\boldsymbol{B}) = 64\,\mathrm{tr}(\boldsymbol{A}^\intercal \boldsymbol{B}) - 32 \int x^\intercal \boldsymbol{B} y \, d\pi_{\boldsymbol{A}}(x,y),$$

as desired.

For the second-order derivative, recall from (4.4) that $\frac{d\pi_{\boldsymbol{A}}}{d\mu_0\otimes\mu_1}(x,y) = e^{\frac{\varphi_0^{\boldsymbol{A}}(x)+\varphi_1^{\boldsymbol{A}}(y)-c_{\boldsymbol{A}}(x,y)}{\varepsilon}}$. As in the proof of Lemma 33, as the map

$$\boldsymbol{A} \in \mathbb{R}^{d_0\times d_1} \mapsto \left((x,y) \in S_0 \times S_1 \mapsto \varphi_0^{\boldsymbol{A}}(x) + \varphi_1^{\boldsymbol{A}}(y) - c_{\boldsymbol{A}}(x,y)\right) \in C(S_0 \times S_1)$$

is differentiable at $\boldsymbol{A} \in \mathbb{R}^{d_0\times d_1}$ with derivative

$$\boldsymbol{C} \in \mathbb{R}^{d_0\times d_1} \mapsto \left((x,y) \in S_0 \times S_1 \mapsto -\left(h_0^{\boldsymbol{A},\boldsymbol{C}}(x) + h_1^{\boldsymbol{A},\boldsymbol{C}}(y) - 32x^\intercal \boldsymbol{C} y\right)\right) \in C(S_0 \times S_1),$$

the expansion

$$\frac{d\pi_{\boldsymbol{A}+\boldsymbol{C}}}{d\mu_0\otimes\mu_1}(x,y) - \frac{d\pi_{\boldsymbol{A}}}{d\mu_0\otimes\mu_1}(x,y) = -\varepsilon^{-1} z_{\boldsymbol{A},\boldsymbol{C}}(x,y) \frac{d\pi_{\boldsymbol{A}}}{d\mu_0\otimes\mu_1}(x,y) + R_{\boldsymbol{C}}(x,y),$$

holds uniformly over $(x,y) \in S_0 \times S_1$, where $R_{\boldsymbol{C}}(x,y) = o(\boldsymbol{C})$ as $\|\boldsymbol{C}\|_F \to 0$ and $z_{\boldsymbol{A},\boldsymbol{C}}(x,y) = h_0^{\boldsymbol{A},\boldsymbol{C}}(x) + h_1^{\boldsymbol{A},\boldsymbol{C}}(y) - 32x^\intercal \boldsymbol{C} y$. Thus,

$$
\begin{aligned}
&\sup_{\|\boldsymbol{B}\|_F=1} \frac{\left|\int x^\intercal \boldsymbol{B} y \, d\pi_{\boldsymbol{A}+\boldsymbol{C}}(x,y) - \int x^\intercal \boldsymbol{B} y \, d\pi_{\boldsymbol{A}}(x,y) - \varepsilon^{-1} \int x^\intercal \boldsymbol{B} y z_{\boldsymbol{A},\boldsymbol{C}}(x,y) d\pi_{\boldsymbol{A}}(x,y)\right|}{\|\boldsymbol{C}\|_F} \\
&= \sup_{\|\boldsymbol{B}\|_F=1} \left| \int x^\intercal \boldsymbol{B} y \|\boldsymbol{C}\|_F^{-1} R_{\boldsymbol{C}}(x,y) d\mu_0 \otimes \mu_1(x,y) \right| \\
&\leq \sup_{(x,y)\in S_1 \times S_2} \|x\|\|y\| \int \|\boldsymbol{C}\|_F^{-1} \left| R_{\boldsymbol{C}}(x,y) \right| d\mu_0 \otimes \mu_1(x,y).
\end{aligned}
$$

As $R_{\boldsymbol{C}}(x,y) = o(\boldsymbol{C})$, this final term converges to 0 as $\|\boldsymbol{C}\|_F \to 0$, so

$$
D^2(\mathsf{OT}_{(\cdot),\varepsilon}(\mu_0,\mu_1))_{[\boldsymbol{A}]}(\boldsymbol{B},\boldsymbol{C}) = 32\varepsilon^{-1} \int x^\intercal \boldsymbol{B} y \left(h_0^{\boldsymbol{A},\boldsymbol{C}}(x) + h_1^{\boldsymbol{A},\boldsymbol{C}}(y) - 32 x^\intercal \boldsymbol{C} y \right) d\pi_{\boldsymbol{A}}(x,y).
$$

As $D(32\|\cdot\|_F^2)_{[\boldsymbol{A}]}(\boldsymbol{B}) = 64\mathrm{tr}(\boldsymbol{A}^\intercal \boldsymbol{B})$, $D^2(32\|\cdot\|_F^2)_{[\boldsymbol{A}]}(\boldsymbol{B},\boldsymbol{C}) = 64\mathrm{tr}(\boldsymbol{C}^\intercal \boldsymbol{B})$. Altogether,

$$
D^2\Phi_{[\boldsymbol{A}]}(\boldsymbol{B},\boldsymbol{C}) = 64\,\mathrm{tr}(\boldsymbol{B}^\intercal \boldsymbol{C}) + 32\varepsilon^{-1} \int x^\intercal \boldsymbol{B} y \left(h_0^{\boldsymbol{A},\boldsymbol{C}}(x) + h_1^{\boldsymbol{A},\boldsymbol{C}}(y) - 32 x^\intercal \boldsymbol{C} y \right) d\pi_{\boldsymbol{A}}(x,y).
$$

Coercivity of Φ is due to nonnegativity of the KL divergence along with the Cauchy-Schwarz inequality,

$$
\begin{aligned}
\mathsf{OT}_{\boldsymbol{A},\varepsilon}(\mu_0,\mu_1) &= \inf_{\pi\in\Pi(\mu_0,\mu_1)} \left\{ \int -4\|x\|^2\|y\|^2 - 32x^\intercal \boldsymbol{A} y \, d\pi(x,y) + \varepsilon \mathsf{D}_{\mathsf{KL}}(\pi\|\mu_0\otimes\mu_1) \right\}, \\
&\geq \inf_{\pi\in\Pi(\mu_0,\mu_1)} \left\{ \int -4\|x\|^2\|y\|^2 - 32\|\boldsymbol{A}\|_F \|x\|\|y\| d\pi(x,y) \right\} \\
&\geq -4\sqrt{M_4(\mu_0)M_4(\mu_1)} - 32\|\boldsymbol{A}\|_F \sqrt{M_2(\mu_0)M_2(\mu_1)},
\end{aligned}
$$

such that $\Phi(\boldsymbol{A}) = 32\|\boldsymbol{A}\|_F^2 + \mathsf{OT}_{\boldsymbol{A},\varepsilon}(\mu_0,\mu_1) \to +\infty$ as $\|\boldsymbol{A}\|_F \to \infty$. □

4.5.2 Proof of Corollary 6

Item (i). The expression for the stationary points follows immediately from Proposition 10. To see that all stationary points are elements of $\mathcal{D}_{M_{\mu_0,\mu_1}}$, observe that if $\boldsymbol{A}$ is a stationary point, then

$$\|\boldsymbol{A}\|_F = \frac{1}{2}\left\|\int xy^\top \, d\pi_{\boldsymbol{A}}(x,y)\right\| \le \frac{1}{2}\int \|x\|\|y\| d\pi_{\boldsymbol{A}}(x,y) \le \frac{1}{2}\sqrt{M_2(\mu_0)M_2(\mu_1)},$$

where the first inequality is due to Jensen's inequality, and the second is due to the Cauchy-Schwarz inequality.

Item (ii). As discussed in Section 4.2.2, if $\pi_\star$ is optimal for S_ε then $\frac{1}{2}\int xy^\top \, d\pi_\star(x,y)$ minimizes Φ. On the other hand, if $\boldsymbol{A}$ minimizes Φ, then we have $\boldsymbol{A} = \frac{1}{2}\int xy^\top \, d\pi_{\boldsymbol{A}}$ and hence

$$\begin{aligned}
\mathsf{S}^2_\varepsilon(\mu_0,\mu_1) &= 8\left\|\int xy^\top \, d\pi_{\boldsymbol{A}}(x,y)\right\|_F^2 - 4\int \|x\|^2\|y\|^2 d\pi_{\boldsymbol{A}}(x,y) \\
&\quad - 32\left\langle \frac{1}{2}\int xy^\top \, d\pi_{\boldsymbol{A}}, \int xy^\top \, d\pi_{\boldsymbol{A}} \right\rangle_F + \varepsilon \mathsf{D}_{\mathsf{KL}}(\pi_{\boldsymbol{A}}\|\mu_0\otimes\mu_1) \\
&= -4\int \|x\|^2\|y\|^2 d\pi_{\boldsymbol{A}(x,y)} - 8\left\|\int xy^\top \, d\pi_{\boldsymbol{A}}(x,y)\right\|_F^2 + \varepsilon \mathsf{D}_{\mathsf{KL}}(\pi_{\boldsymbol{A}}\|\mu_0\otimes\mu_1).
\end{aligned}$$

By (4.6),

$$\begin{aligned}
\mathsf{S}_\varepsilon(\mu_0,\mu_1) &= \mathsf{S}_\varepsilon(\mu_0,\mu_1) + \mathsf{S}^2_\varepsilon(\mu_0,\mu_1) \\
&= \int \left|\|x-x'\|^2 - \|y-y'\|^2\right|^2 + 2\|x-x'\|^2\|y-y'\|^2 d\pi_{\boldsymbol{A}}\otimes\pi_{\boldsymbol{A}}(x,y,x',y') \\
&\quad - 4\int \|x\|^2\|y\|^2 d\mu_0\otimes\mu_1(x,y) - 4\int \|x\|^2\|y\|^2 d\pi_{\boldsymbol{A}}(x,y) \\
&\quad - 8\left\|\int xy^\top \, d\pi_{\boldsymbol{A}}(x,y)\right\|_F^2 + \varepsilon \mathsf{D}_{\mathsf{KL}}(\pi_{\boldsymbol{A}}\|\mu_0\otimes\mu_1).
\end{aligned} \tag{4.14}$$

As $\|x-x'\|^2\|y-y'\|^2 = \left(\|x\|^2 - 2x^\top x' + \|x'\|^2\right)\left(\|y\|^2 - 2y^\top y' + \|y'\|^2\right)$, we have

$$\begin{aligned}
&\int \|x-x'\|^2\|y-y'\|^2 d\pi_{\boldsymbol{A}}\otimes\pi_{\boldsymbol{A}}(x,y,x',y') \\
&= 2\int \|x\|^2\|y\|^2 d\mu_0\otimes\mu_1(x,y) + 2\int \|x\|^2\|y\|^2 d\pi_{\boldsymbol{A}}(x,y) \\
&\quad + 4\int x^\top x' y^\top y' \, d\pi_{\boldsymbol{A}}\otimes\pi_{\boldsymbol{A}}(x,y,x',y'),
\end{aligned}$$

which, together with (4.14) yields

$$\mathsf{S}_\varepsilon(\mu_0,\mu_1) = \int \left|\|x-x'\|^2 - \|y-y'\|^2\right|^2 d\pi_{\boldsymbol{A}}\otimes\pi_{\boldsymbol{A}}(x,y,x',y') + \varepsilon \mathsf{D}_{\mathsf{KL}}(\pi_{\boldsymbol{A}}\|\mu_0\otimes\mu_1),$$

proving optimality of $\pi_{\boldsymbol{A}}$.

Item (iii). Suppose S_ε admits a unique optimal coupling. If two matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ minimize Φ, then $\pi_{\boldsymbol{A}} = \pi_{\boldsymbol{B}}$ by uniqueness, so $\boldsymbol{A} = \frac{1}{2}\int xy^\intercal \, d\pi_{\boldsymbol{A}}(x,y) = \frac{1}{2}\int xy^\intercal \, d\pi_{\boldsymbol{B}}(x,y) = \boldsymbol{B}$. Conversely, suppose Φ admits a unique minimizer $\boldsymbol{A}^\star$. If π is optimal for S_ε, then π solves the EOT problem $\mathsf{OT}_{\boldsymbol{A}^\star,\varepsilon}(\mu_0,\mu_1)$, so $\pi = \pi_{\boldsymbol{A}^\star}$. □

4.5.3 Proof of Corollary 7

We first prove Item (i). The minimal eigenvalue of $D^2\Phi_{[\boldsymbol{A}]}$ is given in variational form as

$$\begin{aligned}
&\inf_{\|\boldsymbol{C}\|_F=1} D^2\Phi_{[\boldsymbol{A}]}(\boldsymbol{C},\boldsymbol{C}) \\
&= \inf_{\|\boldsymbol{C}\|_F=1}\left\{64\|\boldsymbol{C}\|_F^2 + 32\varepsilon^{-1}\int x^\intercal \boldsymbol{C} y\left(h_0^{\boldsymbol{A},\boldsymbol{C}}(x) + h_1^{\boldsymbol{A},\boldsymbol{C}}(y) - 32x^\intercal \boldsymbol{C} y\right) d\pi_{\boldsymbol{A}}(x,y)\right\} \\
&\geq 64 + 32\varepsilon^{-1}\inf_{\|\boldsymbol{C}\|_F=1}\left\{\int x^\intercal \boldsymbol{C} y\left(h_0^{\boldsymbol{A},\boldsymbol{C}}(x) + h_1^{\boldsymbol{A},\boldsymbol{C}}(y) - 32x^\intercal \boldsymbol{C} y\right) d\pi_{\boldsymbol{A}}(x,y)\right\},
\end{aligned}$$

using the formula for $D^2\Phi_{[\boldsymbol{A}]}$ from Proposition 10. Recall that $(h_0^{\boldsymbol{A},\boldsymbol{C}}, h_1^{\boldsymbol{A},\boldsymbol{C}})$ satisfy

$$\begin{aligned}
&\int \left(h_0^{\boldsymbol{A},\boldsymbol{C}}(x) + h_1^{\boldsymbol{A},\boldsymbol{C}}(y) - 32x^\intercal \boldsymbol{C} y\right) e^{\frac{\varphi_0^{\boldsymbol{A}}(x)+\varphi_1^{\boldsymbol{A}}(y)-c_{\boldsymbol{A}}(x,y)}{\varepsilon}} d\mu_1(y) = 0, \quad \forall\, x \in \operatorname{spt}(\mu_0), \\
&\int \left(h_0^{\boldsymbol{A},\boldsymbol{C}}(x) + h_1^{\boldsymbol{A},\boldsymbol{C}}(y) - 32x^\intercal \boldsymbol{C} y\right) e^{\frac{\varphi_0^{\boldsymbol{A}}(x)+\varphi_1^{\boldsymbol{A}}(y)-c_{\boldsymbol{A}}(x,y)}{\varepsilon}} d\mu_0(x) = 0, \quad \forall\, y \in \operatorname{spt}(\mu_1),
\end{aligned}$$

such that, multiplying the top equation by $h_0^{\boldsymbol{A},\boldsymbol{C}}$ and integrating w.r.t. μ_0 and performing the same operations on the lower equation with $h_1^{\boldsymbol{A},\boldsymbol{C}}$ and μ_1 respectively,

$$\begin{aligned}
&\int \left[\left(h_0^{\boldsymbol{A},\boldsymbol{C}}(x)\right)^2 + h_1^{\boldsymbol{A},\boldsymbol{C}}(y)h_0^{\boldsymbol{A},\boldsymbol{C}}(x) - 32x^\intercal \boldsymbol{C} y h_0^{\boldsymbol{A},\boldsymbol{C}}(x)\right] d\pi_{\boldsymbol{A}}(x,y) = 0, \\
&\int \left[h_0^{\boldsymbol{A},\boldsymbol{C}}(x)h_1^{\boldsymbol{A},\boldsymbol{C}}(y) + \left(h_1^{\boldsymbol{A},\boldsymbol{C}}(y)\right)^2 - 32x^\intercal \boldsymbol{C} y h_1^{\boldsymbol{A},\boldsymbol{C}}(y)\right] d\pi_{\boldsymbol{A}}(x,y) = 0.
\end{aligned}$$

Summing these equations gives

$$32\int x^\intercal \boldsymbol{C} y\left(h_0^{\boldsymbol{A},\boldsymbol{C}}(x) + h_1^{\boldsymbol{A},\boldsymbol{C}}(y)\right) d\pi_{\boldsymbol{A}}(x,y) = \int \left(h_0^{\boldsymbol{A},\boldsymbol{C}}(x) + h_1^{\boldsymbol{A},\boldsymbol{C}}(y)\right)^2 d\pi_{\boldsymbol{A}}(x,y),$$

such that

$$\begin{aligned}
&32\int x^\intercal \boldsymbol{C} y\left(h_0^{\boldsymbol{A},\boldsymbol{C}}(x)+h_1^{\boldsymbol{A},\boldsymbol{C}}(y)-32x^\intercal \boldsymbol{C} y\right)d\pi_{\boldsymbol{A}}(x,y)\\
&=\int\left(h_0^{\boldsymbol{A},\boldsymbol{C}}(x)+h_1^{\boldsymbol{A},\boldsymbol{C}}(y)\right)^2 d\pi_{\boldsymbol{A}}(x,y)-32^2\int (x^\intercal \boldsymbol{C} y)^2 d\pi_{\boldsymbol{A}}(x,y),
\end{aligned}$$

which proves the first part of Item (i). It remains to show that

$$\int\left(h_0^{\boldsymbol{A},\boldsymbol{C}}(x)+h_1^{\boldsymbol{A},\boldsymbol{C}}(y)\right)^2 d\pi_{\boldsymbol{A}}(x,y)-32^2\int (x^\intercal \boldsymbol{C} y)^2 d\pi_{\boldsymbol{A}}(x,y)\geq -32^2\mathrm{Var}_{\pi_{\boldsymbol{A}}}[X^\intercal \boldsymbol{C} Y].$$

By Jensen's inequality, we have

$$\begin{aligned}
\int\left(h_0^{\boldsymbol{A},\boldsymbol{C}}(x)+h_1^{\boldsymbol{A},\boldsymbol{C}}(y)\right)^2 d\pi_{\boldsymbol{A}}(x,y)&\geq\left(\int h_0^{\boldsymbol{A},\boldsymbol{C}}(x)+h_1^{\boldsymbol{A},\boldsymbol{C}}(y)d\pi_{\boldsymbol{A}}(x,y)\right)^2\\
&=32^2\left(\int x^\intercal \boldsymbol{C} y\, d\pi_{\boldsymbol{A}}(x,y)\right)^2,
\end{aligned}$$

where the equality follows from (4.13), proving the desired inequality.

To prove the uniform bound on the variance in Item (i), observe that

$$\begin{aligned}
\sup_{\|\boldsymbol{C}\|_F=1}\mathrm{Var}_{\pi_{\boldsymbol{A}}}[X^\intercal \boldsymbol{C} Y]&\leq\sup_{\|\boldsymbol{C}\|_F=1}\mathbb{E}_{\pi_{\boldsymbol{A}}}[(X^\intercal \boldsymbol{C} Y)^2]\\
&\leq\sup_{\|\boldsymbol{C}\|_F=1}\|\boldsymbol{C}\|_F^2\int\|x\|^2\|y\|^2 d\pi_{\boldsymbol{A}}(x,y),\\
&\leq\sqrt{M_4(\mu_0)M_4(\mu_1)}
\end{aligned}$$

where the final two inequalities follow from the Cauchy-Schwarz inequality.

We now prove the upper bound on the maximum eigenvalue of $D^2\Phi_{[\boldsymbol{A}]}$ from Item (ii) again using its variational characterization,

$$\lambda_{\max}\left(D^2\Phi_{[\boldsymbol{A}]}\right)=\sup_{\|\boldsymbol{C}\|_F=1}D^2\Phi_{[\boldsymbol{A}]}(\boldsymbol{C},\boldsymbol{C})=64+\lambda_{\max}\left(D^2\mathsf{OT}_{(\cdot),\varepsilon}(\mu_0,\mu_1)_{[\boldsymbol{A}]}\right).$$

Observe that $\mathsf{OT}_{\boldsymbol{A},\varepsilon}(\mu_0,\mu_1)=\inf_{\pi\in\Pi(\mu_0,\mu_1)}g(\boldsymbol{A},\pi,\mu_0,\mu_1,\varepsilon)$, where g depends on $\boldsymbol{A}$ only through the term $32\mathrm{tr}(\boldsymbol{A}^\intercal\int xy^\intercal\, d\pi(x,y))$ which is linear in $\boldsymbol{A}$. It follows from, e.g., Proposition 2.1.2 in [115] that $\mathsf{OT}_{(\cdot),\varepsilon}(\mu_0,\mu_1)$ is concave. As such, $\lambda_{\max}\left(D^2\mathsf{OT}_{(\cdot),\varepsilon}(\mu_0,\mu_1)_{[\boldsymbol{A}]}\right)\leq 0$, so $\lambda_{\max}\left(D^2\Phi_{[\boldsymbol{A}]}\right)\leq 64$. □

4.5.4 Proof of Theorem 19

We first discuss the convexity properties of Φ. By Corollary 7, $\lambda_{\min}\left(D^2\Phi_{[\boldsymbol{A}]} + \frac{\rho}{2}\|\boldsymbol{A}\|_F^2\right) \geq 64 - 32^2\varepsilon^{-1}\sqrt{M_4(\mu_0)M_4(\mu_1)} + \rho$ for any $\boldsymbol{A} \in \mathbb{R}^{d_0\times d_1}$ and $\rho \geq 0$. When this lower bound is nonnegative, Φ is ρ-weakly convex on $\mathbb{R}^{d_0\times d_1}$ by definition; recall Footnote 5. It follows that Φ is always ρ-weakly convex for $\rho = 32^2\varepsilon^{-1}\sqrt{M_4(\mu_0)M_4(\mu_1)} - 64$. Moreover, if $\sqrt{M_4(\mu_0)M_4(\mu_1)} < \frac{\varepsilon}{16}$, then $\lambda_{\min}\left(D^2\Phi_{[A]}\right) > 0$ such that Φ is strictly convex.

L-smoothness of Φ follows from the mean value inequality (see e.g. Example 2 p. 356 in [8])

$$\begin{aligned}\|D\Phi_{[\boldsymbol{A}]} - D\Phi_{[\boldsymbol{B}]}\|_F &\leq \sup_{\boldsymbol{C}\in[\boldsymbol{A},\boldsymbol{B}]}\sup_{\|\boldsymbol{E}\|_F=1}\left|D^2\Phi_{[\boldsymbol{C}]}(\boldsymbol{A}-\boldsymbol{B},\boldsymbol{E})\right|,\\ &\leq \sup_{\boldsymbol{C}\in[\boldsymbol{A},\boldsymbol{B}]}\left(\left|\lambda_{\min}\left(D^2\Phi_{[\boldsymbol{C}]}\right)\right| \vee \left|\lambda_{\max}\left(D^2\Phi_{[\boldsymbol{C}]}\right)\right|\right)\|\boldsymbol{A}-\boldsymbol{B}\|_F,\end{aligned}$$

for any $\boldsymbol{A},\boldsymbol{B} \in \mathbb{R}^{d_0\times d_1}$, where $[\boldsymbol{A},\boldsymbol{B}]$ denotes the line segment connecting $\boldsymbol{A}$ and $\boldsymbol{B}$. The claimed result then follows by noting that, for any $\boldsymbol{A},\boldsymbol{B} \in \mathcal{D}_M$, $[\boldsymbol{A},\boldsymbol{B}] \subset \mathcal{D}_M$ by convexity and the supremum over $\mathcal{D}_M$ is achieved by compactness and the fact that the objective is continuous. Indeed, the maps $\lambda_{\max}(\cdot), \lambda_{\min}(\cdot)$ are continuous on the space of symmetric matrices, and $D^2\Phi_{[\cdot]}$ is continuous as Φ is smooth. □

4.5.5 Proof of Theorem 20

In this section, we show that Theorem 2.2 in [55] on the convergence rate of Algorithm 5 is applicable in our setting. We particularize their result to a fixed prox-function $d = \frac{1}{2}\|\cdot\|_F^2$ which is smooth and 1-strongly convex.

First, we justify the expressions for the iterates $\boldsymbol{B}_k, \boldsymbol{C}_k$ in Algorithm 5, which are defined in [55] as the proximal operators

$$\boldsymbol{B}_k = \underset{\boldsymbol{V}\in\mathcal{D}_M}{\operatorname{argmin}}\left\{\operatorname{tr}\left(\boldsymbol{G}_k^\intercal \boldsymbol{V}\right) + \frac{L}{2}\|\boldsymbol{V}-\boldsymbol{A}_k\|_F^2\right\},$$
$$\boldsymbol{C}_k = \underset{\boldsymbol{V}\in\mathcal{D}_M}{\operatorname{argmin}}\left\{\operatorname{tr}\left(\boldsymbol{W}_k^\intercal \boldsymbol{V}\right) + \frac{L}{2}\|\boldsymbol{V}\|_F^2\right\}.$$

Rearranging terms, both problems can be written, equivalently, as

$$\underset{\boldsymbol{V}\in\mathcal{D}_M}{\operatorname{argmin}}\left\{\|\boldsymbol{V}-\boldsymbol{U}\|_F^2\right\}, \tag{4.15}$$

for $\boldsymbol{U} = \boldsymbol{A}_k - L^{-1}\boldsymbol{G}_k$ and $\boldsymbol{U} = -L^{-1}\boldsymbol{W}_k$ for the $\boldsymbol{B}_k$ and $\boldsymbol{C}_k$ iterations respectively. The solution of (4.15) is given by $\boldsymbol{V} = \boldsymbol{U}$ if $\boldsymbol{V} = \boldsymbol{U} \in \mathcal{D}_M$, and $\frac{M}{2}\boldsymbol{U}/\|\boldsymbol{U}\|_F$ otherwise.

Next, we show that our notion of δ-oracle yields a δ'-approximate gradient in the sense of Equation (2.3) in [55]. Precisely, we prove that

$$\left|\operatorname{tr}\left(\left(\widetilde{D}\Phi_{[\boldsymbol{A}]} - D\Phi_{[\boldsymbol{A}]}\right)^\intercal(\boldsymbol{B}-\boldsymbol{C})\right)\right| \le \delta', \tag{4.16}$$

for any $\boldsymbol{A},\boldsymbol{B},\boldsymbol{C} \in \mathcal{D}_M$. By the Cauchy-Schwarz inequality,

$$\left|\operatorname{tr}\left(\left(\widetilde{D}\Phi_{[\boldsymbol{A}]} - D\Phi_{[\boldsymbol{A}]}\right)^\intercal(\boldsymbol{B}-\boldsymbol{C})\right)\right| \le M\left\|\widetilde{D}\Phi_{[\boldsymbol{A}]} - D\Phi_{[\boldsymbol{A}]}\right\|_F.$$

Recall that

$$\widetilde{D}\Phi_{[\boldsymbol{A}]} - D\Phi_{[\boldsymbol{A}]} = 32\sum_{\substack{1\le i\le N_0\\1\le j\le N_1}} x^{(i)}\left(y^{(j)}\right)^\intercal\left(\widetilde{\boldsymbol{\Pi}}_{ij}^{\boldsymbol{A}} - \boldsymbol{\Pi}_{ij}^{\boldsymbol{A}}\right),$$

where $\left\|\widetilde{\boldsymbol{\Pi}}^{\boldsymbol{A}} - \boldsymbol{\Pi}^{\boldsymbol{A}}\right\|_\infty < \delta$ uniformly in $\boldsymbol{A} \in \mathcal{D}_M$ by the δ-oracle assumption such that

$$\left\|\widetilde{D}\Phi_{[\boldsymbol{A}]} - D\Phi_{[\boldsymbol{A}]}\right\|_F \le 32\left\|\widetilde{\boldsymbol{\Pi}}^{\boldsymbol{A}} - \boldsymbol{\Pi}^{\boldsymbol{A}}\right\|_\infty \sum_{\substack{1\le i\le N_0\\1\le j\le N_1}}\left\|x^{(i)}\left(y^{(j)}\right)^\intercal\right\|_F < 32\delta\sum_{\substack{1\le i\le N_0\\1\le j\le N_1}}\left\|x^{(i)}\right\|\left\|y^{(j)}\right\|. \tag{4.17}$$

Combining the displayed equations yields

$$\left|\operatorname{tr}\left(\left(\widetilde{D}\Phi_{[\boldsymbol{A}]} - D\Phi_{[\boldsymbol{A}]}\right)^\intercal(\boldsymbol{B}-\boldsymbol{C})\right)\right| \le 32M\delta\sum_{\substack{1\le i\le N_0\\1\le j\le N_1}}\left\|x^{(i)}\right\|\left\|y^{(j)}\right\| = \delta',$$

proving (4.16).

With these preparations Theorem 20 follows from Theorem 2.2 in [55] and the discussion following its proof, noting that $\sum_{i=0}^{k} \frac{i+1}{2} = \frac{(k+1)(k+2)}{4}$. □

4.5.6 Proof of Corollary 8

As $\boldsymbol{A}_k, \boldsymbol{B}_k$ be iterates from Algorithm 6 with $\boldsymbol{B}_k \in \operatorname{int}(\mathcal{D}_M)$ such that $\boldsymbol{B}_k = \boldsymbol{A}_k - \beta_k \widetilde{D}\Phi_{[\boldsymbol{A}_k]}$ by definition. By the triangle inequality,

$$\begin{aligned} \|D\Phi_{[\boldsymbol{A}_k]}\|_F &\leq \|D\Phi_{[\boldsymbol{A}_k]} - \widetilde{D}\Phi_{[\boldsymbol{A}_k]}\|_F + \|\widetilde{D}\Phi_{[\boldsymbol{A}_k]}\|_F \\ &= \|D\Phi_{[\boldsymbol{A}_k]} - \widetilde{D}\Phi_{[\boldsymbol{A}_k]}\|_F + \|\beta_k^{-1}\left(\boldsymbol{B}_k - \boldsymbol{A}_k\right)\|_F. \end{aligned}$$

(4.17) further yields

$$\|D\Phi_{[\boldsymbol{A}_k]} - \widetilde{D}\Phi_{[\boldsymbol{A}_k]}\|_F < 32\delta \sum_{\substack{1\leq i\leq N_0 \\ 1\leq j\leq N_1}} \|x^{(i)}\|\|y^{(j)}\|,$$

proving the claim. □

4.5.7 Proof of Proposition 12

Let $S_i = \operatorname{spt}(\mu_i)$ for $i = 0, 1$. Recall that S_0, S_1 are compact by assumption. The proof of Proposition 12 follows by verifying optimality conditions for minimizing locally Lipschitz functions using the Clarke subdifferential [50]. To this end, we first verify that the objective Φ_0 is locally Lipschitz and prove several auxiliary results to characterize the Clarke subdifferential $\partial\Phi_0$, recalling that the Clarke subdifferential of a locally Lipschitz function $f : \mathbb{R}^{d_0\times d_1} \to \mathbb{R}$ at a point $\boldsymbol{A} \in \mathbb{R}^{d_0\times d_1}$ is given by (cf. e.g. Theorem 2.5.1 in [50])

$$\partial f(\boldsymbol{A}) = \operatorname{conv}\left(\left\{\lim_{n\to\infty} Df_{[\boldsymbol{A}_n]} : U \not\ni \boldsymbol{A}_n \to \boldsymbol{A}\right\}\right), \tag{4.18}$$

where conv(A) denotes the convex hull of the set A, $U \subset \mathbb{R}^{d_0 \times d_1}$ denotes a set of full measure on which f is differentiable, the existence of which is guaranteed by Rademacher's theorem, and we tacitly restrict this definition to convergent sequences of derivatives.

Lemma 36. The function $\boldsymbol{A} \in \mathbb{R}^{d_0 \times d_1} \mapsto \Phi_0(\boldsymbol{A})$ is locally Lipschitz continuous and coercive.

Proof. We start by proving that Φ_0 is locally Lipschitz. Fix a compact set $\boldsymbol{K} \subset \mathbb{R}^{d_0 \times d_1}$ and observe that, for any $\boldsymbol{A}, \boldsymbol{A}' \in \boldsymbol{K}$,

$$\left| \|\boldsymbol{A}\|_F^2 - \|\boldsymbol{A}'\|_F^2 \right| = \left| \|\boldsymbol{A}\|_F - \|\boldsymbol{A}'\|_F \right| (\|\boldsymbol{A}\|_F + \|\boldsymbol{A}'\|_F) \leq 2 \sup_K \| \cdot \|_F \|\boldsymbol{A} - \boldsymbol{A}'\|_F,$$

due to the reverse triangle inequality. This shows that $\| \cdot \|_F^2$ is locally Lipschitz. As for $\mathsf{OT}_{(\cdot),0}(\mu_0, \mu_1)$, let $\pi_{\boldsymbol{A}}, \pi_{\boldsymbol{A}'}$ be solutions of $\mathsf{OT}_{\boldsymbol{A},0}(\mu_0, \mu_1)$, $\mathsf{OT}_{\boldsymbol{A}',0}(\mu_0, \mu_1)$ respectively, then

$$\mathsf{OT}_{\boldsymbol{A},0}(\mu_0, \mu_1) - \mathsf{OT}_{\boldsymbol{A}',0}(\mu_0, \mu_1) \geq \int c_{\boldsymbol{A}} d\pi_{\boldsymbol{A}} - \int c_{\boldsymbol{A}'} d\pi_{\boldsymbol{A}} = -32 \left\langle \boldsymbol{A} - \boldsymbol{A}', \int x y^\intercal d\pi_{\boldsymbol{A}}(x, y) \right\rangle_F, \tag{4.19}$$

and similarly $\mathsf{OT}_{\boldsymbol{A},0}(\mu_0, \mu_1) - \mathsf{OT}_{\boldsymbol{A}',0}(\mu_0, \mu_1) \leq -32 \langle \boldsymbol{A} - \boldsymbol{A}', \int x y^\intercal d\pi_{\boldsymbol{A}'}(x, y) \rangle_F$. Thus,

$$\left| \mathsf{OT}_{\boldsymbol{A},0}(\mu_0, \mu_1) - \mathsf{OT}_{\boldsymbol{A}',0}(\mu_0, \mu_1) \right| \leq 16 M \|\boldsymbol{A} - \boldsymbol{A}'\|_F,$$

recalling that, for any $\pi \in \Pi(\mu_0, \mu_1)$, $\int x y^\intercal d\pi \in \mathcal{D}_M$. This proves that Φ_0 is locally Lipschitz. Coercivity follows by adapting the proof of Proposition 10. $\square$

By Lemma 36 and Rademacher's theorem, Φ_0 is differentiable almost everywhere on $\mathbb{R}^{d_0 \times d_1}$. As the squared Frobenius norm is smooth, we study differentiability of $\mathsf{OT}_{\boldsymbol{A},0}(\mu_0, \mu_1)$. To simplify notation, let $\Pi^\star_{\boldsymbol{A},0}$ denote the set of optimal solutions to $\mathsf{OT}_{\boldsymbol{A},0}(\mu_0, \mu_1)$.

Lemma 37. The function $\boldsymbol{A} \in \mathbb{R}^{d_0 \times d_1} \mapsto -\mathsf{OT}_{\boldsymbol{A},0}(\mu_0, \mu_1)$ is convex; its subdifferential[11] at $\boldsymbol{A}$ contains $\left\{ 32 \int x y^\intercal d\pi : \pi \in \Pi^\star_{\boldsymbol{A},0} \right\}$.

[11]The subdifferential of a convex function $f : \mathbb{R}^{d_0 \times d_1} \mapsto \mathbb{R}$ at $\boldsymbol{A}$ consists of all $\boldsymbol{\Xi} \in \mathbb{R}^{d_0 \times d_1}$ for which $f(\boldsymbol{A}') - f(\boldsymbol{A}) \geq \langle \boldsymbol{A}' - \boldsymbol{A}, \boldsymbol{\Xi} \rangle_F$ for every $\boldsymbol{A}' \in \mathbb{R}^{d_0 \times d_1}$.

Proof. As $\mathsf{OT}_{(\cdot),0}(\mu_0,\mu_1)$ is the infimum of a family of affine functions, it is concave (see Theorem 5.5 in [167]). The second claim follows directly from (4.19). $\square$

With Lemma 37, it is easy to classify the points at which $\mathsf{OT}_{(\cdot),0}(\mu_0,\mu_1)$ is differentiable.

Lemma 38. $-\mathsf{OT}_{(\cdot),0}(\mu_0,\mu_1)$ is differentiable at $\boldsymbol{A}$ with derivative $-D(\mathsf{OT}_{(\cdot),0}(\mu_0,\mu_1))_{[\boldsymbol{A}]} = 32\int xy^\intercal d\pi(x,y)$ for any $\pi\in\Pi^\star_{\boldsymbol{A},0}$ if and only if all couplings in $\Pi^\star_{\boldsymbol{A},0}$ admit the same cross-correlation matrix.

Proof. We recall that a convex function is differentiable at $\boldsymbol{A}$ precisely when its subdifferential at $\boldsymbol{A}$ is a singleton, see Theorem 25.1 in [167]. If the proposed condition on the cross-correlation matrices fails, Lemma 38 implies that the subdifferential is not a singleton, so differentiability fails.

To prove the other direction, assume that all couplings in $\Pi^\star_{\boldsymbol{A},0}$ admit the same cross-correlation matrix. Fix a sequence $\boldsymbol{H}_n\in\mathbb{R}^{d_0\times d_1}\backslash\{\boldsymbol{0}\}$ with $\|\boldsymbol{H}_n\|_F\downarrow 0$. From (4.19), we have, for any $\pi_{\boldsymbol{A}+\boldsymbol{H}_n}\in\Pi^\star_{\boldsymbol{A}+\boldsymbol{H}_n,0}$ and $\pi_{\boldsymbol{A}}\in\Pi^\star_{\boldsymbol{A},0}$,

$$\begin{aligned}\|\boldsymbol{H}_n\|_F^{-1}&\left|\mathsf{OT}_{\boldsymbol{A}+\boldsymbol{H}_n,0}(\mu_0,\mu_1)-\mathsf{OT}_{\boldsymbol{A},0}(\mu_0,\mu_1)+32\left\langle\boldsymbol{H}_n,\int xy^\intercal d\pi_{\boldsymbol{A}}(x,y)\right\rangle_F\right|\\&\leq 32\left\|\int xy^\intercal d\pi_{\boldsymbol{A}+\boldsymbol{H}_n}(x,y)-\int xy^\intercal d\pi_{\boldsymbol{A}}(x,y)\right\|_F,\end{aligned}$$

As $c_{\boldsymbol{A}+\boldsymbol{H}_n}$ converges uniformly to $c_{\boldsymbol{A}}$ on $S_0\times S_1$, for any subsequence n' of n there exists a further subsequence n'' along which $\pi_{\boldsymbol{A}+\boldsymbol{H}_{n''}}\xrightarrow{w}\pi\in\Pi^\star_{\boldsymbol{A},0}$ by Theorem 5.20 in [208]. Since $(x,y)\in S_0\times S_1\mapsto x_iy_j$ is continuous and bounded for any $1\leq i\leq d_0, 1\leq j\leq d_1$, $\left\|\int xy^\intercal d\pi_{\boldsymbol{A}+\boldsymbol{H}_{n''}}(x,y)-\int xy^\intercal d\pi(x,y)\right\|_F = \left\|\int xy^\intercal d\pi_{\boldsymbol{A}+\boldsymbol{H}_{n''}}(x,y)-\int xy^\intercal d\pi_{\boldsymbol{A}}(x,y)\right\|_F\to 0$. As this limit is independent of the choice of subsequence by assumption, it holds along the original sequence $\boldsymbol{H}_n$ such that $D\left(\mathsf{OT}_{(\cdot),0}(\mu_0,\mu_1)\right)_{[\boldsymbol{A}]}=-32\int xy^\intercal d\pi_{\boldsymbol{A}}(x,y)$. $\square$

Those auxiliary results lead to finding the Clarke subdifferential $\partial\Phi_0$, as given below.

Lemma 39. The Clarke subdifferential of Φ_0 at $\boldsymbol{A} \in \mathbb{R}^{d_0 \times d_1}$ is given by

$$\partial\Phi_0(\boldsymbol{A}) = \left\{64\boldsymbol{A} - 32\int xy^\intercal d\pi(x,y) : \pi \in \Pi^\star_{\boldsymbol{A},0}\right\}.$$

Proof. We first characterize the Clarke subdifferential of $\mathsf{OT}_{(\cdot),0}(\mu_0,\mu_1)$ at $\boldsymbol{A}$. By Propositions 2.2.7 and 2.3.1 in [50] along with Lemma 37,

$$\partial\left(\mathsf{OT}_{(\cdot),0}(\mu_0,\mu_1)\right)(\boldsymbol{A}) \supset \left\{-32\int xy^\intercal d\pi(x,y) : \pi \in \Pi^\star_{\boldsymbol{A},0}\right\}. \tag{4.20}$$

As Lemma 38 establishes the set $U \subset \mathbb{R}^{d_0\times d_1}$ on which Φ_0 is differentiable, we study the Clarke subdifferential of Φ_0 through the lens of (4.18). Observe that if $U \not\ni \boldsymbol{A}_n \to \boldsymbol{A}$ and $\pi_{\boldsymbol{A}_n} \in \Pi^\star_{\boldsymbol{A}_n,0}$, then $\pi_{\boldsymbol{A}_n} \stackrel{w}{\to} \pi_{\boldsymbol{A}} \in \Pi^\star_{\boldsymbol{A},0}$ as in the proof of Lemma 38 (up to a subsequence). Moreover,

$$\lim_{n\to\infty} D\left(\mathsf{OT}_{(\cdot),0}(\mu_0,\mu_1)\right)_{[\boldsymbol{A}_n]} = \lim_{n\to\infty} -32\int xy^\intercal d\pi_{\boldsymbol{A}_n}(x,y) = -32\int xy^\intercal d\pi_{\boldsymbol{A}}(x,y)$$

along this subsequence. Thus, if $\boldsymbol{A}_n' \notin U$ converges to $\boldsymbol{A}$ and $\lim_{n\to\infty} D\left(\mathsf{OT}_{(\cdot),0}(\mu_0,\mu_1)\right)_{[\boldsymbol{A}_n']}$ exists, then the limit is given by $-32\int xy^\intercal d\pi(x,y)$ for some $\pi \in \Pi^\star_{\boldsymbol{A},0}$. It is easy to see that $\Pi^\star_{\boldsymbol{A},0}$ is convex, so (4.20) and (4.18) together imply that

$$\partial\left(\mathsf{OT}_{(\cdot),0}(\mu_0,\mu_1)\right)(\boldsymbol{A}) = \left\{-32\int xy^\intercal d\pi(x,y) : \pi \in \Pi^\star_{\boldsymbol{A},0}\right\}.$$

Conclude by applying the subdifferential sum rule (Corollary 1 on p. 39 of [50]). □

Proof of Proposition 12. By Proposition 2.3.2 in [50], if Φ_0 attains a local minimum at $\boldsymbol{A}^\star$, then $0 \in \partial\Phi_0(\boldsymbol{A}^\star)$, i.e., there exists $\pi \in \Pi^\star_{\boldsymbol{A}^\star,0}$ for which $\boldsymbol{A}^\star = \frac{1}{2}\int xy^\intercal d\pi(x,y)$ by Lemma 39. The second result on optimality for the original GW problem follows directly from the proof of Corollary 6. □

4.5.8 Proof of Theorem 22

Throughout, let $\mathcal{X}_i \subset \mathbb{R}^{d_i}$ be a closed ball with finite radius $r > 0$ for which $\mathrm{spt}(\mu_i) \subset \mathcal{X}_i$ for $i = 0, 1$. By the Bolzano-Weierstrass theorem, $\boldsymbol{A}_\varepsilon^\star$ admits a limit point $\boldsymbol{A}_0^\star$ as $\varepsilon \downarrow 0$. Let $\varepsilon_k \downarrow 0$ be a sequence along which $\boldsymbol{A}_{\varepsilon_k}^\star \to \boldsymbol{A}_0^\star$ and define

$$F_k : \pi \in \mathcal{P}(\mathcal{X}_0 \times \mathcal{X}_1) \mapsto \begin{cases} \int c_{\boldsymbol{A}_{\varepsilon_k}^\star} d\pi + \varepsilon_k \mathsf{D}_{\mathsf{KL}}(\pi \| \mu_0 \otimes \mu_1), & \text{if } \pi \in \Pi(\mu_0, \mu_1), \\ +\infty, & \text{otherwise,} \end{cases}$$

$$F : \pi \in \mathcal{P}(\mathcal{X}_0 \times \mathcal{X}_1) \mapsto \begin{cases} \int c_{\boldsymbol{A}_0^\star} d\pi, & \text{if } \pi \in \Pi(\mu_0, \mu_1), \\ +\infty, & \text{otherwise.} \end{cases}$$

Throughout, we endow $\mathcal{P}(\mathcal{X}_0 \times \mathcal{X}_1)$ with the topology induced by the weak convergence of probability measures which is metrized by the 2-Wasserstein distance, W_2, for instance; $(\mathcal{P}(\mathcal{X}_0 \times \mathcal{X}_1), \mathsf{W}_2)$ is notably a separable metric space, see Theorem 6.18 in [208].

We first show that F_k Γ-converges to F according to Definition 2.3 in [29]. To this end, we show that

$$F(\pi) \leq \liminf_{k \to \infty} F_k(\pi_k), \text{ for every } \pi_k \xrightarrow{w} \pi. \tag{4.21}$$

and exhibit a sequence $\pi_k \xrightarrow{w} \pi$ satisfying

$$F(\pi) \geq \limsup_{k \to \infty} F_k(\pi_k). \tag{4.22}$$

To prove (4.21), first assume that $\pi \in \Pi(\mu_0, \mu_1)$. Then, if $\pi_k \xrightarrow{w} \pi$, $F_k(\pi_k) \geq \int c_{\boldsymbol{A}_{\varepsilon_k}^\star} d\pi_k$, by nonnegativity of the KL divergence. As $c_{\boldsymbol{A}_{\varepsilon_k}^\star}$ converges to $c_{\boldsymbol{A}_0^\star}$ uniformly on $\mathcal{X}_0 \times \mathcal{X}_1$ and $c_{\boldsymbol{A}_0^\star}$ is continuous and bounded,

$$\left| \int c_{\boldsymbol{A}_{\varepsilon_k}^\star} d\pi_k - \int c_{\boldsymbol{A}_0^\star} d\pi \right| \leq \left| \int c_{\boldsymbol{A}_{\varepsilon_k}^\star} - c_{\boldsymbol{A}_0^\star} d\pi_k \right| + \left| \int c_{\boldsymbol{A}_0^\star} d\pi - \int c_{\boldsymbol{A}_0^\star} d\pi_k \right| \to 0. \tag{4.23}$$

This shows that, if $\pi \in \Pi(\mu_0, \mu_1)$, then

$$\liminf_{k \to \infty} F_k(\pi_k) \geq \lim_{k \to \infty} \int c_{\boldsymbol{A}_{\varepsilon_k}^\star} d\pi_k = F(\pi).$$

If $\pi \notin \Pi(\mu_0, \mu_1)$, then $\pi_k \notin \Pi(\mu_0, \mu_1)$ for all k sufficiently large as $\Pi(\mu_0, \mu_1)$ is compact for the weak convergence on $\mathcal{P}(\mathcal{X}_0 \times \mathcal{X}_1)$ (cf. e.g. the proof of Theorem 1.4 in [175]) so the bound holds vacuously, proving (4.21).

As for (4.22), if $\pi \notin \Pi(\mu_0, \mu_1)$, then there is nothing to show. For $\pi \in \Pi(\mu_0, \mu_1)$, we consider a block approximation; cf. [37, 87]. First, for $i = 0, 1$, partition $\mathbb{R}^{d_i}$ by hypercubes of length $\ell > 0$,

$$\{H_{i,q,\ell} = [k_1\ell, (q_1 + 1)\ell) \times \cdots \times [q_{d_i}\ell, (q_{d_i} + 1)\ell) : q = (q_1, \ldots, q_{d_i}) \in \mathbb{Z}^{d_i}\},$$

and define the block approximation, π_ℓ, of π by

$$\pi_\ell = \sum_{(q,q') \in \mathbb{Z}^{d_0} \times \mathbb{Z}^{d_1}} \pi_\ell|_{H_{0,q,\ell} \times H_{1,q',\ell}},$$

$$\pi_\ell|_{H_{0,q,\ell} \times H_{1,q',\ell}} = \frac{\pi(H_{0,q,\ell} \times H_{1,q',\ell})}{\mu_0 \otimes \mu_1(H_{0,q,\ell} \times H_{1,q',\ell})} (\mu_0|_{H_{0,q,\ell}} \otimes \mu_1|_{H_{1,q',\ell}}),$$

for every $(q, q') \in \mathbb{Z}^{d_0} \times \mathbb{Z}^{d_1}$ with the convention $\frac{0}{0} = 0$. Here, $\mu_0|_{H_{0,q,\ell}}(A) = \mu_0(A \cap H_{0,q,\ell})$ and similarly for the other restrictions. Of note is that $\pi_\ell \ll \mu_0 \otimes \mu_1$ and that $\pi_\ell \in \Pi(\mu_0, \mu_1)$ (see the discussion surrounding Definition 1 in [87]).

Lemma 40. The block approximation π_ℓ of π converges weakly to π as $\ell \downarrow 0$.

Proof. Let $Q = \left\{(q, q') \in \mathbb{Z}^{d_0} \times \mathbb{Z}^{d_1} : \pi(H_{0,q,\ell} \times H_{1,q',\ell}) > 0\right\}$ and $\gamma = \sum_{(q,q') \in Q} \pi(H_{0,q,\ell} \times H_{1,q',\ell})\gamma_{q,q'}$ where $\gamma_{q,q'}$ is any coupling of the measures $\left(\frac{\pi|_{H_{0,q,\ell} \times H_{1,q',\ell}}}{\pi(H_{0,q,\ell} \times H_{1,q',\ell})}, \frac{\pi_\ell|_{H_{0,q,\ell} \times H_{1,q',\ell}}}{\pi_\ell(H_{0,q,\ell} \times H_{1,q',\ell})}\right)$ with support in $\overline{H_{0,q,\ell} \times H_{1,q',\ell}}$ (the closure of $H_{0,q,\ell} \times H_{1,q',\ell}$). As $\pi_\ell(H_{0,q,\ell} \times H_{1,q',\ell}) = \pi(H_{0,q,\ell} \times H_{1,q',\ell})$ by construction, it is readily seen that $\gamma \in \Pi(\pi, \pi_\ell)$. Thus,

$$\begin{aligned} \mathsf{W}_2^2(\pi, \pi_\ell) \leq \int \|x - y\|^2 d\gamma(x, y) &= \sum_{(q,q') \in Q} \pi(H_{0,q,\ell} \times H_{1,q',\ell}) \int \|x - y\|^2 d\gamma_{q,q'}(x, y) \\ &\leq (d_0 + d_1)\ell^2, \end{aligned}$$

noting that $\operatorname{diam}(H_{0,q,\ell} \times H_{1,q',\ell}) = \sqrt{d_0 + d_1}\ell$ is the diameter of a hypercube in $\mathbb{R}^{d_0+d_1}$ of length ℓ. Conclude that $\mathsf{W}_2(\pi, \pi_\ell) \to 0$ as $\ell \downarrow 0$ such that $\pi_\ell \xrightarrow{w} \pi$. □

Setting $\ell_k = \varepsilon_k$, Lemma 40 yields

$$\lim_{k\to\infty} \int c_{A^\star_{\varepsilon_k}} d\pi_{\varepsilon_k} = \int c_{A^\star_0} d\pi,$$

by a simple adaption of (4.23). It remains to show that $\lim_{k\to\infty} \varepsilon_k \mathsf{D}_{\mathsf{KL}}(\pi_{\varepsilon_k} \| \mu_0 \otimes \mu_1) = 0$ such that

$$\lim_{k\to\infty} F_k(\pi_{\varepsilon_k}) \to F(\pi).$$

Lemma 41. The block approximation π_ℓ of π satisfies $\varepsilon_k \mathsf{D}_{\mathsf{KL}}(\pi_{\varepsilon_k} \| \mu_0 \otimes \mu_1) \to 0$ as $k \to \infty$.

Proof. As $\pi(H_{0,q,\ell} \times H_{1,q',\ell}) \le 1$,

$$\begin{aligned}
\mathsf{D}_{\mathsf{KL}}(\pi^\ell \| \mu_0 \otimes \mu_1) &= \sum_{(q,q')\in\mathbb{Z}^{d_0}\times\mathbb{Z}^{d_1}} \int \log\left(\frac{\pi\left(H_{0,q,\ell} \times H_{1,q',\ell}\right)}{\mu_0 \otimes \mu_1(H_{0,q,\ell} \times H_{1,q',\ell})}\right) d\pi_\ell|_{H_{0,q,\ell}\times H_{1,q',\ell}} \\
&\le \sum_{(q,q')\in\mathbb{Z}^{d_0}\times\mathbb{Z}^{d_1}} \int -\log(\mu_0(H_{0,q,\ell})) - \log(\mu_1(H_{1,q',\ell})) d\pi_\ell|_{H_{0,q,\ell}\times H_{1,q',\ell}} \\
&= \sum_{(q,q')\in\mathbb{Z}^{d_0}\times\mathbb{Z}^{d_1}} \Big(-\log(\mu_0(H_{0,q,\ell})) - \log(\mu_1(H_{1,q',\ell}))\Big) \pi(H_{0,q,\ell} \times H_{1,q',\ell}) \\
&= -\sum_{q\in\mathbb{Z}^{d_0}} \log(\mu_0(H_{0,q,\ell}))\mu_0(H_{0,q,\ell}) - \sum_{q'\in\mathbb{Z}^{d_1}} \log(\mu_1(H_{1,q',\ell}))\mu_1(H_{1,q',\ell}).
\end{aligned} \tag{4.24}$$

Observe that

$$\begin{aligned}
&-\sum_{q\in\mathbb{Z}^{d_0}} \log(\mu_0(H_{0,q,\ell}))\mu_0(H_{0,q,\ell}) \\
&= -\int \log\left(\frac{\mu_0(H_{0,q,\ell})}{\ell^{d_0}} \mathbb{1}_{H_{0,q,\ell}}\right) \frac{\mu_0(H_{0,q,\ell})}{\ell^{d_0}} \mathbb{1}_{H_{0,q,\ell}} d\lambda - d_0 \log(\ell),
\end{aligned}$$

where λ denotes the Lebesgue measure. The first term on the last line is the differential entropy of a probability distribution supported on $\mathcal{X}_{0,\ell} = \cup_{q\in I} H_{0,q,\ell}$, where $I = \left\{q \in \mathbb{Z}^{d_0} : H_{0,q,\ell} \cap \mathcal{X}_0 \neq \emptyset\right\}$. This quantity is maximized among all probability distributions supported on $\mathcal{X}_{0,\ell}$ which are absolutely continuous w.r.t. the Lebesgue measure by the uniform distribution on $\mathcal{X}_{0,\ell}$ with value $d_0 \log(\lambda(\mathcal{X}_{0,\ell}))$. With this and (4.24), we obtain

$$\mathsf{D}_{\mathsf{KL}}(\pi^\ell \| \mu_0 \otimes \mu_1) \le d_0 \log\left(\frac{\lambda(\mathcal{X}_{0,\ell})}{\ell}\right) + d_1 \log\left(\frac{\lambda(\mathcal{X}_{1,\ell})}{\ell}\right).$$

Conclude that

$$0 \le \varepsilon_k \mathsf{D}_{\mathsf{KL}}(\pi_{\varepsilon_k} \| \mu_0 \otimes \mu_1) \le \varepsilon_k \left(d_0 \log \left(\frac{\lambda(\mathcal{X}_{0,\varepsilon_k})}{\varepsilon_k} \right) + d_1 \log \left(\frac{\lambda(\mathcal{X}_{1,\varepsilon_k})}{\varepsilon_k} \right) \right) \to 0,$$

as $\lambda(\mathcal{X}_{i,\varepsilon_k}) \ge \lambda(\mathcal{X}_i) > 0$ for $i = 0, 1$. □

With this, we have shown that F_k Γ-converges to F. Now, let π_{ε_k} minimize F_k. As $\Pi(\mu_0, \mu_1)$ is compact, π_{ε_k} admits a cluster point π_0 which minimizes F by Theorem 2.1 in [29]. Thus, π_0 is an optimal solution of $\mathsf{OT}_{\boldsymbol{A}_0^\star}(\mu_0, \mu_1)$ and, along the subsequence where $\pi_{\varepsilon_k} \xrightarrow{w} \pi_0$,

$$\left\| 64\boldsymbol{A}_0^\star - 32 \int xy^\intercal d\pi_0 \right\|_F = \lim_{k \to \infty} \left\| 64\boldsymbol{A}_{\varepsilon_k}^\star - 32 \int xy^\intercal d\pi_{\varepsilon_k} \right\|_F \le \delta.$$

This concludes the proof of the first result.

For the results pertaining to local/global optimality, let $\varepsilon_k \downarrow 0$ be arbitrary and define

$$G_k : \boldsymbol{A} \in \mathbb{R}^{d_0 \times d_1} \mapsto \begin{cases} \Phi_{\varepsilon_k}, & \text{if } \boldsymbol{A} \in \mathcal{D}_M, \\ +\infty, & \text{otherwise,} \end{cases}, \quad G : \boldsymbol{A} \in \mathbb{R}^{d_0 \times d_1} \mapsto \begin{cases} \Phi_0, & \text{if } \boldsymbol{A} \in \mathcal{D}_M, \\ +\infty, & \text{otherwise.} \end{cases}$$

We now show that G_k Γ-converges to G. Observe that, for any $\boldsymbol{A}_k \to \boldsymbol{A} \in \mathcal{D}_M$,

$$\Phi_{\varepsilon_k}(\boldsymbol{A}_k) = 32\|\boldsymbol{A}_k\|_F^2 + \mathsf{OT}_{\boldsymbol{A}_k, \varepsilon_k}(\mu_0, \mu_1) \ge 32\|\boldsymbol{A}_k\|_F^2 + \mathsf{OT}_{\boldsymbol{A}_k, 0}(\mu_0, \mu_1),$$

where the inequality is due to nonnegativity of the KL divergence. As $c_{\boldsymbol{A}_k} \to c_{\boldsymbol{A}}$ uniformly, $32\|\boldsymbol{A}_k\|_F^2 + \mathsf{OT}_{\boldsymbol{A}_k,0}(\mu_0, \mu_1) \to 32\|\boldsymbol{A}\|_F^2 + \mathsf{OT}_{\boldsymbol{A},0}(\mu_0, \mu_1) = \Phi_0(\boldsymbol{A})$. If $\boldsymbol{A} \notin \mathcal{D}_M$, then $\boldsymbol{A}_k \notin \mathcal{D}_M$ for every k sufficiently large, so the lim inf condition (4.21) holds.

For the lim sup condition (4.22), if $\boldsymbol{A} \notin \mathcal{D}_M$, then the bound is vacuous. If $\boldsymbol{A} \in \mathcal{D}_M$, then let $\pi_{\boldsymbol{A}}$ be an optimal solution of $\mathsf{OT}_{\boldsymbol{A},0}(\mu_0, \mu_1)$ and π_{ε_k} be the block approximation of $\pi_{\boldsymbol{A}}$ with $\ell = \varepsilon_k$. Then,

$$0 \le \mathsf{OT}_{\boldsymbol{A},\varepsilon_k}(\mu_0, \mu_1) - \mathsf{OT}_{\boldsymbol{A},0}(\mu_0, \mu_1) \le \int c_{\boldsymbol{A}} d\pi_{\varepsilon_k} + \varepsilon_k \mathsf{D}_{\mathsf{KL}}(\pi_{\varepsilon_k} \| \mu_0 \otimes \mu_1) - \int c_{\boldsymbol{A}} d\pi_{\boldsymbol{A}}.$$

We have shown previously that the rightmost term converges to 0 as $k \to \infty$, so $\Phi_{\varepsilon_k}(\boldsymbol{A}) \to \Phi_0(\boldsymbol{A})$, such that G_k Γ-converges to G. By Theorem 2.1 in [29], any cluster point of a sequence of minizers of G_k minimizes G, proving the claim.

Finally, consider the case where $(\boldsymbol{A}_{\varepsilon_k})_{k\in\mathbb{N}}$ satisfies the conditions of part 3 of Theorem 22, i.e., $\boldsymbol{A}_{\varepsilon_k} \to \boldsymbol{A}^\star$ and $\boldsymbol{A}_{\varepsilon_k}$ minimizes Φ_{ε_k} on a closed ball of fixed radius $r > 0$ centred at $\boldsymbol{A}^\star$, say $B_r^\star$. A simple adaptation of the previous arguments yields that G_k restricted to $B_r^\star$ Γ-converges to G restricted to $B_r^\star$. As such, $\boldsymbol{A}^\star$ minimizes Φ_0 over $B_r^\star$ and is thus locally minimal for Φ_0.

4.6 Concluding Remarks

This work studied efficient computation of the quadratic EGW alignment problem between Euclidean spaces subject to non-asymptotic convergence guarantees. Despite the availability of various heuristic methods for computing EGW, formal guarantees beyond asymptotic convergence to a stationary point were absent until now. To develop our algorithms, we leveraged the variational form of the EGW distance that ties it to the well-understood EOT problem with a certain parametrized cost function. By analyzing the stability of the variational problem, its convexity and smoothness properties were established, which led to two new efficient algorithms for computing the EGW distance. The complexity of our algorithms agree with the best known complexity of $O(N^2)$ for computing the quadratic EGW distance directly, but unlike previous approaches, our methods are subject to non-asymptotic convergence rate guarantees to global/local solutions in the convex/non-convex regime. As the first derivative of the objective function depends on the solution to an EOT problem which must be solved numerically, we quantify both the error incurred by Sinkhorn's algorithm and the resulting effect on the convergence of both algorithms. Moreover, we establish a suitable notion of convergence of solutions to the variational EGW problem to those of the variational GW problem in the limit of vanishing regularization. Below, we discuss possible extensions and future research direction stemming from this work.

Algorithmic improvements. The stability analysis of the variational problem lays the groundwork for solving the EGW problem via smooth optimization methods. Consequently, improvements or alternatives to the proposed accelerated gradient methods are of great practical interest. For instance, marked improvements can be attained by analyzing the tradeoff between the per iteration cost associated with updating the step size parameter and the resulting decrease in the number of iterations required for convergence. Similarly, establishing sharper bounds on the eigenvalues of the Hessian would improve our characterization of the smoothness and convexity properties of the objective.

Expanding duality theory. To the best of our knowledge, the current duality theory for the EGW problem is limited to the quadratic and inner product costs over Euclidean spaces. As the present work makes heavy use of this duality theory, we anticipate that these results could be extended to the EGW problem with other costs and/or spaces once an adequate duality theory has been established. Non-Euclidean spaces are not only of theoretical, but also of practical interest under the GW paradigm as they allow comparing/aligning important examples such as manifold or graph data.

CHAPTER 5

CONCLUSION

This thesis presented several contributions to the study of statistics and computation for optimal transport-based distances. It introduced a general framework for deriving limit theorems for quantities of interest in optimal transport which was applied to both regularized optimal transport distances and later extended to the Gromov-Wasserstein distance. Beyond the statistical analysis of the Gromov-Wasserstein distance, new algorithms were proposed for its entropically regularized variant, achieving the same per-iteration complexity as existing methods whilst benefiting from first-of-their kind convergence rate guarantees for this problem.

Although the primary focus of this thesis is of a theoretical nature, optimal transport and Gromov-Wasserstein distances have found applications across a range fields. For example, this work explored the use of the Gromov-Wasserstein distance in testing for isomorphisms of distributions on graphs. As computation is still an important bottleneck, improving algorithms for solving the Gromov-Wasserstein problem is expected to make it more accessible to practitioners and hence broaden applications.

Finally, this work identified many promising directions for future research. A key open question in the study of Gromov-Wasserstein distances is the development of a suitable duality theory for arbitrary metric measure spaces. Notably, the variational formulation used herein applies only to the $(2,2)$-Gromov-Wassertein distance when the metrics are induced by inner products and hence cannot be readily extended to the general setting. Resolving this issue would not only advance the theoretical foundations of the distance, but would also enable proving general statistical results and developing new algorithms for Gromov-Wasserstein problems.

APPENDIX A

SUPPLEMENT TO CHAPTER 2

A.1 Proofs for Sections 2.2 and 2.3

A.1.1 Proof of Lemma 2

The lemma follows from Proposition 2.1 in [79]. We include its proof for completeness. Pick any $h \in \mathfrak{D}_0, t_n \to 0, h_n \to h$ in $\mathfrak{D}$ such that $\theta + t_n h_n \in \Theta$. For any subsequence n' of n, there exists a further subsequence n'' such that (i) $t_{n''} > 0$ for all n'' or (ii) $t_{n''} < 0$ for all n''. If (i) holds, then $(\phi(\theta + t_{n''}h_{n''}) - \phi(\theta))/t_{n''} \to \phi'(h)$, while if (ii) holds, then $(\phi(\theta + t_{n''}h_{n''}) - \phi(\theta))/t_{n''} = -(\phi(\theta + (-t_{n''})(-h)) - \phi(\theta))/(-t_{n''}) \to -\phi'(-h) = \phi'(h)$. Since the limit does not depend on the choice of subsequence, we have the result. □

A.1.2 Proof of Proposition 1

Let $\mathcal{P}_{0,\epsilon} = \{\nu \in \mathcal{P}_0 : \|\nu - \mu\|_{\infty,\mathcal{F}} < \epsilon\}$. Since $\mathcal{P}_{0,\epsilon}$ is convex, $\bigcup_{t>0} t(\mathcal{P}_{0,\epsilon} - \mu) = \bigcup_{t>0} t(\mathcal{P}_0 - \mu)$, and $\mu_n \in \mathcal{P}_{0,\epsilon}$ with inner probability approaching one, it suffices to prove the proposition with $\mathcal{P}_0$ replaced by $\mathcal{P}_{0,\epsilon}$, i.e., we may assume without loss of generality that $\delta : \mathcal{P}_0 \to \mathfrak{E}$ is globally Lipschitz w.r.t. $\| \cdot \|_{\infty,\mathcal{F}}$,

$$\|\delta(\nu) - \delta(\nu')\|_{\mathfrak{E}} \le C\|\nu - \nu'\|_{\infty,\mathcal{F}}, \quad \forall \nu, \nu' \in \mathcal{P}_0$$

for some constant $C < \infty$. We apply the extended functional delta method, Lemma 1, by identifying $\mathcal{P}_0$ as a subset of $\ell^\infty(\mathcal{F})$. Consider the map $\tau : \mathcal{P}_0 \ni \nu \mapsto (f \mapsto \int f d\nu)_{f\in\mathcal{F}} \in \ell^\infty(\mathcal{F})$. Define the map $\bar{\delta} : \tau\mathcal{P}_0 \subset \ell^\infty(\mathcal{F}) \to \mathfrak{E}$ by $\bar{\delta}(\tau\nu) = \delta(\nu)$ for $\nu \in \mathcal{P}_0$. The map $\bar{\delta}$ is well-defined since, whenever $\tau\nu = \tau\nu'$, we have $\|\delta(\nu) - \delta(\nu')\|_{\mathfrak{E}} \le C\|\tau\nu - \tau\nu'\|_{\infty,\mathcal{F}} = 0$, i.e., $\delta(\nu) = \delta(\nu')$. In what follows, we identify δ with $\bar{\delta}$.

Under such identification, we apply Lemma 1 with $\mathfrak{D} = \ell^\infty(\mathcal{F}), \Theta = \mathcal{P}_0, \theta = \mu$, and $\phi(\nu) = \delta(\nu)$ for $\nu \in \mathcal{P}_0$. We will verify that $\nu \mapsto \delta(\nu)$ is Hadamard directionally differentiable at $\nu = \mu$ with derivative coinciding with δ'_μ on $\mathcal{P}_0 - \mu$. This (essentially) follows from the fact that δ is Lipschitz continuous together with the fact that it is Gâteaux directionally differentiable at $\nu = \mu$ with derivative δ'_μ; cf. [181], p. 483–484. However, in our application, the Gâteaux derivative is a priori defined only on $\mathcal{P}_0 - \mu$, and to extend δ'_μ to the tangent cone $\mathcal{T}_{\mathcal{P}_0}(\mu)$, we require $\mathfrak{E}$ to be complete. For completeness, we provide its full proof.

Extend δ'_μ to $\bigcup_{t>0} t(\mathcal{P}_0 - \mu)$ as

$$\delta'_\mu(t(\nu - \mu)) = t\delta'_\mu(\nu - \mu),\ t > 0, \nu \in \mathcal{P}_0.$$

This extension is well-defined. Indeed, suppose that $t(\nu - \mu) = t'(\nu' - \mu)$ (as elements of $\ell^\infty(\mathcal{F})$) for some $t' > 0$ and $\nu' \in \mathcal{P}_0$; then

$$\begin{aligned} t'\delta'_\mu(\nu' - \mu) &= t' \lim_{s\downarrow 0} \frac{\delta(\mu + st'(\nu' - \mu)) - \delta(\mu)}{st'} \\ &= \lim_{s\downarrow 0} \frac{\delta(\mu + st'(\nu' - \mu)) - \delta(\mu)}{s} \\ &= \lim_{s\downarrow 0} \frac{\delta(\mu + st(\nu - \mu)) - \delta(\mu)}{s} \\ &= t \lim_{s\downarrow 0} \frac{\delta(\mu + st(\nu - \mu)) - \delta(\mu)}{st} = t\delta'_\mu(\nu - \mu). \end{aligned}$$

Pick any $m_n \in \bigcup_{t>0} t(\mathcal{P}_0 - \mu)$ with $m_n \to m$ in $\ell^\infty(\mathcal{F})$ and $t_n \downarrow 0$. Also, pick any $\epsilon' > 0$ and $m' \in \bigcup_{t>0} t(\mathcal{P}_0 - \mu)$ such that $\|m - m'\|_{\infty,\mathcal{F}} < \epsilon'$. For sufficiently large n, $t_n m_n, t_n m' \in \mathcal{P}_0 - \mu$

(as $\mathcal{P}_0$ is convex), so that

$$\begin{aligned}\|\delta(\mu + t_n m_n) - \delta(\mu) - t_n\delta'_\mu(m')\|_{\mathfrak{E}} &\le \|\delta(\mu + t_n m_n) - \delta(\mu + t_n m')\|_{\mathfrak{E}} \\ &\quad + \underbrace{\|\delta(\mu + t_n m') - \delta(\mu) - t_n\delta'_\mu(m')\|_{\mathfrak{E}}}_{=o(t_n)} \\ &\le Ct_n\|m_n - m'\|_{\infty,\mathcal{F}} + o(t_n) \\ &\le C\epsilon' t_n + o(t_n),\end{aligned}$$

which implies that

$$\limsup_{n,n'\to\infty}\left\|\frac{\delta(\mu + t_n m_n) - \delta(\mu)}{t_n} - \frac{\delta(\mu + t_{n'} m_{n'}) - \delta(\mu)}{t_{n'}}\right\|_{\mathfrak{E}} \le 2C\epsilon'.$$

Since $\epsilon' > 0$ is arbitrary, $t_n^{-1}[\delta(\mu + t_n m_n) - \delta(\mu)]$ is Cauchy, and since $\mathfrak{E}$ is a Banach space, the limit

$$\lim_{n\to\infty}\frac{\delta(\mu + t_n m_n) - \delta(\mu)}{t_n}$$

exists in $\mathfrak{E}$. Thus, we have proved that the map $\nu \mapsto \delta(\nu)$ from $\mathcal{P}_0$ into $\mathfrak{E}$ is Hadamard directionally differentiable at $\nu = \mu$. The conclusion of the proposition then follows from the functional delta method. We note that the second claim that $G_\mu \in \mathcal{T}_{\mathcal{P}_0}(\mu)$ with probability one follows from the portmanteau theorem. □

A.1.3 Proof of Corollary 1

By Lemma 3.1 in [202], $\mathrm{spt}(G_\mu)$ is the closure in $\ell^\infty(\mathcal{F})$ of the reproducing kernel Hilbert space for G_μ and thus a vector subspace of $\ell^\infty(\mathcal{F})$. The second claim follows by Lemma 2. □

A.1.4 Proof of Proposition 2

By direct calculation, for a given $h \in \dot{\mathcal{P}}_{0,\mu}$, the submodel $\{\mu_t : 0 \leq t < \epsilon'\}$ with $\mu_t = (1 + th)\mu$ for sufficiently small $\epsilon' > 0$ is differentiable in quadratic mean at $t = 0$ with score function h, i.e.,

$$\int \left[\frac{d\mu_t^{1/2} - d\mu^{1/2}}{t} - \frac{1}{2} h d\mu^{1/2} \right]^2 \to 0, \quad t \downarrow 0.$$

Thus, $\dot{\mathcal{P}}_{0,\mu}$ is the tangent space at μ corresponding to all such submodels.

By assumption, we see that, for the above submodel μ_t,

$$\frac{\delta(\mu_t) - \delta(\mu)}{t} \to \delta'_\mu(h\mu).$$

Observe that the $L^2(\mu)$ closure of $\dot{\mathcal{P}}_{0,\mu}$ is $L^2_0(\mu) = \{h \in L^2(\mu) : \mu(h) = 0\}$. Since $\mathcal{F}$ is μ-pre-Gaussian, the variance function $f \mapsto \mathrm{Var}_\mu(f)$ is bounded (cf. Section 1.5 in [201]), so the Cauchy-Schwarz inequality implies that the mapping

$$h \mapsto (f \mapsto \mu(fh))$$

is continuous from $L^2_0(\mu)$ into $\ell^\infty(\mathcal{F})$. Thus, the mapping $h \mapsto \delta'_\mu(h\mu)$ is a continuous linear functional on $L^2_0(\mu)$, and the functional $\delta : \mathcal{P}_0 \to \mathbb{R}$ is differentiable at μ in the sense of [200, p. 363] relative to $\dot{\mathcal{P}}_{0,\mu}$ with derivative $h \mapsto \delta'_\mu(h\mu)$. By Theorem 25.20 and Lemma 25.19 in [200], the semiparametric efficiency bound for estimating δ at μ is given by

$$\sup_{h \in L^2_0(\mu), h \neq 0} \frac{(\delta'_\mu(h\mu))^2}{\|h\|^2_{L^2(\mu)}}. \tag{A.1}$$

We will show that (A.1) coincides with $\mathrm{Var}(\delta'_\mu(G_\mu))$.

Let $C_u(\mathcal{F})$ be the space of uniformly continuous functions on $\mathcal{F}$ w.r.t. the pseudo-metric $d_\mu(f, g) = \sqrt{\mathrm{Var}_\mu(f - g)}$. Then $\mathcal{F}$ is totally bounded w.r.t. d_μ and $G_\mu \in C_u(\mathcal{F})$ with probability one (cf. Example 1.5.10 in [201]). The completion $\bar{\mathcal{F}}$ of $\mathcal{F}$ w.r.t. d_μ is compact and denote the (unique) extension of G_μ to $\bar{\mathcal{F}}$ by the same symbol. Also the restriction of δ'_μ to $C_u(\mathcal{F})$ uniquely extends to $C(\bar{\mathcal{F}})$ (the space of continuous functions on $\bar{\mathcal{F}}$), which we denote by the same symbol δ'_μ. Then, by the Riesz representation theorem, there exists a signed Borel measure $\mathfrak{m}$ on $\bar{\mathcal{F}}$ such that

$$\delta'_\mu(z) = \int_{\bar{\mathcal{F}}} z(f) d\mathfrak{m}(f), \quad z \in C(\bar{\mathcal{F}}).$$

Since $\bar{\mathcal{F}}$ is compact, for any $k = 1, 2, \ldots$, one can find disjoint sets $B_{k,j}$, $j = 1, \ldots, N_k$ in $\bar{\mathcal{F}}$ with radius at most $1/k$ such that $\bigcup_{j=1}^{N_k} B_{k,j} = \bar{\mathcal{F}}$. Approximate $\mathfrak{m}$ by $\sum_{j=1}^{N_k} \mathfrak{m}(B_{k,j})\delta_{f_{k,j}}$ where each $f_{k,j}$ is an arbitrary point in $B_{k,j}$. Then,

$$\left| \delta'_\mu(z) - \sum_{j=1}^{N_k} \mathfrak{m}(B_{k,j}) z(f_{k,j}) \right| \le \sup_{d_\mu(f,g) \le 1/k} |z(f) - z(g)| \times |\mathfrak{m}|(\bar{\mathcal{F}}),$$

where $|\mathfrak{m}|$ is the total variation of $\mathfrak{m}$. For any $h \in L_0^2(\mu)$,

$$|\mu(fg) - \mu(gh)| \le \|h\|_{L^2(\mu)} d_\mu(f, g),$$

so that, uniformly in $h \in L_0^2(\mu)$ such that $\|h\|_{L^2(\mu)} = 1$,

$$\sum_{j=1}^{N_k} \mathfrak{m}(B_{k,j}) \mu(f_{k,j} h) \to \delta'_\mu(h\mu).$$

The left-hand side can be written as $\langle h, \sum_{j=1}^{N_k} \mathfrak{m}(B_{k,j})(f_{k,j} - \mu(f_{k,j})) \rangle_{L^2(\mu)}$, the supremum of which w.r.t. $h \in L_0^2(\mu)$ such that $\|h\|_{L^2(\mu)} = 1$ is

$$\Big\| \sum_{j=1}^{N_k} \mathfrak{m}(B_{k,j}) \big(f_{k,j} - \mu(f_{k,j})\big) \Big\|_{L^2(\mu)}.$$

Thus,

$$\begin{aligned} \sup_{h \in L_0^2(\mu), h \neq 0} \frac{(\delta'_\mu(h\mu))^2}{\|h\|^2_{L^2(\mu)}} &= \sup_{h \in L_0^2(\mu), \|h\|_{L^2(\mu)} = 1} (\delta'_\mu(h\mu))^2 = \Big(\sup_{h \in L_0^2(\mu), \|h\|_{L^2(\mu)} = 1} \delta'_\mu(h\mu) \Big)^2 \\ &= \lim_{k \to \infty} \Big\| \sum_{j=1}^{N_k} \mathfrak{m}(B_{k,j}) \big(f_{k,j} - \mu(f_{k,j})\big) \Big\|^2_{L^2(\mu)}. \end{aligned}$$

On the other hand,

$$\delta'_\mu(G_\mu) = \lim_{k\to\infty} \sum_{j=1}^{N_k} \mathfrak{m}(B_{k,j}) G_\mu(f_{k,j}) = \lim_{k\to\infty} G_\mu\Big(\sum_{j=1}^{N_k} \mathfrak{m}(B_{k,j}) f_{k,j}\Big)$$

a.s., where the second equality follows from the fact that G_μ is a μ-Brownian bridge. This implies that

$$\begin{aligned}\operatorname{Var}(\delta'_\mu(G_\mu)) &= \lim_{k\to\infty} \operatorname{Var}\bigg(G_\mu\Big(\sum_{j=1}^{N_k} \mathfrak{m}(B_{k,j}) f_{k,j}\Big)\bigg)\\ &= \lim_{k\to\infty} \Big\| \sum_{j=1}^{N_k} \mathfrak{m}(B_{k,j})(f_{k,j} - \mu(f_{k,j}))\Big\|^2_{L^2(\mu)}.\end{aligned}$$

Conclude that (A.1) coincides with $\operatorname{Var}(\delta'_\mu(G_\mu))$. □

Remark 25 (Relation with Theorem 3.1 in [198])**.** Consider the assumption of Proposition 1 with $\mathfrak{E} = \mathbb{R}$ and assume that the function class $\mathcal{F}$ is μ-Donsker and the derivative δ'_μ is linear on any subspace of $\mathcal{T}_{\mathcal{P}_0}(\mu)$. If, in addition, $(1+th)\mu \in \mathcal{P}_0$ for sufficiently small $t > 0$ for any $h \in \dot{\mathcal{P}}_{0,\mu}$, then the fact that $\operatorname{Var}(\delta'_\mu(G_\mu))$ (with G_μ being a μ-Brownian bridge) agrees with the semiparametric efficiency bound for estimating δ at μ can be deduced from Theorem 3.1 in [198]. Indeed, recall the map $\tau : \mathcal{P}_0 \ni \nu \mapsto (f \mapsto \nu(f)) \in \ell^\infty(\mathcal{F})$ and $\bar{\delta} : \tau\mathcal{P}_{0,\epsilon} \to \mathbb{R}$ defined by $\bar{\delta}(\tau\nu) = \delta(\nu)$. Then, the map τ is differentiable at μ relative to $\dot{\mathcal{P}}_{0,\mu}$ with derivative $\dot{\tau}_\mu : L^2(\mu) \to \ell^\infty(\mathcal{F})$ given by $\dot{\tau}_\mu(h) = (f \mapsto \mu(fh))$, and the empirical distribution $\hat{\mu}_n = n^{-1}\sum_{i=1}^n \delta_{X_i}$ is asymptotically efficient for estimating τ at μ tangentially to $\dot{\mathcal{P}}_{0,\mu}$; see Section 3.11 in [201] for the result and its precise meaning. Theorem 3.1 in [198] shows that, if $\bar{\delta}$ is Hadamard differentiable at $\tau\mu$ tangentially to $\overline{\dot{\tau}_\mu(\dot{\mathcal{P}}_{0,\mu})}^{\ell^\infty(\mathcal{F})}$, then $\delta(\hat{\mu}_n) = (\bar{\delta}\circ\tau)(\hat{\mu}_n)$ is asymptotically efficient, implying that $\operatorname{Var}(\bar{\delta}'_{\tau\mu}(G_\mu))$ agrees with the semiparametric efficiency bound. Identifying δ with $\bar{\delta}$ and μ with $\tau\mu$ as in the proof of Proposition 1, we have $\operatorname{Var}(\bar{\delta}'_{\tau\mu}(G_\mu)) = \operatorname{Var}(\delta'_\mu(G_\mu))$. In our notation, $\dot{\tau}_\mu(\dot{\mathcal{P}}_{0,\mu}) = \{h\mu : h \in \dot{\mathcal{P}}_{0,\mu}\}$, and from the assumption that $(1+th)\mu \in \mathcal{P}_0$ for sufficiently small $t > 0$ for any $h \in \dot{\mathcal{P}}_{0,\mu}$, we see that $\dot{\tau}_\mu(\dot{\mathcal{P}}_{0,\mu}) \subset \mathcal{T}_{\mathcal{P}_0}(\mu)$, and as $\mathcal{T}_{\mathcal{P}_0}(\mu)$ is closed in $\ell^\infty(\mathcal{F})$, it contains $\overline{\dot{\tau}_\mu(\dot{\mathcal{P}}_{0,\mu})}^{\ell^\infty(\mathcal{F})}$. Finally, as the derivative $\bar{\delta}'_\mu$ is linear on $\overline{\dot{\tau}_\mu(\dot{\mathcal{P}}_{0,\mu})}^{\ell^\infty(\mathcal{F})}$

by assumption, $\bar{\delta}$ is Hadamard differentiable at $\tau\mu$ tangentially to $\overline{\dot{\tau}_\mu(\dot{\mathcal{P}}_{0,\mu})}^{\ell^\infty(\mathcal{F})}$ by Lemma 2.

A.1.5 Proof of Corollary 2

Fix $h_1 \oplus h_2 \in \dot{\mathcal{P}}_{0,\mu} \oplus \dot{\mathcal{P}}_{0,\nu}$. Observe that the submodel $\{\mu_t \otimes \nu_t : 0 \le t < \epsilon'\}$ with $\mu_t = (1+th_1)\mu$ and $\nu_t = (1+th_2)\nu$ for sufficiently small $\epsilon' > 0$ is differentiable in quadratic mean at $t = 0$ with score function $h_1 \oplus h_2$ (note: $(1+th_1(x))(1+th_2(y)) = 1 + t(h_1(x) + h_2(y)) + o(t)$).

Further, $h_1 \oplus h_2 \mapsto \delta'_\mu(h_1\mu) + \delta'_\nu(h_2\nu)$ is a continuous linear functional on $(\dot{\mathcal{P}}_{0,\mu} \oplus \dot{\mathcal{P}}_{0,\nu}, \|\cdot\|_{L^2(\mu\otimes\nu)})$, and thus the semiparametric efficiency bound is given by

$$\sup_{h_1\oplus h_2 \in \dot{\mathcal{P}}_{0,\mu}\oplus\dot{\mathcal{P}}_{0,\nu}, h_1\oplus h_2 \neq 0} \frac{(\delta'_\mu(h_1\mu) + \delta'_\nu(h_2\nu))^2}{\|h_1 \oplus h_2\|^2_{L^2(\mu\otimes\nu)}}.$$

The rest of the proof is analogous to Proposition 2, upon observing that $\mathcal{F}$ is totally bounded w.r.t. $d_{\mu,\nu}(f,g) = \sqrt{\mathrm{Var}_\mu(f-g)} + \sqrt{\mathrm{Var}_\nu(f-g)}$ (if a set T is totally bounded for two pseudometrics d_1 and d_2, then T is totally bounded for $\mathsf{d} = \mathsf{d}_1 + \mathsf{d}_2$), and G_μ and G_ν have $d_{\mu,\nu}$-uniformly continuous paths, combined with the fact that, for $(f,g) \in L^2(\mu) \times L^2(\nu)$ and $(h_1,h_2) \in \dot{\mathcal{P}}_{0,\mu} \times \dot{\mathcal{P}}_{0,\nu}$, we have

$$\mu(h_1 f) + \nu(h_2 g) = (\mu\otimes\nu)((h_1 \oplus h_2)(f \oplus g)),$$

and the supremum of the right-hand side w.r.t. $(h_1,h_2) \in \dot{\mathcal{P}}_{0,\mu} \times \dot{\mathcal{P}}_{0,\nu}$ such that $\|h_1 \oplus h_2\|^2_{L^2(\mu\otimes\nu)} = 1$ is $\sqrt{\mathrm{Var}_{\mu\otimes\nu}(f\oplus g)} = \sqrt{\mathrm{Var}_\mu(f) + \mathrm{Var}_\nu(g)}$. □

A.2 Proofs for Section 2.4

A.2.1 Proofs of Theorems 1 and 2 and Lemma 3

Preliminary results

The section presents preliminary results needed for the derivation of Theorem 1, followed by their proofs. In what follows, we fix $1 < p < \infty$.

Lemma 42 (OT potentials)**.** The following hold.

(i) Suppose $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$ are supported in $B(0, M)$ for some $M > 0$ (i.e, $\operatorname{spt}(\mu), \operatorname{spt}(\nu) \subset B(0, M)$). Then, there exists an OT potential φ from μ to ν for W_p such that $|\varphi(x) - \varphi(x')| \le C_{p,M}\|x - x'\|$ for all $x, x' \in B(0, M)$, where $C_{p,M} < \infty$ is a constant that depends only on p, M.

(ii) If $\mu_0, \mu_1, \nu \in \mathcal{P}(\mathbb{R}^d)$ are supported in $B(0, M)$ for some $M > 0$, then

$$\left|\mathsf{W}_p^p(\mu_1, \nu) - \mathsf{W}_p^p(\mu_0, \nu)\right| \le C_{p,M}\mathsf{W}_1(\mu_0, \mu_1).$$

Proof. **Part (i).** Since $\operatorname{spt}(\mu), \operatorname{spt}(\nu) \subset B(0, M)$, there exists an OT potential φ such that

$$\varphi(x) = \inf_{y \in B(0,M)} [\|x - y\|^p - \varphi^c(y)], \; x \in B(0, M).$$

Indeed, regard μ and ν as probability measures on $B(0, M)$, and pick any OT potential φ_M from μ to ν as probability measures on $B(0, M)$. This φ_M is defined only on $B(0, M)$ so extend it to $\mathbb{R}^d$ by setting $\varphi(x) = \varphi_M(x)$ if $x \in B(0, M)$ and $\varphi(x) = -\infty$ if $x \notin B(0, M)$.

Observe that

$$
\begin{aligned}
\varphi^c(y) &= \inf_{x\in\mathbb{R}^d}[\|x-y\|^p - \varphi(x)] \\
&= \inf_{x\in B(0,M)}[\|x-y\|^p - \varphi(x)] \\
&= \inf_{x\in B(0,M)}[\|x-y\|^p - \varphi_M(x)] = \varphi_M^c(y).
\end{aligned}
$$

Thus, whenever $x \in B(0, M)$,

$$
\begin{aligned}
\varphi(x) = \varphi_M(x) &= \varphi_M^{cc}(x) \\
&= \inf_{y\in B(0,M)}[\|x-y\|^p - \varphi_M^c(y)] \\
&= \inf_{y\in B(0,M)}[\|x-y\|^p - \varphi^c(y)].
\end{aligned}
$$

Now, for $x, x' \in B(0, M)$,

$$
|\varphi(x) - \varphi(x')| \leq \sup_{y\in B(0,M)} \Big|\|x-y\|^p - \|x'-y\|^p\Big| \leq p2^{p-1}M^{p-1}\|x-x'\|.
$$

Part (ii). Let φ_i denote an OT potential from μ_i to ν. Observe that

$$
\begin{aligned}
\mathsf{W}_p^p(\mu_1,\nu) &\geq \int \varphi_0 d\mu_1 + \int \varphi_0^c d\nu \\
&= \int \varphi_0 d\mu_0 + \int \varphi_0^c d\nu + \int \varphi_0 d(\mu_1 - \mu_0) \\
&= \mathsf{W}_p^p(\mu_0,\nu) + \int \varphi_0 d(\mu_1 - \mu_0).
\end{aligned}
$$

Likewise, we have

$$
\begin{aligned}
\mathsf{W}_p^p(\mu_1,\nu) &= \int \varphi_1 d\mu_1 + \int \varphi_1^c d\nu \\
&= \int \varphi_1 d\mu_0 + \int \varphi_1^c d\nu + \int \varphi_1 d(\mu_1 - \mu_0) \\
&\leq \mathsf{W}_p^p(\mu_0,\nu) + \int \varphi_1 d(\mu_1 - \mu_0).
\end{aligned}
$$

The conclusion follows from the fact that $\varphi_i, i = 1, 2$ are $C_{p,M}$-Lipschitz on $B(0, M)$ and any Lipschitz function on $B(0, M)$ can be extended to $\mathbb{R}^d$ without changing the Lipschitz constant (cf. [72], Theorem 6.1.1), followed by the Kantorovich-Rubinstein duality. □

Part (ii) of the preceding lemma implies the following corollary.

Corollary 9. If $\mu_0, \mu_1, \nu \in \mathcal{P}(\mathbb{R}^d)$ are supported in $B(0, M)$ for some $M > 0$, then

$$\left|\underline{\mathsf{W}}_p^p(\mu_1, \nu) - \underline{\mathsf{W}}_p^p(\mu_0, \nu)\right| \bigvee \left|\overline{\mathsf{W}}_p^p(\mu_1, \nu) - \overline{\mathsf{W}}_p^p(\mu_0, \nu)\right| \leq C_{p,M}\overline{\mathsf{W}}_1(\mu_0, \mu_1).$$

Proof. This follows from Part (ii) of the preceding lemma, the definitions of $\underline{\mathsf{W}}_p$ and $\overline{\mathsf{W}}_p$, and the fact that $\mathfrak{p}^\theta_\sharp\mu_0$, $\mathfrak{p}^\theta_\sharp\mu_1$, $\mathfrak{p}^\theta_\sharp\nu$ are supported in $[-M, M]$. □

Lemma 43 (Regularity of sliced OT potential)**.** Let $\mu, \nu \in \mathcal{P}_p(\mathbb{R}^d)$ and assume that μ is absolutely continuous with convex support. Set $\mathcal{X} = \mathrm{int}(\mathrm{spt}(\mu))$. For each $\theta \in \mathbb{S}^{d-1}$, let φ^θ be an OT potential from $\mathfrak{p}^\theta_\sharp\mu$ to $\mathfrak{p}^\theta_\sharp\nu$ for W_p. Pick and fix any $x_0 \in \mathcal{X}$. Then the following hold.

(i) The map $\mathcal{X} \times \mathbb{S}^{d-1} \ni (x, \theta) \mapsto \varphi^\theta(\theta^\intercal x) - \varphi^\theta(\theta^\intercal x_0)$ is measurable in x and continuous in θ, and thus jointly measurable in (x, θ).

(ii) The map $\mathcal{X}\times\mathbb{S}^{d-1} \ni (x, \theta) \mapsto \varphi^\theta(\theta^\intercal x)-\varphi^\theta(\theta^\intercal x_0)$ is unique, in the sense that if, for each $\theta \in \mathbb{S}^{d-1}$, $\tilde{\varphi}^\theta$ is another OT potential from $\mathfrak{p}^\theta_\sharp\mu$ to $\mathfrak{p}^\theta_\sharp\nu$ for W_p, then $\varphi^\theta(\theta^\intercal x)-\varphi^\theta(\theta^\intercal x_0) = \tilde{\varphi}^\theta(\theta^\intercal x) - \tilde{\varphi}^\theta(\theta^\intercal x_0)$ for all $(x, \theta) \in \mathcal{X} \times \mathbb{S}^{d-1}$.

(iii) If, in addition, $\mathrm{spt}(\mu)$ and $\mathrm{spt}(\nu)$ are contained in $B(0, M)$ for some $M > 0$, then $|\varphi^\theta(\theta^\intercal x) - \varphi^\theta(\theta^\intercal x_0)| \leq C_{p,M}$ for all $(x, \theta) \in \mathcal{X} \times \mathbb{S}^{d-1}$, where $C_{p,M}$ is given in Lemma 42. In particular, for any finite sighed Borel measure ρ on $\mathbb{R}^d$ with total mass 0, the map $\mathbb{S}^{d-1} \ni \theta \mapsto \int \varphi^\theta(\theta^\intercal x)d\rho(x)$ is continuous.

Proof. **Part (i).** Measurability of the map $x \mapsto \varphi^\theta(\theta^\intercal x)$ follows by measurability of the map $t \mapsto \varphi^\theta(t)$ and continuity of $x \mapsto \theta^\intercal x$. We shall prove continuity of the map $\theta \mapsto \varphi^\theta(\theta^\intercal x) - \varphi^\theta(\theta^\intercal x_0)$. By convexity of the support and absolute continuity of μ, we have $\mu(\mathcal{X}) = 1$, as the boundary of any convex set has Lebesgue measure zero. Let $\mathcal{X}_\theta = \{\theta^\intercal x : x \in \mathcal{X}\}$. Then $\mathcal{X}_\theta$ is an open interval and its closure coincides with $\mathrm{spt}(\mathfrak{p}^\theta_\sharp\mu)$;

see Lemma 44 below. Pick any $x \in \mathcal{X}$ and let $\theta_n \to \theta$ in $\mathbb{S}^{d-1}$. Then, for all large n, $\theta_n^\intercal x$ lies in a compact neighborhood of $\theta^\intercal x$ contained in $\mathcal{X}_\theta$. Since $\mathfrak{p}^{\theta_n}_\sharp \mu \stackrel{w}{\to} \mathfrak{p}^{\theta}_\sharp \mu$, by Theorem 3.4 in [61], $\varphi^{\theta_n} - \varphi^{\theta_n}(\theta^\intercal x_0) \to \varphi^{\theta} - \varphi^{\theta}(\theta^\intercal x_0)$ uniformly on each compact subset of $\mathcal{X}_\theta$. This implies that $\varphi^{\theta_n}(\theta_n^\intercal x) - \varphi^{\theta_n}(\theta_n^\intercal x_0) \to \varphi^{\theta}(\theta^\intercal x) - \varphi^{\theta}(\theta^\intercal x_0)$. The last claim follows from, e.g., Lemma 4.51 in [3].

Part (ii). This follows from Corollary 2.7 in [61].

Part (iii). The first claim follows from Lemma 42 Part (i) (note: $\mathrm{spt}(\mathfrak{p}^{\theta}_\sharp \mu)$ and $\mathrm{spt}(\mathfrak{p}^{\theta}_\sharp \nu)$ are contained in $[-M, M]$) together with the uniqueness result from Part (ii) of this lemma. The second claim follows from the fact that $\int \varphi^{\theta}(\theta^\intercal x) d\rho(x) = \int [\varphi^{\theta}(\theta^\intercal x) - \varphi^{\theta}(\theta^\intercal x_0)] d\rho(x)$ and the dominated convergence theorem. ◻

We shall prove the following technical result used in the preceding proof.

Lemma 44 (Support of projected distribution)**.** Let $\mu \in \mathcal{P}(\mathbb{R}^d)$ be absolutely continuous with convex support and let $\mathbb{A} \in \mathbb{R}^{m \times d}$ be of rank m (thus $m \le d$). Set $\mathcal{X} = \mathrm{int}(\mathrm{spt}(\mu))$. Define $\nu \in \mathcal{P}(\mathbb{R}^m)$ by $\nu(B) = \mu(\{x : \mathbb{A}x \in B\})$. Then $\mathbb{A}\mathcal{X} = \{\mathbb{A}x : x \in \mathcal{X}\} \subset \mathbb{R}^m$ is open and convex, and its closure coincides with $\mathrm{spt}(\nu)$.

Proof. Since the mapping $x \mapsto \mathbb{A}x$ is open, the set $\mathbb{A}\mathcal{X}$ is open. Convexity of $\mathbb{A}\mathcal{X}$ is trivial. Pick any $x \in \mathcal{X}$ and any open neighborhood N_y of $y = \mathbb{A}x$. Then, since $\mathbb{A}^{-1}N_y$ is open and contains x, $\nu(N_y) = \mu(\mathbb{A}^{-1}N_y) > 0$, which implies that $\mathbb{A}\mathcal{X} \subset \mathrm{spt}(\nu)$. Since clearly $\nu(\mathbb{A}\mathcal{X}) = 1$, the closure of $\mathbb{A}\mathcal{X}$ coincides with $\mathrm{spt}(\nu)$. ◻

Lemma 45 (Gâteaux derivatives of $\underline{\mathsf{W}}_p$ and $\overline{\mathsf{W}}_p$)**.** Suppose that $\mu_0, \mu_1, \nu \in \mathcal{P}(\mathbb{R}^d)$ are compactly supported, that μ_0 is absolutely continuous, and that the support of μ_0 is convex and and its interior has μ_1-probability one, i.e., $\mu_1(\mathrm{int}(\mathrm{spt}(\mu_0))) = 1$. For every

$\theta \in \mathbb{S}^{d-1}$, let φ_0^θ be an OT potential from $\mathfrak{p}^\theta_\sharp \mu_0$ to $\mathfrak{p}^\theta_\sharp \nu$ for W_p. Let $\rho = \mu_1 - \mu_0$. Then we have

$$\frac{d}{dt^+} \underline{\mathsf{W}}_p^p(\mu_0 + t\rho, \nu)\Big|_{t=0} = \int_{\mathbb{S}^{d-1}} \int \varphi_0^\theta(\theta^\intercal x) d\rho(x) d\sigma(\theta), \tag{A.2}$$

$$\frac{d}{dt^+} \overline{\mathsf{W}}_p^p(\mu_0 + t\rho, \nu)\Big|_{t=0} = \sup_{\theta \in \mathfrak{S}_{\mu,\nu}} \int \varphi_0^\theta(\theta^\intercal x) d\rho(x). \tag{A.3}$$

The right-hand sides of (A.2) and (A.3) are well-defined and finite.

Proof. We first prove (A.2). Let φ_t^θ denote an OT potential from $\mathfrak{p}^\theta_\sharp \mu_t$ to $\mathfrak{p}^\theta_\sharp \nu$ with $\mu_t = \mu_0 + t\rho$. We may assume without loss of generality that $\varphi_t^\theta(\theta^\intercal x_0) = 0$, where x_0 is an arbitrary (fixed) interior point of $\mathrm{spt}(\mu_0)$. Then,

$$\begin{aligned}
\underline{\mathsf{W}}_p^p(\mu_t, \nu) &\geq \int_{\mathbb{S}^{d-1}} \int \varphi_0^\theta(\theta^\intercal x)\, d\mu_t(x)\, d\sigma(\theta) + \int_{\mathbb{S}^{d-1}} \int [\varphi_0^\theta]^c(\theta^\intercal y)\, d\nu(y)\, d\sigma(\theta) \\
&= \underline{\mathsf{W}}_p^p(\mu_0, \nu) + t \int_{\theta \in \mathbb{S}^{d-1}} \int \varphi_0^\theta(\theta^\intercal x)\, d\rho(x)\, d\sigma(\theta), \quad \text{and} \\
\underline{\mathsf{W}}_p^p(\mu_t, \nu) &\leq \underline{\mathsf{W}}_p^p(\mu_0, \nu) + t \int_{\theta \in \mathbb{S}^{d-1}} \int \varphi_t^\theta(\theta^\intercal x)\, d\rho(x)\, d\sigma(\theta).
\end{aligned}$$

The first inequality implies that

$$\liminf_{t \downarrow 0} \frac{\underline{\mathsf{W}}_p^p(\mu_t, \nu) - \underline{\mathsf{W}}_p^p(\mu_0, \nu)}{t} \geq \int_{\mathbb{S}^{d-1}} \int \varphi_0^\theta(\theta^\intercal x) d\rho(x) d\sigma(\theta).$$

To prove the upper bound, we note that $\varphi_t^\theta(\theta^\intercal \cdot)$ is uniformly bounded on $\mathrm{spt}(\mu_0)$ by Lemma 43 Part (iii), and by Theorem 3.4 in [61], $\varphi_t^\theta(\theta^\intercal x) \to \varphi_0^\theta(\theta^\intercal x)$ as $t \downarrow 0$ for each $x \in \mathrm{int}(\mathrm{spt}(\mu_0))$ and $\theta \in \mathbb{S}^{d-1}$. Thus, by the dominated convergence theorem, we have

$$\limsup_{t \downarrow 0} \frac{\underline{\mathsf{W}}_p^p(\mu_t, \nu) - \underline{\mathsf{W}}_p^p(\mu_0, \nu)}{t} \leq \int_{\mathbb{S}^{d-1}} \int \varphi_0^\theta(\theta^\intercal x) d\rho(x) d\sigma(\theta).$$

Next we prove (A.3). To save space, we write $\int \varphi_t^\theta(\theta^\intercal x) d\mu_t(x) = \int \varphi_t^\theta d\mu_t$ etc. Observe that for $t \in (0,1)$,

$$\begin{aligned}
&\frac{1}{t}\left[\overline{\mathsf{W}}_p^p(\mu_t,\nu) - \overline{\mathsf{W}}_p^p(\mu_0,\nu)\right] \\
&= \frac{1}{t}\left[\sup_{\theta\in\mathbb{S}^{d-1}}\left(\int \varphi_t^\theta\, d\mu_t + \int [\varphi_t^\theta]^c d\nu\right) - \sup_{\theta\in\mathbb{S}^{d-1}}\left(\int \varphi_0^\theta\, d\mu_0 + \int [\varphi_0^\theta]^c\, d\nu\right)\right] \\
&\le \sup_{\theta\in\mathbb{S}^{d-1}}\left\{\int \varphi_t^\theta\, d\rho + \frac{1}{t}\left(\int \varphi_0^\theta\, d\mu_0 + \int [\varphi_0^\theta]^c\, d\nu\right)\right\} \\
&\qquad - \frac{1}{t}\sup_{\theta\in\mathbb{S}^{d-1}}\left(\int \varphi_0^\theta\, d\mu_0 + \int [\varphi_0^\theta]^c\, d\nu\right).
\end{aligned}$$

Let $S_\epsilon = \{\theta \in \mathbb{S}^{d-1} : \mathsf{W}_p^p(\mathfrak{p}^\theta_\sharp\mu_0, \mathfrak{p}^\theta_\sharp\nu) \ge \overline{\mathsf{W}}_p^p(\mu_0,\nu) - \epsilon\}$. Then, for $\theta \notin S_\epsilon$,

$$\begin{aligned}
\int \varphi_t^\theta\, d\rho + &\frac{1}{t}\left(\int \varphi_0^\theta\, d\mu_0 + \int [\varphi_0^\theta]^c\, d\nu\right) \\
&- \frac{1}{t}\sup_{\vartheta\in\mathbb{S}^{d-1}}\left(\int \varphi_0^\vartheta\, d\mu_0 + \int [\varphi_0^\vartheta]^c\, d\nu\right) \le \int \varphi_t^\theta\, d\rho - \frac{\epsilon}{t},
\end{aligned}$$

which tends to $-\infty$ uniformly over θ as $t \downarrow 0$. Hence, for every $\epsilon > 0$, we have

$$\limsup_{t\downarrow 0}\frac{1}{t}\left[\overline{\mathsf{W}}_p^p(\mu_t,\nu) - \overline{\mathsf{W}}_p^p(\mu_0,\nu)\right] \le \limsup_{t\downarrow 0}\sup_{\theta\in S_\epsilon}\int \varphi_t^\theta\, d\rho.$$

Taking $\epsilon \downarrow 0$ above, since the right-hand side is decreasing in ϵ, we have

$$\limsup_{t\downarrow 0}\frac{1}{t}\left[\overline{\mathsf{W}}_p^p(\mu_t,\nu) - \overline{\mathsf{W}}_p^p(\mu_0,\nu)\right] \le \lim_{\epsilon\downarrow 0}\limsup_{t\downarrow 0}\sup_{\theta\in S_\epsilon}\int \varphi_t^\theta\, d\rho.$$

We will show that $\int \varphi_t^\theta\, d\rho \to \int \varphi_0^\theta\, d\rho$ uniformly over $\theta \in \mathbb{S}^{d-1}$ as $t \downarrow 0$. To this end, by compactness of $\mathbb{S}^{d-1}$ and continuity of $\int \varphi_0^\theta\, d\rho$ in θ, it suffices to show that for any sequences $t_n \downarrow 0$ and $\theta_n \to \theta$, $\int \varphi_{t_n}^{\theta_n}\, d\rho \to \int \varphi_0^\theta\, d\rho$.[1] Since, for any bounded continuous

[1]This can be verified as follows. Suppose on the contrary that $c := \liminf_{t\downarrow 0}\sup_\theta |\int(\varphi_t^\theta - \varphi_0^\theta)d\rho| > 0$. Then there exist sequences $t_n \downarrow 0$ and $\theta_n \in \mathbb{S}^{d-1}$ such that $|\int(\varphi_{t_n}^{\theta_n} - \varphi_0^{\theta_n})d\rho| \to c$. Since $\mathbb{S}^{d-1}$ is compact, there exists a subsequence of θ_n converging to some $\theta \in \mathbb{S}^{d-1}$. Thus, along with the subsequence, we have $|\int \varphi_{t_n}^{\theta_n} d\rho - \int \varphi_0^\theta d\rho| \to c$, a contradiction.

function g on $\mathbb{R}$,

$$\begin{aligned}\int g d(\mathfrak{p}^{\theta_n}_{\sharp}\mu_{t_n}) &= \int g(\theta_n^\intercal x) d\mu_{t_n}(x)\\ &= \int g(\theta_n^\intercal x) d\mu_0(x) + t_n \int g(\theta_n^\intercal x) d\rho(x)\\ &= \int g(\theta^\intercal x) d\mu_0(x) + o(1) \quad \text{(as } g \text{ is bounded and continuous)}\\ &= \int g d(\mathfrak{p}^{\theta}_{\sharp}\mu_0) + o(1),\end{aligned}$$

we have $\mathfrak{p}^{\theta_n}_{\sharp}\mu_{t_n} \xrightarrow{w} \mathfrak{p}^{\theta}_{\sharp}\mu_0$. Thus, by Theorem 3.4 in [61], $\varphi^{\theta_n}_{t_n} \to \varphi^{\theta}_0$ uniformly on each compact subset of $\mathrm{int}(\mathrm{spt}(\mathfrak{p}^{\theta}_{\sharp}\mu_0))$, which implies that $\varphi^{\theta_n}_{t_n}(\theta_n^\intercal x) \to \varphi^{\theta}_0(\theta^\intercal x)$ for each $x \in \mathrm{int}(\mathrm{spt}(\mu_0))$. Since φ^{ϑ}_t is bounded on $\mathrm{int}(\mathrm{spt}(\mu_0))$ uniformly in (t, ϑ), we conclude from the dominated convergence theorem that $\int \varphi^{\theta_n}_{t_n} d\rho \to \int \varphi^{\theta}_0 d\rho$. This yields that

$$\limsup_{t\downarrow 0} \frac{1}{t}\left[\overline{\mathsf{W}}^p_p(\mu_t, \nu) - \overline{\mathsf{W}}^p_p(\mu_0, \nu)\right] \le \lim_{\epsilon\downarrow 0} \sup_{\theta\in S_\epsilon} \int \varphi^{\theta}_0 d\rho.$$

To prove the reverse inequality, we note that

$$\begin{aligned}\frac{1}{t}\left[\overline{\mathsf{W}}^p_p(\mu_t, \nu) - \overline{\mathsf{W}}^p_p(\mu_0, \nu)\right] \ge \frac{1}{t}\Bigg[&\sup_{\theta\in\mathbb{S}^{d-1}}\left(\int \varphi^{\theta}_0 d\mu_t + \int [\varphi^{\theta}_0]^c d\nu\right)\\ &- \sup_{\theta\in\mathbb{S}^{d-1}}\left(\int \varphi^{\theta}_0 d\mu_0 + \int [\varphi^{\theta}_0]^c d\nu\right)\Bigg].\end{aligned}$$

For $\theta \in S_\epsilon$,

$$\left(\int \varphi^{\theta}_0 d\mu_t + \int [\varphi^{\theta}_0]^c d\nu\right) - \sup_{\theta\in\mathbb{S}^{d-1}}\left(\int \varphi^{\theta}_0 d\mu_0 + \int [\varphi^{\theta}_0]^c d\nu\right) \ge t\int \varphi^{\theta}_0 d\rho - \epsilon.$$

Choosing $\epsilon = \epsilon_t = t^2$, we have

$$\limsup_{t\downarrow 0} \frac{1}{t}\left[\overline{\mathsf{W}}_p(\mu_t, \nu) - \overline{\mathsf{W}}_p(\mu_0, \nu)\right] \ge \lim_{\epsilon\downarrow 0} \sup_{\theta\in S_\epsilon} \int \varphi^{\theta}_0 d\rho.$$

Since $\mathbb{S}^{d-1}$ is compact and $\int \varphi^{\theta}_0 d\rho$ is continuous in θ, we further have

$$\lim_{\epsilon\downarrow 0} \sup_{\theta\in S_\epsilon} \int \varphi^{\theta}_0 d\rho = \sup_{\theta\in\mathfrak{S}_{\mu,\nu}} \int \varphi^{\theta}_0 d\rho.$$

This completes the proof. $\square$

Corollary 9 shows that the mapping $\mu \mapsto \underline{\mathsf{W}}_p(\mu, \nu)$ and $\mu \mapsto \overline{\mathsf{W}}_p(\mu, \nu)$ are Lipschitz continuous w.r.t. $\overline{\mathsf{W}}_1$. By the Kantorovich-Rubinstein duality, $\overline{\mathsf{W}}_1$ can be represented as $\overline{\mathsf{W}}_1(\mu, \nu) = \|\mu - \nu\|_{\infty,\mathcal{F}}$ with $\mathcal{F} = \{\varphi \circ \mathfrak{p}^\theta : \theta \in \mathbb{S}^{d-1}, \varphi \in \mathsf{Lip}_{1,0}(\mathbb{R})\}$. It is not difficult to show that $\mathcal{F}$ is μ-Donsker if μ is compactly supported. For a later purpose, we allow μ to be unbounded in the following lemma.

Lemma 46 (Bracketing entropy bound for projected Lipschitz functions)**.** Let $\mathcal{F} = \{\varphi \circ \mathfrak{p}^\theta : \theta \in \mathbb{S}^{d-1}, \varphi \in \mathsf{Lip}_{1,0}(\mathbb{R})\}$. For any $\mu \in \mathcal{P}_{2+\epsilon}(\mathbb{R}^d)$ for some $\epsilon > 0$, we have

$$\log N_{[\,]}(t, \mathcal{F}, L^2(\mu)) \lesssim C^{1/\epsilon} t^{-(1+2/\epsilon)} + d \log(2C/t), \quad t \in (0, 1) \tag{A.4}$$

up to a constant that depends only on ϵ, where $C = \int \|x\|^{2+\epsilon} d\mu(x)$. In particular, $\mathcal{F}$ is μ-Donsker provided that $\mu \in \mathcal{P}_{4+\epsilon}(\mathbb{R}^d)$ for some $\epsilon > 0$.

In what follows, $N(\epsilon, \mathcal{F}, d)$ denotes the ϵ-covering number of a function class $\mathcal{F}$ w.r.t. a pseudometric d, and $N_{[\,]}(\epsilon, \mathcal{F}, d)$ is the corresponding bracketing number. The proof of this lemma relies on the following two auxiliary lemmas.

Lemma 47 (Uniform entropy for 1-Lipschitz functions)**.** For any $K > 0$,

$$\log N(t, \mathsf{Lip}_{1,0}([-K, K]), \|\cdot\|_\infty) \le \lceil K/t \rceil 2 \log 2, \ t \in (0, 1), \tag{A.5}$$

where $\mathsf{Lip}_{1,0}([-K, K])$ denotes the class of 1-Lipschitz functions f on $[-K, K]$ with $f(0) = 0$.

Proof of Lemma 47. Let $N = \lceil K/t \rceil$. Partition $[0, K]$ into N subintervals with length at most t. Enumerate the endpoints of these subintervals and label them as $0 = x_0 < \cdots < x_N = K$. For each vector $j = (j_{-N}, \ldots, j_{-1}, j_1, \ldots, j_N) \in \{-1, 1\}^{2N}$ indexed from $-N$ to N

excluding 0, define the function

$$f_j(x) = \begin{cases} 0 & \text{if } x = 0 \\ \sum_{l=1}^{k} j_l(x_k - x_{k-1}) & \text{if } x = x_k, 1 \le k \le N \\ \sum_{l=1}^{k} j_{-l}(x_k - x_{k-1}) & \text{if } x = -x_k, 1 \le k \le N \\ \text{piecewise linear} & \text{otherwise.} \end{cases} \tag{A.6}$$

These form a t-net of $\mathsf{Lip}_{1,0}([-K, K])$ w.r.t. $\|\cdot\|_\infty$. □

The following result is standard and thus we omit the proof.

Lemma 48 (Covering number for $\mathbb{S}^{d-1}$)**.** For any $t \in (0, 1)$, $N(t, \mathbb{S}^{d-1}, \|\cdot\|) \le (5/t)^d$.

Proof of Lemma 46. Pick any $t \in (0, 1)$ and choose $M > 0$ such that

$$\sqrt{\mathbb{E}[\|X\|^2 \mathbb{1}_{\{\|X\|>M\}}]} \le t/4$$

for $X \sim \mu$. Consider a minimal $t/8$-net $f_1, \dots, f_m$ of $\mathsf{Lip}_{1,0}([-M, M])$ w.r.t. $\|\cdot\|_\infty$, and a minimal $t/(8M)$-net $\theta_1, \dots, \theta_n$ of $\mathbb{S}^{d-1}$ w.r.t. $\|\cdot\|$. Consider the following function brackets for $i = 1, \dots, m$ and $j = 1, \dots, n$:

$$\Big[(f_i \circ \mathfrak{p}^{\theta_j} - t/4)\mathbb{1}_{\{\|x\|\le M\}} - \|x\|\mathbb{1}_{\{\|x\|>M\}},\ (f_i \circ \mathfrak{p}^{\theta_j} + t/4)\mathbb{1}_{\{\|x\|\le M\}} + \|x\|\mathbb{1}_{\{\|x\|>M\}}\Big].$$

It is not difficult to verify that these brackets cover $\mathcal{F}$. The $L^2(\mu)$ width of the brackets is bounded above by t. By Lemmas 47 and 48,

$$\log mn = \log m + \log n \le (8M/t + 1)2\log 2 + \log d + (d-1)\log 3 + (d-1)\log\left(\frac{8M}{t}\right).$$

We are left to determine M. Observe that

$$\mathbb{E}[\|X\|^2 \mathbb{1}_{\{\|X\|>M\}}] \le \frac{\mathbb{E}[\|X\|^{2+\epsilon}]}{M^\epsilon} = \frac{C}{M^\epsilon}.$$

Hence $M = (4C^{1/2}/t)^{2/\epsilon}$ is a feasible choice. This leads to the first claim.

For the second claim, by Theorem 2.5.6 in [201], it is enough to show that

$$\int_0^\infty \sqrt{N_{[\,]}(t, \mathcal{F}, L^2(\mu))}\, dt < \infty.$$

Provided that μ has finite $(4+\epsilon)$th moment for some $\epsilon > 0$, $\sqrt{N_{[\,]}(t, \mathcal{F}, L^2(\mu))}$ is integrable near 0. Since $\mathcal{F}$ has envelope function $F(x) = \|x\|$ which is also in $L^2(\mu)$ under the given moment condition, $\mathcal{F}$ is μ-Donsker. $\square$

Proof of Theorem 1

Part (i). We shall apply Proposition 1 with $\mathcal{P}_0 = \{\rho \in \mathcal{P}(\mathbb{R}^d) : \rho(\mathcal{X}) = 1\}$, $\mathcal{X} = \mathrm{int}(\mathrm{spt}(\mu))$, $\mathcal{F} = \{\varphi \circ \mathfrak{p}^\theta : \theta \in \mathbb{S}^{d-1}, \varphi \in \mathsf{Lip}_{1,0}(\mathbb{R})\}$, $F(x) = \|x\|$, and $\delta(\rho) = \underline{\mathsf{W}}_p^p(\rho, \nu)$. It is not difficult to verify that $\mathcal{P}_0$ is convex and contains $\hat{\mu}_n$ with probability one. Further, by Lemma 46, $\mathcal{F}$ is μ-Donsker, $\sqrt{n}(\hat{\mu}_n - \mu) \xrightarrow{d} G_\mu$ in $\ell^\infty(\mathcal{F})$. Combined with Corollary 9 (recall that $\overline{\mathsf{W}}_1(\mu_0, \mu_1) = \|\mu_1 - \mu_0\|_{\infty,\mathcal{F}}$ by the Kantorovich-Rubinstein duality) and Lemma 45, we have verified all conditions in Proposition 1 with $\delta'_\mu(\rho - \mu) = \int_{\mathbb{S}^{d-1}} (\rho - \mu)(\varphi^\theta \circ \mathfrak{p}^\theta) d\sigma(\theta)$.

Let $M > 0$ such that $\mathcal{X} \subset B(0, M)$. Each $\varphi^\theta|_{[-M,M]}$ is c-Lipshitz for some constant $\mathsf{c} < \infty$ independent of θ by Lemma 42. Extend $\varphi^\theta|_{[-M,M]}$ to $\mathbb{R}$ without changing the Lipschitz constant and redefine φ^θ by this extension. In view of Lemma 43 Part (iii) and noting that $(\rho - \mu)(\varphi^\theta \circ \mathfrak{p}^\theta - a) = (\rho - \mu)(\varphi^\theta \circ \mathfrak{p}^\theta)$, δ'_μ extends to $\mathcal{T}_{\mathcal{P}_0}(\mu)$ as

$$\delta'_\mu(m) = \int_{\mathbb{S}^{d-1}} \mathsf{c} \cdot m(\bar{\varphi}^\theta \circ \mathfrak{p}^\theta) d\sigma(\theta) \tag{A.7}$$

with $\bar{\varphi}^\theta = (\varphi^\theta - \varphi^\theta(0))/\mathsf{c}$. Conclude that

$$\sqrt{n}(\underline{\mathsf{W}}_p^p(\hat{\mu}_n, \nu) - \underline{\mathsf{W}}_p^p(\mu, \nu)) \xrightarrow{d} \delta'_\mu(G_\mu) = \int_{\mathbb{S}^{d-1}} \mathsf{c} \cdot G_\mu(\bar{\varphi}^\theta \circ \mathfrak{p}^\theta) d\sigma(\theta).$$

Set $\mathbb{G}_\mu(\theta) = \mathsf{c} \cdot G_\mu(\bar{\varphi}^\theta \circ \mathfrak{p}^\theta)$. The process $(\mathbb{G}_\mu(\theta))_{\theta \in \mathbb{S}^{d-1}}$ is centered Gaussian with covariance function $\mathrm{Cov}(\mathbb{G}_\mu(\theta), \mathbb{G}_\mu(\vartheta)) = \mathrm{Cov}_\mu(\varphi^\theta \circ \mathfrak{p}^\theta, \varphi^\vartheta \circ \mathfrak{p}^\vartheta)$. Further, since $\theta \mapsto \bar{\varphi}^\theta \circ \mathfrak{p}^\theta$ is continuous relative to the pseudometric $d_\mu(f, g) = \sqrt{\mathrm{Var}_\mu(f - g)}$ (cf. Lemma 43(i)) and

G_μ has uniformly continuous paths relative to d_μ, we see that $\mathbb{G}_\mu$ has continuous paths. By Riemann approximation, we see that $\int_{\mathbb{S}^{d-1}} \mathbb{G}_\mu d\sigma$ is centered Gaussian with variance $\int_{\mathbb{S}^{d-1}} \int_{\mathbb{S}^{d-1}} \mathrm{Cov}(\mathbb{G}_\mu(\theta), \mathbb{G}_\mu(\vartheta)) d\sigma(\theta) d\sigma(\vartheta) = v_p^2$.

Second, we show that v_p^2 coincides with the semiparametric efficiency bound by invoking Proposition 2. We have shown that the function class $\mathcal{F}$ is μ-Donsker, which implies μ-pre-Gaussianity. Also, for any bounded μ-mean zero function h, $\mathrm{spt}((1+th)\mu) = \mathrm{spt}(\mu)$ for sufficiently small $t > 0$, so that $(1 + th)\mu \in \mathcal{P}_0$. It remains to show that the derivative δ'_μ in (A.2) suitably extends to a continuous linear functional on $\ell^\infty(\mathcal{F})$. Let $C_u(\mathcal{F})$ denote the space of uniformly continuous functions on $\mathcal{F}$ relative to d_μ. As $\mathcal{F}$ is μ-pre-Gaussian, $\mathcal{F}$ is totally bounded for d_μ. Since $\theta \mapsto \varphi^\theta \circ \mathfrak{p}^\theta$ is continuous relative to d_μ, the expression (A.7) makes sense for any $m \in C_u(\mathcal{F})$, and the map δ'_μ is continuous and linear on $(C_u(\mathcal{F}), \|\cdot\|_{\infty,\mathcal{F}})$. Extend δ'_μ from $C_u(\mathcal{F})$ to a continuous linear functional on $\ell^\infty(\mathcal{F})$ by the Hahn-Banach theorem. For this extension, since $h\mu \in C_u(\mathcal{F})$ for any bounded μ-mean zero function h, we have $t^{-1}(\delta_\mu((1 + th)\mu) - \delta(\mu)) \to \delta'_\mu(h\mu)$ as $t \downarrow 0$. Applying Proposition 2, we see that the semiparametric efficiency bound is given by $\mathrm{Var}(\delta'_\mu(G_\mu))$, which coincides with v_p^2 from the computation above.

Finally, we show the bootstrap consistency. Since the derivative δ'_μ in (A.2) is linear, $\rho \mapsto \delta(\rho)$ is Hadamard differentiable (w.r.t. $\|\cdot\|_{\infty,\mathcal{F}}$) at $\rho = \mu$ tangentially to $\mathrm{spt}(G_\mu)$ by Corollary 1. Thus the bootstrap consistency follows from Theorem 23.9 in [200].

Part (ii). To account for the two-sample case, we consider the following setting. Set $\mathcal{P}_0 = \{\rho_1 \otimes \rho_2 \in \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d) : (\rho_1 \otimes \rho_2)(\mathcal{X} \times \mathcal{Y}) = 1\}$ with $\mathcal{X} = \mathrm{int}(\mathrm{spt}(\mu))$ and $\mathcal{Y} = \mathrm{int}(\mathrm{spt}(\nu))$, $\mathcal{F} = \{f_1 \oplus f_2 : f_i \in \mathcal{F}_0, i = 1, 2\}$ with $\mathcal{F}_0 = \{\varphi \circ \mathfrak{p}^\theta : \theta \in \mathbb{S}^{d-1}, \varphi \in \mathsf{Lip}_{1,0}(\mathbb{R})\}$, and $\delta(\rho_1 \otimes \rho_2) = \underline{\mathsf{W}}_p^p(\rho_1, \rho_2)$. Observe that, for $f_1, f_2 \in \mathcal{F}_0$, $\rho_1 \otimes \rho_2, \rho'_1 \otimes \rho'_2 \in \mathcal{P}_0$, and $t \in [0, 1]$, we have

$$(t(\rho_1 \otimes \rho_2) + (1-t)(\rho'_1 \otimes \rho'_2))(f_1 \oplus f_2) = \Big((t\rho_1 + (1-t)\rho'_1) \otimes (t\rho_2 + (1-t)\rho'_2)\Big)(f_1 \oplus f_2).$$

Thus

$$t(\rho_1 \otimes \rho_2) + (1-t)(\rho_1' \otimes \rho_2') = \big(t\rho_1 + (1-t)\rho_1'\big) \otimes \big(t\rho_2 + (1-t)\rho_2'\big)$$

as elements of $\ell^\infty(\mathcal{F}_0)$, so that $\mathcal{P}_0$ is covex as a subset of $\ell^\infty(\mathcal{F})$. It is not difficult to see from Corollary 9 that

$$\begin{aligned} |\delta(\mu_0 \otimes \nu_0) - \delta(\mu_1 \otimes \nu_1)| &\lesssim \overline{\mathsf{W}}_1(\mu_0, \mu_1) + \overline{\mathsf{W}}_1(\nu_0, \nu_1) \\ &\le 2\|(\mu_0 \otimes \nu_0) - (\mu_1 \otimes \nu_1)\|_{\infty, \mathcal{F}}. \end{aligned}$$

Also, for $\mu_t = (1-t)\mu + t\rho_1$ and $\nu_t = (1-t)\nu + t\rho_2$ with $\rho_1 \otimes \rho_2 \in \mathcal{P}_0$, arguing as in the proof of Lemma 45 and using the fact that $[\varphi^\theta]^{cc} = \varphi^\theta$, we have

$$\begin{aligned} &\lim_{t \downarrow 0} \frac{\delta(\mu_t \otimes \nu_t) - \delta(\mu \otimes \nu)}{t} \\ &= \int_{\mathbb{S}^{d-1}} \Big[(\rho_1 - \mu)(\varphi^\theta \circ \mathfrak{p}^\theta) + (\rho_2 - \nu)([\varphi^\theta]^c \circ \mathfrak{p}^\theta)\Big] d\sigma(\theta) \\ &= \int_{\mathbb{S}^{d-1}} \int (\varphi^\theta \circ \mathfrak{p}^\theta) \oplus ([\varphi^\theta]^c \circ \mathfrak{p}^\theta) d(\rho_1 \otimes \rho_2 - \mu \otimes \nu) d\sigma(\theta). \end{aligned}$$

Here recall that $\mu_t \otimes \nu_t$ can be identified with $(1-t)(\mu \otimes \nu) + t(\rho_1 \otimes \rho_2)$ as elements of $\ell^\infty(\mathcal{F})$. Let $\psi^\theta = [\varphi^\theta]^c$. Again, choose versions of φ^θ and ψ^θ such that, for some constant c independent of θ, $\bar{\varphi}^\theta = (\varphi^\theta - \varphi^\theta(0))/\mathsf{c}$, $\bar{\psi}^\theta = (\psi^\theta - \psi^\theta(0))/\mathsf{c} \in \mathsf{Lip}_{1,0}(\mathbb{R})$ for every $\theta \in \mathbb{S}^{d-1}$. Then, the derivative $\delta'_{\mu\otimes\nu}$ extends to $\mathcal{T}_{\mathcal{P}_0}(\mu \otimes \nu)$ as

$$\delta'_{\mu\otimes\nu}(m) = \int_{\mathbb{S}^{d-1}} \mathsf{c} \cdot m((\bar{\varphi}^\theta \circ \mathfrak{p}^\theta) \oplus (\bar{\psi}^\theta \circ \mathfrak{p}^\theta)) d\sigma(\theta).$$

We shall verify that the process $\sqrt{n}(\hat{\mu}_n \otimes \hat{\nu}_n - \mu \otimes \nu)$ converges weakly to a tight Gaussian process in $\ell^\infty(\mathcal{F})$. Since $\mathcal{F}_0$ is Donsker w.r.t. μ and ν, and X_i's and Y_i's are independent, by Example 1.4.6 in [201] (combined with Lemma 3.2.4 in [73] concerning measurable covers), we see that

$$\big(\sqrt{n}(\hat{\mu}_n - \mu), \sqrt{n}(\hat{\nu}_n - \nu)\big) \xrightarrow{d} (G_\mu, G'_\nu) \quad \text{in } \ell^\infty(\mathcal{F}_0) \times \ell^\infty(\mathcal{F}_0), \tag{A.8}$$

where G_μ and G'_ν are independent and the weak limits of $\sqrt{n}(\hat{\mu}_n - \mu)$ and $\sqrt{n}(\hat{\nu}_n - \nu)$ in $\ell^\infty(\mathcal{F}_0)$, respectively. Since the map $\ell^\infty(\mathcal{F}_0) \times \ell^\infty(\mathcal{F}_0) \ni (z_1, z_2) \mapsto (z_1(f_1) + z_2(f_2))_{f_1 \oplus f_2 \in \mathcal{F}} \in$

$\ell^\infty(\mathcal{F})$ is continuous, we have

$$\sqrt{n}(\hat{\mu}_n \otimes \hat{\nu}_n - \mu \otimes \nu) \xrightarrow{d} G_{\mu\otimes\nu} \quad \text{in } \ell^\infty(\mathcal{F})$$

where $G_{\mu\otimes\nu}(f_1 \oplus f_2) = G_\mu(f_1) + G'_\nu(f_2)$ for $f = f_1 \oplus f_2 \in \mathcal{F}$. Applying Proposition 1 (see also Remark 2), we conclude that

$$\begin{aligned}\sqrt{n}(\delta(\hat{\mu}_n \otimes \hat{\nu}_n) - \delta(\mu \otimes \nu)) &\xrightarrow{d} \int_{\mathbb{S}^{d-1}} \left[\mathsf{c} \cdot G_\mu(\bar{\varphi}^\theta \circ \mathfrak{p}^\theta) + \mathsf{c} \cdot G'_\nu(\bar{\psi}^\theta \circ \mathfrak{p}^\theta)\right] d\sigma(\theta) \\ &= \int_{\mathbb{S}^{d-1}} \mathbb{G}_\mu d\sigma + \int_{\mathbb{S}^{d-1}} \mathbb{G}'_\nu d\sigma.\end{aligned}$$

where $\mathbb{G}_\mu$ is given before and $\mathbb{G}'_\nu$ is defined as $\mathbb{G}'_\nu(\theta) = \mathsf{c} \cdot G'_\nu(\bar{\psi}^\theta \circ \mathfrak{p}^\theta)$. Arguing as in Part (i), the right-hand side is Gaussian with mean zero and variance $v_p^2 + w_p^2$.

Second, the fact that $v_p^2 + w_p^2$ coincides with the semiparametric efficiency bound follows from Corollary 2. The argument is similar to the one-sample case and omitted for brevity.

Finally, we show the bootstrap consistency. The argument is analogous to the one-sample case above. To apply Theorem 23.9 in [200], we need to verify that, as maps into $\ell^\infty(\mathcal{F})$, the sequence $\sqrt{n}(\hat{\mu}_n^B \otimes \hat{\nu}_n^B - \hat{\mu}_n \otimes \hat{\nu}_n)$ is asymptotically measurable and converges conditionally in distribution to $(G_\mu(f_1) + G'_\nu(f_2))_{f_1\oplus f_2\in\mathcal{F}}$ given $(X_1, Y_1), (X_2, Y_2), \dots$. Since $\mathcal{F}_0$ is Donsker w.r.t. μ and ν, each of $\sqrt{n}(\hat{\mu}_n^B - \hat{\mu}_n)$ and $\sqrt{n}(\hat{\nu}_n^B - \hat{\nu}_n)$ is asymptotically measurable and converges conditionally in distribution (cf. Chapter 3.6 in [201]). By Lemma 1.4.4 and Example 1.4.6 in [201], as maps into $\ell^\infty(\mathcal{F}_0) \times \ell^\infty(\mathcal{F}_0)$, the sequence $(\sqrt{n}(\hat{\mu}_n^B - \hat{\mu}_n), \sqrt{n}(\hat{\nu}_n^B - \hat{\nu}_n))$ is asymptotically measurable and converges conditionally in distribution to (G_μ, G'_ν). Since the map $\ell^\infty(\mathcal{F}_0) \times \ell^\infty(\mathcal{F}_0) \ni (z_1, z_2) \mapsto (z_1(f_1) + z_2(f_2))_{f_1\oplus f_2\in\mathcal{F}} \in \ell^\infty(\mathcal{F})$ is continuous, we see that, as maps into $\ell^\infty(\mathcal{F})$, $\sqrt{n}(\hat{\mu}_n^B \otimes \hat{\nu}_n^B - \hat{\mu}_n \otimes \hat{\nu}_n)$ is asymptotically measurable and converges conditionally in distribution to $(G_\mu(f_1) + G'_\nu(f_2))_{f_1\oplus f_2\in\mathcal{F}}$, as desired. The rest follows from Corollary 1 and Theorem 23.9 in [200]. □

Proof of Theorem 2

Given the Gâteaux derivative result for $\overline{\mathsf{W}}_p$ (cf. Lemma 45), the proof is analogous to Theorem 1. We omit the details for brevity. $\square$

Proof of Lemma 3

We follow the notation used in the proof of Theorem 1(ii). Let $S_n^B = \sqrt{n}(\overline{\mathsf{W}}_p^p(\hat{\mu}_{m,n}^B, \hat{\nu}_{m,n}^B) - \overline{\mathsf{W}}_p^p(\hat{\mu}_n, \hat{\nu}_n))$ and S be a random variable following the limit law in (2.6). By Theorem 3.6.2 combined with Example 1.4.6 in [201], we have

$$(\sqrt{m}(\hat{\mu}_{m,n}^B - \hat{\mu}_n), \sqrt{m}(\hat{\nu}_{m,n}^B - \hat{\nu}_n) \xrightarrow{d} (G_\mu, G_\nu) \quad \text{in } \ell^\infty(\mathcal{F}_0) \times \ell^\infty(\mathcal{F}_0)$$

given almost every sequence $X_1, X_2, \ldots, Y_1, Y_2, \ldots$ Since $\mathbb{E}[\|\sqrt{m}(\hat{\mu}_n - \mu)\|_{\infty,\mathcal{F}_0} \vee \|\sqrt{m}(\hat{\mu}_n - \mu)\|_{\infty,\mathcal{F}_0}] = O(\sqrt{m/n}) = o(1)$ by Theorem 2.14.1 in [201] and the assumption that $m = o(n)$, we have $\|\sqrt{m}(\hat{\mu}_n - \mu)\|_{\infty,\mathcal{F}_0} \vee \|\sqrt{m}(\hat{\nu}_n - \mu)\|_{\infty,\mathcal{F}_0} \to 0$ in probability. Pick any subsequence n' of n and choose a further subsequence n'' such that $\sqrt{m''}(\overline{\mathsf{W}}_p^p(\hat{\mu}_{n''}, \hat{\nu}_{n''}) - \overline{\mathsf{W}}_p^p(\mu, \nu)) \to 0$ a.s. (cf. Theorem 2(ii)) and $\|\sqrt{m''}(\hat{\mu}_{n''} - \mu)\|_{\infty,\mathcal{F}_0} \vee \|\sqrt{m''}(\hat{\nu}_{n''} - \mu)\|_{\infty,\mathcal{F}_0} \to 0$ a.s. with $m'' = m_{n''}$. Arguing as in the proof of Theorem 1(ii), we have

$$\sqrt{m}(\hat{\mu}_{m'',n''}^B \otimes \hat{\nu}_{m'',n''}^B - \mu \otimes \nu) \xrightarrow{d} G_{\mu\otimes\nu} \quad \text{in } \ell^\infty(\mathcal{F})$$

given almost every sequence $X_1, X_2, \ldots, Y_1, Y_2, \ldots$ By directional Hadamard differentiability of $\delta(\rho_1 \otimes \rho_2) = \overline{\mathsf{W}}_p^p(\rho_1, \rho_2)$ at $\rho_1 \otimes \rho_2 = \mu \otimes \nu$, we have

$$\underbrace{\sqrt{m''}(\delta(\hat{\mu}_{m'',n''}^B \otimes \hat{\nu}_{m'',n''}^B) - \delta(\mu \otimes \nu))}_{= \sqrt{m''}\left(\overline{\mathsf{W}}_p^p(\hat{\mu}_{m'',n''}^B, \hat{\nu}_{m'',n''}^B) - \overline{\mathsf{W}}_p^p(\mu,\nu)\right)} \xrightarrow{d} \delta'_{\mu\otimes\nu}(G_{\mu\otimes\nu}) \stackrel{d}{=} S$$

given almost every sequence $X_1, X_2, \ldots, Y_1, Y_2, \ldots$ Since $\sqrt{m''}(\overline{\mathsf{W}}_p^p(\hat{\mu}_{n''}, \hat{\nu}_{n''}) - \overline{\mathsf{W}}_p^p(\mu, \nu)) \to 0$ a.s. by construction, we conclude that

$$S_{n''}^B = \sqrt{m''}(\overline{\mathsf{W}}_p^p(\hat{\mu}_{m'',n''}^B, \hat{\nu}_{m'',n''}^B) - \overline{\mathsf{W}}_p^p(\hat{\mu}_{n''}, \hat{\nu}_{n''})) \xrightarrow{d} S$$

given almost every sequence $X_1, X_2, \dots, Y_1, Y_2, \dots$, that is

$$\sup_{g \in \mathsf{BL}_1(\mathbb{R})} \left| \mathbb{E}^B[g(S^B_{n''})] - \mathbb{E}[g(S)] \right| \to 0$$

a.s. This completes the proof. □

A.2.2 Proof of Theorem 3

Preliminaries

Lemma 49 (CLT in L^1 space)**.** Let $(S, \mathcal{S}, \mu)$ be a σ-finite measure space with $\mathcal{S}$ being countably generated, and let $Z_1, Z_2, \dots$ be i.i.d. $L^1(\mu)$-valued random variables. If $\mathbb{P}(\|Z_1\|_{L^1(\mu)} > t) = o(t^{-2})$ as $t \to \infty$ and $\int_S \sqrt{\mathbb{E}[Z_1^2(x)]}\, d\mu(x) < \infty$, then there exists a centered jointly measurable Gaussian process G with paths in $L^1(\mu)$ such that $n^{-1/2} \sum_{i=1}^n Z_i \xrightarrow{d} \mathsf{G}$ in $L^1(\mu)$.

Proof. See [129], Theorem 10.10. □

Lemma 50 (Directional derivative of L^1 norm)**.** Let $(S, \mathcal{S}, \mu)$ be a measure space. Then, for any $f, g \in L^1(\mu)$ and $t_n \downarrow 0$, we have

$$\lim_{n\to\infty} \frac{\|f + t_n g\|_{L^1(\mu)} - \|f\|_{L^1(\mu)}}{t_n} = \int [\mathrm{sign}(f)] g \, d\mu + \int_{\{f=0\}} |g| \, d\mu.$$

Proof of Lemma 50. Let $S^+ = \{f > 0\}$ and $S^- = \{f < 0\}$, and set $h = \frac{g}{|f|} \mathbb{1}_{\{f \neq 0\}}$ and $d\gamma = |f| d\mu$. Observe that

$$\begin{aligned} &t_n^{-1}(\|f + t_n g\|_{L^1(\mu)} - \|f\|_{L^1(\mu)}) \\ &\quad = t_n^{-1} \int_{S^+} (|1 + t_n h| - 1)\, d\gamma + t_n^{-1} \int_{S_-} (|-1 + t_n h| - 1)\, d\gamma + \int_{\{f=0\}} |g| \, d\mu. \end{aligned}$$

We next show that the first and second terms converge to $\int_{\{f>0\}} g d\mu$ and $\int_{\{f<0\}} (-g) d\mu$, respectively. The arguments are analogous, so we only prove the former.

By normalization, we may assume without loss of generality that $\gamma(S^+) = 1$. Consider the set $A_n = \{t_n|h| < 1\}$. Since $h \in L^1(\gamma)$, by Markov's inequality we have $\gamma(A_n^c \cap S^+) \to 0$. Consequently

$$\left| \int_{S^+} |1 + t_n h| d\gamma - \int_{S^+} |1 + t_n h \mathbb{1}_{A_n}| d\gamma \right| \le t_n \int_{S^+} |h| \mathbb{1}_{A_n^c} d\gamma = o(t_n).$$

Since $1 + t_n h \mathbb{1}_{A_n} \ge 0$, we have

$$\int_{S^+} (|1 + t_n h \mathbb{1}_{A_n}| - 1) d\gamma = t_n \int_{S^+} h \mathbb{1}_{A_n} d\gamma = t_n \int_{S^+} h d\gamma + o(t_n).$$

As $\int_{S^+} h d\gamma = \int_{\{f>0\}} g d\mu$, we obtain the desired result. □

Lemma 51 (Directional derivative of supremum functional)**.** Let S be a nonempty set. Consider the functional $\phi : \ell^\infty(S) \to \mathbb{R}$ defined by $\phi(f) = \sup_{x \in S} f(x)$. Suppose that there exists a pseudometric d on S such that (S, d) is totally bounded, and let $C_u(S, \mathsf{d})$ denote the set of all uniformly d-continuous functions on S. Then ϕ is Hadamard directionally differentiable at $f \in C_u(S, \mathsf{d})$ with derivative

$$\phi'_f(g) = \sup_{x \in \overline{S} : f(x) = \phi(f)} g(x), \quad g \in C_u(S, \mathsf{d}),$$

where $\overline{S}$ is the completion of S with respect to d [note that every $f \in C_u(S, \mathsf{d})$ has the unique continuous extension to $\overline{S}$, which we denote by the same symbol f].

Proof. See [35], Corollary 2.5. □

Proof of Theorem 3

Part (i). Recall the expression $\underline{\mathsf{W}}_1(\hat{\mu}_n, \mu) = \|F_{\hat{\mu}_n} - F_\mu\|_{L^1(\lambda \otimes \sigma)}$. We divide the proof into two steps.

<u>Step 1</u>. We will first show that there exists a centered, jointly measurable Gaussian process $\mathsf{G}_\mu = (\mathsf{G}_\mu(t, \theta))_{(t,\theta) \in \mathbb{R} \times \mathbb{S}^{d-1}}$ with paths in $L^1(\lambda \otimes \sigma)$ and covariance function (2.9)

such that

$$\sqrt{n}(F_{\hat{\mu}_n}(t;\theta) - F_\mu(t;\theta))_{(t,\theta)\in\mathbb{R}\times\mathbb{S}^{d-1}} \xrightarrow{d} \mathsf{G}_\mu \quad \text{in } L^1(\lambda\otimes\sigma).$$

We will apply the CLT in the L^1 space (Lemma 49) to prove this claim. Let $Z_i(t,\theta) = \mathbb{1}(\theta^\intercal X_i \le t) - F_\mu(t;\theta)$. Then $\sqrt{n}(F_{\hat{\mu}_n}(t;\theta) - F_\mu(t;\theta)) = n^{-1/2}\sum_{i=1}^n Z_i(t,\theta)$. Since $((t,\theta),\omega) \mapsto Z_i(t,\theta,\omega) = \mathbb{1}(\theta^\intercal X_i(\omega) \le t) - F_\mu(t;\theta)$ is (jointly) measurable, Z_i can be regarded an an $L^1(\lambda\otimes\sigma)$-valued random variable as long as $\|Z_i\|_{L^1(\lambda\otimes\sigma)} < \infty$ a.s. [33], which will be verified below.

Observe that $\|Z_i\|_{L^1(\lambda\otimes\sigma)}$ can be evaluated as

$$\begin{aligned}
&\int_{\mathbb{S}^{d-1}}\int_{\mathbb{R}} |\mathbb{1}(\theta^\intercal X_i \le t) - \mathbb{P}(\theta^\intercal X_i \le t)|\, dt\, d\sigma(\theta) \\
&= \int_{\mathbb{S}^{d-1}}\Bigg(\int_{-\infty}^0 |\mathbb{1}(\theta^\intercal X_i \le t) - \mathbb{P}(\theta^\intercal X_i \le t)|\, dt \\
&\qquad\qquad + \int_0^\infty |\mathbb{1}(\theta^\intercal X_i > t) - \mathbb{P}(\theta^\intercal X_i > t)|\, dt\Bigg) d\sigma(\theta) \\
&\le \int_{\mathbb{S}^{d-1}}\Bigg[(-\theta^\intercal X_i \vee 0) + \int_{-\infty}^0 \mathbb{P}(\theta^\intercal X_i \le t)\, dt \\
&\qquad\qquad + (\theta^\intercal X_i \vee 0) + \int_0^\infty \mathbb{P}(\theta^\intercal X_i > t)\, dt\Bigg] d\sigma(\theta) \\
&\le \|X_i\| + \mathbb{E}[\|X_i\|].
\end{aligned}$$

This implies that $Z_i \in L^1(\lambda\otimes\sigma)$. Further, for $t > 2\mathbb{E}[\|X_i\|]$,

$$\begin{aligned}
\mathbb{P}(\|Z_i\|_{L^1(\lambda\otimes\sigma)} > t) &\le \mathbb{P}\left(\|X_i\| + \mathbb{E}[\|X_i\|] > t\right) \\
&\le \mathbb{P}\left(\|X_i\| > t/2\right) \\
&\le \frac{2^{2+\epsilon}\mathbb{E}[\|X_i\|^{2+\epsilon}]}{t^{2+\epsilon}} = o(t^{-2})
\end{aligned}$$

as $t \to \infty$. Finally, letting $K = \sup_{\theta \in \mathbb{S}^{d-1}} \mathbb{E}[|\theta^\intercal X_i|^{2+\epsilon}] < \infty$, we have

$$\begin{aligned}
&\int_{\mathbb{S}^{d-1}} \int_{\mathbb{R}} \sqrt{\mathbb{E}[Z_i(t,\theta)^2]}\, dt\, d\sigma(\theta) \\
&= \int_{\mathbb{S}^{d-1}} \int_{-\infty}^{\infty} \sqrt{F_\mu(t;\theta)(1 - F_\mu(t;\theta))}\, dt\ d\sigma(\theta) \\
&\le \int_{\mathbb{S}^{d-1}} \left(\int_{-\infty}^{0} \sqrt{F_\mu(t;\theta)}\, dt\ + \int_0^\infty \sqrt{1 - F_\mu(t;\theta)}\, dt \right) d\sigma(\theta) \\
&\le \int_{\mathbb{S}^{d-1}} \left(2 \int_0^\infty \sqrt{\mathbb{P}(|\theta^\intercal X_i| \ge t)}\, dt \right) d\sigma(\theta) \\
&\le 2 \left(1 + \int_1^\infty \sqrt{\frac{K}{t^{2+\epsilon}}}\, dt \right) \\
&\le 2 + 4\frac{\sqrt{K}}{\epsilon} < \infty.
\end{aligned} \tag{A.9}$$

Thus, by Lemma 49, we have $n^{-1/2} \sum_{i=1}^n Z_i \xrightarrow{d} \mathsf{G}_\mu$ in $L^1(\lambda \otimes \sigma)$.

Step 2. We will apply the extended functional delta method to derive the limit distribution for $\sqrt{n}(\underline{\mathsf{W}}_1(\hat{\mu}_n, \nu) - \underline{\mathsf{W}}_1(\mu, \nu))$. Let $\phi : L^1(\lambda \otimes \sigma) \to \mathbb{R}$ be defined by

$$\phi(f) = \|f\|_{L^1(\lambda \otimes \sigma)},\ f \in L^1(\lambda \otimes \sigma).$$

By Lemma 50 and the fact that ϕ is trivially Lipshitz in $\|\cdot\|_{L^1(\lambda\otimes\sigma)}$, ϕ is Hadamard directionally differentialble at $f \in L^1(\lambda \otimes \sigma)$ with derivative

$$\phi'_f(g) = \int [\mathrm{sign}(f)] g d(\lambda \otimes \sigma) + \int_{\{f=0\}} |g| d(\lambda \otimes \sigma).$$

Thus, by the extended functional delta method, we have

$$\begin{aligned}
\sqrt{n}(\underline{\mathsf{W}}_1(\hat{\mu}_n, \nu) - \underline{\mathsf{W}}_1(\mu, \nu)) &= \sqrt{n}(\phi(F_\mu - F_\nu + F_{\hat{\mu}_n} - F_\mu) - \phi(F_\mu - F_\nu)) \\
&\xrightarrow{d} \phi'_{F_\mu - F_\nu}(\mathsf{G}_\mu).
\end{aligned}$$

The right-hand side coincides with the limit distribution in (2.10).

Part (ii). Observe that $\underline{\mathsf{W}}_1(\hat{\mu}_n, \hat{\nu}_n) = \|F_{\hat{\mu}_n} - F_{\hat{\nu}_n}\|_{L^1(\lambda\otimes\sigma)}$. From the proof of Part (i), we have $\sqrt{n}(F_{\hat{\mu}_n} - F_{\hat{\nu}_n} - F_\mu + F_\nu) \xrightarrow{d} \mathsf{G}_\mu - \mathsf{G}'_\nu$ in $L^1(\lambda \otimes \sigma)$. The rest of the proof is analogous to Step 2 in the proof of Part (i).

Part (iii). Observe that $\overline{\mathsf{W}}_1(\mu,\nu) = \sup_{f\in\mathcal{F}}(\mu-\nu)(f)$ by the Kantorovich-Rubinstein duality. By Lemma 46, the function class $\mathcal{F}$ is μ-Donsker under the given moment condition, $\sqrt{n}(\hat{\mu}_n - \mu) \stackrel{d}{\to} G_\mu$ in $\ell^\infty(\mathcal{F})$.

Then, we apply the functional delta method to the supremum functional, Lemma 51. Since, relative to the standard deviation metric, $f \mapsto \int f d(\mu-\nu) = \mu(f-\nu(f))$ is uniformly continuous, $\mathcal{F}$ is totally bounded, G_μ has uniformly continuous paths (cf. Example 1.5.10 in [201]), we obtain the desired result.

Part (iv). The proof is analogous to Part (iii) and thus omitted. □

A.2.3 Proof of Corollary 3

The limit distribution results directly follow from Theorem 3. Under the setting of Corollary 3, inspection of the proof of Theorem 3 shows that the map $f \mapsto \|f\|_{L^1(\lambda\otimes\lambda)}$ is Hadamard differentiable at $f = F_\mu - F_\nu$ with derivative

$$\phi'_{F_\mu - F_\nu}(g) = \int [\mathrm{sign}(F_\mu - F_\nu)] g d(\lambda\otimes\sigma),\ g \in L^1(\lambda\otimes\sigma).$$

Combined with the equivalence between the CLT and the bootstrap consistency for i.i.d. random variables with values in a separable Banach space (cf. Remark 2.5 in [97]), the bootstrap consistency result for $\underline{\mathsf{W}}_1$ (like that in Theorem 1) follows via Theorem 23.9 in [200].

Regarding semiparametric efficiency, for any bounded measurable function h on $\mathbb{R}^d$

with μ-mean zero,

$$
\begin{aligned}
&\frac{1}{s}(\|F_{(1+sh)\mu} - F_\nu\|_{L^1(\lambda\otimes\sigma)} - \|F_\mu - F_\nu\|_{L^1(\lambda\otimes\sigma)}) \\
&\to \int_{\mathbb{R}\times\mathbb{S}^{d-1}} [\mathrm{sign}(F_\mu - F_\nu)(t;\theta)] \left\{ \int h(x)\mathbb{1}(x^\intercal\theta \le t) d\mu(x) \right\} dt d\sigma(\theta) \quad (s \downarrow 0) \\
&= \int_{\mathbb{R}\times\mathbb{S}^{d-1}} [\mathrm{sign}(F_\mu - F_\nu)(t;\theta)] \left\{ \int h(x)\mathfrak{g}(x,t,\theta) d\mu(x) \right\} dt d\sigma(\theta) \\
&= \left\langle h, \int_{\mathbb{R}\times\mathbb{S}^{d-1}} [\mathrm{sign}(F_\mu - F_\nu)(t;\theta)]\mathfrak{g}(\cdot,t,\theta) dt d\sigma(\theta) \right\rangle_{L^2(\mu)},
\end{aligned}
$$

where $\mathfrak{g}(x,t,\theta) = \mathbb{1}(x^\intercal\theta \le t) - \mathbb{1}(t > 0)$ and the second equality follows as h has μ-mean zero. Hence, the semiparametric efficiency bound for estimating $\rho \mapsto \underline{\mathsf{W}}_1(\rho,\nu)$ at $\rho = \mu$ is given by the μ-variance of

$$
\int_{\mathbb{R}\times\mathbb{S}^{d-1}} [\mathrm{sign}(F_\mu - F_\nu)(t;\theta)]\mathfrak{g}(\cdot,t,\theta) dt d\sigma(\theta),
$$

which agrees with v_1^2 above. Likewise, in the two-sample case, it is shown that the semiparametric efficiency bound agrees with $v_1^2 + w_1^2$; see the proof of Corollary 2. □

A.3 Proofs for Section 2.5

A.3.1 Proof of Proposition 3

Lemma 5.2 of [175] establishes that convolution acts as a contraction for W_p, $\mathsf{W}_p^\sigma \le \mathsf{W}_p$. For any $\rho \in \mathcal{P}(\mathbb{R}^d)$, the coupling $\pi := \chi_\sigma(x-y)d\lambda(x)d\rho(y) \in \Pi(\rho, \rho * \eta_\sigma)$ satisfies

$$
\int_{\mathbb{R}^d\times\mathbb{R}^d} \|x-y\|^p d\pi(x,y) = \sigma^p \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} \|z\|^p \chi_1(z) d\lambda(z) d\mu(y) = \sigma^p \mathbb{E}_{\eta_1}[\|X\|]^p,
$$

so $\mathsf{W}_p(\mu, \mu * \eta_\sigma) \le (\mathbb{E}[\|X_{\eta_1}\|^p])^{1/p}$ by Lemma 7.1.10 in [6]. Conclude by applying the triangle inequality, $\mathsf{W}_p \le \mathsf{W}_p^\sigma + 2(\mathbb{E}_{\eta_1}[\|X\|^p])^{1/p}$. □

A.3.2 Proof of Proposition 5

Lemma 52. For $\sigma > 0$, the characteristic function ϕ_{η_σ} of η_σ is zero on a set of negligible (Lebesgue) measure.

Proof. The Fourier-Laplace transform of χ_σ is defined by

$$\widehat{\chi_\sigma} : t \in \mathbb{C}^d \mapsto \int_{\mathbb{R}^d} e^{-it^\intercal x} \chi_\sigma(x) dx.$$

By Theorem 7.1.14 in [118], $\widehat{\chi_\sigma}$ is complex-analytic, so $\phi_{\eta_\sigma} = \widehat{\chi_\sigma}|_{\mathbb{R}^d}(-\cdot)$ is real analytic. The conclusion follows since the set of zeros of any real analytic function on $\mathbb{R}^d$ that is not identically zero has zero Lebesgue measure [80, p.240]. □

Lemma 53. For $p \geq 1$, $\sigma > 0$, $(\mu_k)_{k\in\mathbb{N}} \subset \mathcal{P}_p(\mathbb{R}^d)$ and $\mu \in \mathcal{P}_p(\mathbb{R}^d)$, $\mu_k * \eta_\sigma \xrightarrow{w} \mu * \eta_\sigma$ as $k \to \infty$ if and only if $\mu_k \xrightarrow{w} \mu$ as $k \to \infty$.

Proof. First, if $\mu_k \xrightarrow{w} \mu$, $\phi_{\mu_k * \eta_\sigma} = \phi_{\mu_k}\phi_{\eta_\sigma} \to \phi_\mu \phi_{\eta_\sigma} = \phi_{\mu*\eta_\sigma}$ pointwise, so $\mu_k * \eta_\sigma \xrightarrow{w} \mu * \eta_\sigma$. To prove the opposite direction, if $\mu_k * \eta_\sigma \xrightarrow{w} \mu * \eta_\sigma$, arguing as above yields $\phi_{\mu_k} \to \phi_\mu$ pointwise almost everywhere, as $\phi_{\eta_\sigma} = 0$ on a set of negligible (Lebesgue) measure by Lemma 52. It is not difficult to show that pointwise almost everywhere convergence of characteristic functions implies pointwise convergence thereof and hence that $\mu_k \xrightarrow{w} \mu$. □

Proof of Proposition 5. The fact that W_p^σ is a metric follows essentially from the proof of Proposition 1 in [152], with a slight modification using Lemma 52 and the fact that characteristic functions are uniformly continuous. Similarly, the fact that W_p^σ and W_p induce the same topology on $\mathcal{P}_p(\mathbb{R}^d)$ follows from the proof of Proposition 1 in [152] along with Lemma 53. □

A.3.3 Proof of Theorem 4

We only prove the one-sample result, as the two-sample result follows from essentially the same reasoning coupled with the fact that φ^c is the unique OT potential from $\nu * \eta_\sigma$ to $\mu * \eta_\sigma$ for W_p up to additive constants under the assumption that ν has connected support; see also the proof of Theorem 1 Part (ii). In what follows, we fix $\sigma > 0$ and $1 < p < \infty$. Recall that B denotes the unit ball in $L^2(\mathcal{X}_\sigma)$. We first prove the weak convergence of the smoothed empirical process $\sqrt{n}(\hat{\mu}_n - \mu) * \eta_\sigma$ and the associated bootstrap process in $\ell^\infty(B)$.

Lemma 54. For any $\mu \in \mathcal{P}(\mathcal{X})$, we have $\sqrt{n}(\hat{\mu}_n - \mu) * \eta_\sigma \xrightarrow{d} \mathbb{G}_\mu$ in $\ell^\infty(B)$, where $\mathbb{G}_\mu = \big(\mathbb{G}_\mu(f)\big)_{f \in B}$ is a tight centered Gaussian process in $\ell^\infty(B)$ with the covariance function $\mathrm{Cov}(\mathbb{G}_\mu(f), \mathbb{G}_\mu(g)) = \mathrm{Cov}_\mu(f * \chi_\sigma, g * \chi_\sigma)$. Further, conditionally on $X_1, X_2, \dots,$ $\sqrt{n}(\hat{\mu}_n^B - \hat{\mu}_n) * \eta_\sigma \xrightarrow{d} \mathbb{G}_\mu$ in $\ell^\infty(B)$ for almost every realization of $X_1, X_2, \dots$

Proof. We will show weak convergence of $\sqrt{n}(\hat{\mu}_n - \mu) * \eta_\sigma$ in $(L^2(\mathcal{X}_\sigma))'$, the topological dual of $L^2(\mathcal{X}_\sigma)$. Observe that, for every $f \in L^2(\mathcal{X}_\sigma)$ and $x \in \mathrm{spt}(\mu)$,

$$|(f * \chi_\sigma)(x)| \le \|f\|_{L^2(\mathcal{X}_\sigma)} \|\chi_\sigma(x - \cdot)\|_{L^2(\mathcal{X}_\sigma)} \le C\|f\|_{L^2(\mathcal{X}_\sigma)}, \tag{A.10}$$

where $C = \|\chi_\sigma\|_\infty \sqrt{\lambda(B(0,\sigma))} < \infty$.

Set $Z_i := \delta_{X_i} * \eta_\sigma$ for $i \in \mathbb{N}$. The inequality (A.10) implies that $\|Z_i\|_{(L^2(\mathcal{X}_\sigma))'} \le C$. It is not difficult to see that the Borel σ-field on $(L^2(\mathcal{X}_\sigma))'$ agrees with the smallest σ-field for which the projections $z \mapsto z(f), f \in L^2(\mathcal{X}_\sigma)$ are measurable. Thus, Z_i can be thought as i.i.d. random variables with values in $(L^2(\mathcal{X}_\sigma))'$ with (Bochner) expectation $\mathbb{E}[Z_i] = \mu * \eta_\sigma$ (cf. Lemma 5.1.1 in [188] for Bochner integrals). Since $(L^2(\mathcal{X}_\sigma))'$ is a separable Hilbert space, by Example 1.8.5 in [201], we have $\sqrt{n}(\hat{\mu}_n - \mu) * \eta_\sigma = n^{-1/2} \sum_{i=1}^n (Z_i - \mathbb{E}[Z_i]) \xrightarrow{d} \mathbb{G}_\mu^\circ$, where $\mathbb{G}_\mu^\circ$ is a centered Gaussian random variable in $\big(L^2(\mathcal{X}_\sigma)\big)'$.

In addition, let $Z_i^B = \delta_{X_i^B} * \eta_\sigma$ for $i = 1, \dots, n$. Then, $Z_1^B, \dots, Z_n^B$ are an i.i.d. sample from the empirical distribution $n^{-1}\sum_{i=1}^n \delta_{Z_i}$, so that by Remark 2.5 in [97], we have that, conditionally on $X_1, X_2, \dots,$ $\sqrt{n}(\hat{\mu}_n^B - \hat{\mu}_n) * \eta_\sigma = n^{-1/2}\sum_{i=1}^n (Z_i^B - \overline{Z}_n) \xrightarrow{d} \mathbb{G}_\mu$ in $\left(L^2(\mathcal{X}_\sigma)\right)'$ for almost every realization of $X_1, X_2, \dots$, where $\overline{Z}_n = n^{-1}\sum_{i=1}^n Z_i$.

The conclusion of the lemma follows by noting that the map $(L^2(\mathcal{X}_\sigma))' \ni z \mapsto z|_B \in \ell^\infty(B)$ is an isometry. $\square$

Second, we shall establish Lipschitz continuity of $\mathsf{V}_p^\sigma(\cdot, \nu)$ w.r.t. $\|\cdot * \eta_\sigma\|_{\infty,B}$.

Lemma 55. For every $\mu_0, \mu_1 \in \mathcal{P}(\mathcal{X})$,

$$|\mathsf{V}_p^\sigma(\mu_1, \nu) - \mathsf{V}_p^\sigma(\mu_0, \nu)| \le \sqrt{\lambda(\mathcal{X}_\sigma)}\, \mathrm{diam}(\mathcal{X}_\sigma)^p \|(\mu_1 - \mu_0) * \eta_\sigma\|_{\infty,B}.$$

Proof. Recall $\mathsf{V}_p^\sigma(\mu_i, \nu) = [\mathsf{W}_p(\mu_i * \eta_\sigma, \nu * \eta_\sigma)]^p$. Let φ_i be an OT potential from $\mu_i * \eta_\sigma$ to $\nu * \eta_\sigma$ for W_p satisfying $0 \le \varphi_i \le \mathrm{diam}(\mathcal{X}_\sigma)^p$ on $\mathcal{X}_\sigma$ for $i = 0, 1$ as in Remark 1.13 of [207]. Then, by duality,

$$\begin{aligned} \mathsf{V}_p^\sigma(\mu_1, \nu) - \mathsf{V}_p^\sigma(\mu_0, \nu) &\le \int_{\mathcal{X}_\sigma} \varphi_1 d((\mu_1 - \mu_0) * \eta_\sigma) \\ &\le \|\varphi_1\|_{L^2(\mathcal{X}_\sigma)} \|(\mu_1 - \mu_0) * \eta_\sigma\|_{\infty,B}. \end{aligned}$$

Likewise,

$$\mathsf{V}_p^\sigma(\mu_1, \nu) - \mathsf{V}_p^\sigma(\mu_0, \nu) \ge -\|\varphi_0\|_{L^2(\mathcal{X}_\sigma)} \|(\mu_1 - \mu_0) * \eta_\sigma\|_{\infty,B},$$

so that

$$|\mathsf{V}_p^\sigma(\mu_1, \nu) - \mathsf{V}_p^\sigma(\mu_0, \nu)| \le (\|\varphi_0\|_{L^2(\mathcal{X}_\sigma)} \vee \|\varphi_1\|_{L^2(\mathcal{X}_\sigma)}) \|(\mu_1 - \mu_0) * \eta_\sigma\|_{\infty,B}.$$

The conclusion of the lemma follows by noting that

$$\|\varphi_0\|_{L^2(\mathcal{X}_\sigma)} \vee \|\varphi_1\|_{L^2(\mathcal{X}_\sigma)} \le \sqrt{\lambda(\mathcal{X}_\sigma)}\, \mathrm{diam}(\mathcal{X}_\sigma)^p$$

by construction. $\square$

The following lemma concerns the Gâteaux directional derivative of $\mathsf{V}_p^\sigma(\cdot, \nu)$.

Lemma 56. Let $\mu_0, \mu_1 \in \mathcal{P}(\mathcal{X})$ be such that $\mathrm{int}(\mathrm{spt}(\mu_0))$ is connected with negligible boundary and $\mathrm{spt}(\mu_1) \subset \mathrm{spt}(\mu_0)$. Then,

$$\lim_{t \downarrow 0} \frac{\mathsf{V}_p^\sigma(\mu_0 + t(\mu_1 - \mu_0), \nu) - \mathsf{V}_p^\sigma(\mu_0, \nu)}{t} = ((\mu_1 - \mu_0) * \eta_\sigma)(\varphi_0),$$

where φ_0 is an OT potential from $\mu * \eta_\sigma$ to $\nu * \eta_\sigma$ for W_p

The proof of Lemma 56 relies on the following technical lemma concerning the support of $\mu * \eta_\sigma$.

Lemma 57. For any $\mu \in \mathcal{P}(\mathcal{X})$, $\mathrm{int}(\mathrm{spt}(\mu * \eta_\sigma))$ has negligible boundary. Further, $\mathrm{int}(\mathrm{spt}(\mu * \eta_\sigma))$ is connected provided $\mathrm{spt}(\mu)$ is connected.

Proof of Lemma 57. By [136, p. 106],

$$\mathrm{spt}(\mu * \eta_\sigma) = \overline{\mathrm{spt}(\mu) + \overline{B(0, \sigma)}} = \cup_{x \in \mathrm{spt}(\mu)} \overline{B(x, \sigma)},$$

where the second equality is due to the fact that a union of closed balls with centers in a closed set is closed.

Let $\Gamma = \{z \in \mathbb{R}^d : \mathrm{dist}(z, \mathrm{spt}(\mu)) = \sigma\}$ and $\Lambda = \cup_{x \in \mathrm{spt}(\mu)} B(x, \sigma)$. Observe that, since $\mathrm{spt}(\mu)$ is closed, $\mathrm{dist}(z, \mathrm{spt}(\mu)) = \|z - x\|$ for some $x \in \mathrm{spt}(\mu)$. It is clear that $\Gamma \subset \partial\Lambda$. On the other hand, if $z \notin \Gamma \cup \Lambda$, then $\mathrm{dist}(z, \mathrm{spt}(\mu)) > \sigma$ in which case $z \in \mathbb{R}^d \setminus \cup_{x \in \mathrm{spt}(\mu)} \overline{B(x, \sigma)}$. Since $\cup_{x \in \mathrm{spt}(\mu)} \overline{B(x, \sigma)}$ is closed, we have $z \notin \partial\Lambda$ and $\Gamma = \partial\Lambda$. The same argument implies that $\cup_{x \in \mathrm{spt}(\mu)} \overline{B(x, \sigma)} \subset \Gamma \cup \Lambda$ and the other inclusion is trivial, so $\cup_{x \in \mathrm{spt}(\mu)} B(x, \sigma) \subset \mathrm{int}(\mathrm{spt}(\mu * \eta_\sigma))$ and the closure of both sets coincide with $\mathrm{spt}(\mu * \eta_\sigma)$.

Example 1 in [171] implies that Λ has a (locally) Lipschitz boundary, which is negligible by Section 4.9 and Section 4.11 of [1]. As such, $\mathrm{int}(\mathrm{spt}(\mu * \eta_\sigma))$ has negligible boundary.

Assume that Λ is not connected, then there exist nonempty sets $A, B \subset \mathrm{spt}(\mu)$ such that $A \cup B = \mathrm{spt}(\mu)$, and $\cup_{x \in A} B(x, \sigma)$ and $\cup_{y \in B} B(y, \sigma)$ are disjoint, which contradicts the assumption that $\mathrm{spt}(\mu)$ is connected. Hence $\mathrm{int}(\mathrm{spt}(\mu * \eta_\sigma))$ is connected, as $\Gamma \subset \mathrm{int}(\mathrm{spt}(\mu * \eta_\sigma)) \subset \overline{\Gamma} = \mathrm{spt}(\mu * \eta_\sigma)$. □

Proof of Lemma 56. Set $\mu_t = \mu_0 + t(\mu_1 - \mu_0)$, and let φ_t° be OT potentials from $\mu_t * \eta_\sigma$ to $\nu * \eta_\sigma$ for W_p satisfying $0 \le \varphi_t^\circ \le \mathrm{diam}(\mathcal{X}_\sigma)^p$ on $\mathcal{X}_\sigma$; cf. Remark 1.13 in [207]. For any fixed $x_0 \in \mathrm{int}(\mathrm{spt}(\mu_0 * \eta_\sigma))$, $\varphi_t := \varphi_t^\circ - \varphi_t^\circ(x_0)$ is again an OT potential from $\mu_t * \eta_\sigma$ to $\nu * \eta_\sigma$ for W_p satisfying $\varphi_t(x_0) = 0$ and $|\varphi_t| \le \mathrm{diam}(\mathcal{X}_\sigma)^p$ on $\mathcal{X}_\sigma$.

Observe that, by duality,

$$\begin{aligned}
\mathsf{V}_p^\sigma(\mu_t, \nu) - \mathsf{V}_p^\sigma(\mu_0, \nu) &= [\mathsf{W}_p(\mu_t * \eta_\sigma, \nu * \eta_\sigma)]^p - [\mathsf{W}_p(\mu_t * \eta_\sigma, \nu * \eta_\sigma)]^p \\
&\le \int_{\mathcal{X}_\sigma} \varphi_t d(\mu_t * \eta_\sigma) - \int_{\mathcal{X}_\sigma} \varphi_t d(\mu_0 * \eta_\sigma) \\
&\le t \int_{\mathcal{X}_\sigma} \varphi_t d((\mu_1 - \mu_0) * \eta_\sigma),
\end{aligned}$$

so that

$$\limsup_{t \downarrow 0} \frac{\mathsf{V}_p^\sigma(\mu_t, \nu) - \mathsf{V}_p^\sigma(\mu_0, \nu)}{t} \le \limsup_{t \downarrow 0} \int_{\mathcal{X}_\sigma} \varphi_t d((\mu_1 - \mu_0) * \eta_\sigma).$$

Since $\mathrm{int}(\mathrm{spt}(\mu_0 * \eta_\sigma))$ is connected with negligible boundary by Lemma 57, and $\mu_{t_n} * \eta_\sigma \xrightarrow{w} \mu * \eta_\sigma$ for any sequence $t_n \downarrow 0$, Theorem 3.4 in [61] implies that $\varphi_{t_n} - a_n \to \varphi$ pointwise on $\mathrm{int}(\mathrm{spt}(\mu_0 * \eta_\sigma))$ for some sequence of constants a_n. Since $\varphi_{t_n}(x_0) = \varphi(x_0) = 0$, we must have that $a_n \to 0$, so in fact $\varphi_{t_n} \to \varphi$ pointwise on $\mathrm{int}(\mathrm{spt}(\mu_0 * \eta_\sigma))$. Since $\mathrm{int}(\mathrm{spt}(\mu_0 * \eta_\sigma))$ has negligible boundary by Lemma 57 and $\mathrm{spt}(\mu_1) \subset \mathrm{spt}(\mu_0)$, applying the dominated convergence theorem yields

$$\int_{\mathcal{X}_\sigma} \varphi_{t_n} d((\mu_1 - \mu_0) * \eta_\sigma) \to \int_{\mathcal{X}_\sigma} \varphi_0 d((\mu_1 - \mu_0) * \eta_\sigma).$$

Since $t_n \downarrow 0$ is arbitrary, we have

$$\limsup_{t \downarrow 0} \frac{\mathsf{V}_p^\sigma(\mu_t, \nu) - \mathsf{V}_p^\sigma(\mu_0, \nu)}{t} \le \int \varphi_0 d((\mu_1 - \mu_0) * \eta_\sigma).$$

For the reverse inequality, observe that

$$\begin{aligned}\mathsf{V}_p^\sigma(\mu_t,\nu)-\mathsf{V}_p^\sigma(\mu_0,\nu) &\geq \int_{\mathcal{X}_\sigma}\varphi d(\mu_t*\eta_\sigma)-\int_{\mathcal{X}_\sigma}\varphi_0 d(\mu_0*\eta_\sigma)\\ &= t\int_{\mathcal{X}_\sigma}\varphi_0 d\left((\mu_1-\mu_0)*\eta_\sigma\right),\end{aligned}$$

so that

$$\liminf_{t\downarrow 0}\frac{\mathsf{V}_p^\sigma(\mu_t,\nu)-\mathsf{V}_p^\sigma(\mu_0,\nu)}{t}\geq\int_{\mathcal{X}_\sigma}\varphi_0 d\left((\mu_1-\mu_0)*\eta_\sigma\right).$$

Conclude that $\lim_{t\downarrow 0} t^{-1}(\mathsf{V}_p^\sigma(\mu_t,\nu)-\mathsf{V}_p^\sigma(\mu_0,\nu)) = ((\mu_1-\mu_0)*\eta_\sigma)(\varphi_0)$. The derivative is uniquely defined, as φ_0 is unique up to additive constants on $\mathrm{int}(\mathrm{spt}(\mu*\eta_\sigma))$ and $(\mu_1-\mu_0)*\eta_\sigma$ has total mass zero. □

Proof of Theorem 4. As noted before, we only prove the one-sample result. We shall apply Proposition 1 by identifying $\mathsf{V}_p^\sigma(\rho,\nu)=\mathsf{W}_p^p(\rho*\eta_\sigma,\nu*\eta_\sigma)$ as a functional defined on $\mathcal{P}_0*\eta_\sigma=\{\rho*\eta_\sigma:\rho\in\mathcal{P}_0\}$ with $\mathcal{P}_0=\{\rho\in\mathcal{P}(\mathcal{X}):\mathrm{spt}(\rho)\subset\mathrm{spt}(\mu)\}$, and taking $\mathcal{F}=B$, $F\equiv C$ with constant C given in (A.10), and $\mu_n=\hat{\mu}_n*\eta_\sigma$. The set $\mathcal{P}_0*\eta_\sigma$ is convex and contains $\hat{\mu}_n*\eta_\sigma$ with probability one. Combining Lemmas 54–56, we have verified all conditions in Proposition 1. Pick a version of φ that is bounded on $\mathcal{X}_\sigma$; cf. Remark 1.13 in [208]. Choose a constant $\mathsf{c}>0$ such that $\bar{\varphi}=\varphi/\mathsf{c}\in B$. Observe that the derivative formula in Lemma 56 agrees with $\mathsf{c}\cdot((\mu_1-\mu_0)*\eta_\sigma)(\bar{\varphi})$. Conclude that

$$\sqrt{n}(\mathsf{V}_p^\sigma(\hat{\mu}_n,\nu)-\mathsf{V}_p^\sigma(\mu,\nu))\stackrel{d}{\to}\mathsf{c}\cdot\mathbb{G}_\mu(\bar{\varphi})\sim N(0,\mathrm{Var}_\mu(\varphi*\chi_\sigma)).$$

Pertaining to semiparametric efficiency, think of $\mathsf{V}_p^\sigma(\cdot,\nu)$ now as a functional defined on $\mathcal{P}_0$. It is not difficult to see that, for any bounded measurable function h on $\mathbb{R}^d$ with μ-mean zero, $(1+th)\mu\in\mathcal{P}_0$ for sufficiently small $t>0$. Lemma 56 implies that $t^{-1}(\mathsf{V}_p^\sigma((1+th)\mu,\nu)-\mathsf{V}_p^\sigma(\mu,\nu))\to((h\mu)*\eta_\sigma)(\varphi)=(h\mu)(\varphi*\chi_\sigma)$ as $t\downarrow 0$, which is the point evaluation at $\varphi*\chi_\sigma\in L^2(\mu)$. Hence, by Proposition 1, the semiparametric efficiency bound for estimating $\mathsf{V}_p^\sigma(\cdot,\nu)$ at μ agrees with $\mathrm{Var}_\mu(\varphi*\chi_\sigma)$.

Finally, the bootstrap consistency follows by linearity of the derivative in Lemma 56 and the second claim in Lemma 54, combined with Corollary 1 and Theorem 23.9 in [200]. □

A.3.4 Proof of Proposition 6

Let $Z_i = \delta_{X_i} * \eta_\sigma$ be i.i.d. $(L^2(\mathcal{X}_\sigma))'$-valued random variables as in the proof of Lemma 54. By the proof of Lemma 55,

$$\begin{aligned}\left|\mathsf{V}_p^\sigma(\hat{\mu}_n, \nu) - \mathsf{V}_p^\sigma(\mu, \nu)\right| &\leq \sqrt{\lambda(\mathcal{X}_\sigma)}\,\mathrm{diam}(\mathcal{X}_\sigma)^p \|(\hat{\mu}_n - \mu) * \eta_\sigma\|_{(L^2(\mathcal{X}_\sigma))'},\\ &= \sqrt{\lambda(\mathcal{X}_\sigma)}\,\mathrm{diam}(\mathcal{X}_\sigma)^p \left\| n^{-1} \sum_{i=1}^n (Z_i - \mathbb{E}[Z_i]) \right\|_{(L^2(\mathcal{X}_\sigma))'}.\end{aligned}$$

Since $(L^2(\mathcal{X}_\sigma))'$ is a Hilbert space,

$$\mathbb{E}\left[\left\| n^{-1} \sum_{i=1}^n (Z_i - \mathbb{E}[Z_i]) \right\|^2_{(L^2(\mathcal{X}_\sigma))'}\right] = n^{-1}\mathbb{E}[\|Z_1 - \mathbb{E}[Z_1]\|^2_{(L^2(\mathcal{X}_\sigma))'}] \leq 4n^{-1}C^2,$$

where $C = \|\chi_\sigma\|_\infty \sqrt{\lambda(B(0,\sigma))}$ is a constant given in (A.10). Applying Jensen's inequality in the previous display and combining both bounds yields

$$\mathbb{E}\left[\left|\mathsf{V}_p^\sigma(\hat{\mu}_n, \nu) - \mathsf{V}_p^\sigma(\mu, \nu)\right|\right] \leq 2C\sqrt{\lambda(\mathcal{X}_\sigma)}\,\mathrm{diam}(\mathcal{X}_\sigma)^p n^{-1/2}.$$

Applying the inequality $|x - y| \leq y^{1-p}\,|x^p - y^p|$ yields

$$\mathbb{E}\left[\left|\mathsf{W}_p^\sigma(\hat{\mu}_n, \nu) - \mathsf{W}_p^\sigma(\mu, \nu)\right|\right] \leq 2C\sqrt{\lambda(\mathcal{X}_\sigma)}\,\mathrm{diam}(\mathcal{X}_\sigma)^p [\mathsf{W}_p^\sigma(\mu, \nu)]^{1-p} n^{-1/2},$$

as desired. □

A.3.5 Proof of Theorem 5

In what follows, we fix $\sigma > 0$. We will first show that the smoothed empirical process $\sqrt{n}(\hat{\mu}_n - \mu) * \eta_\sigma$ can be regarded as the scaled sum of i.i.d. random variables with values in $\dot{H}^{-1,2}(\mu * \eta_\sigma)$.

Lemma 58. If $\mu \in \mathcal{P}(\mathcal{X})$ is such that $\mu * \eta_\sigma$ satisfies a 2-Poincaré inequality, then $(\delta_X - \mu) * \eta_\sigma \in \dot{H}^{-1,2}(\mu * \eta_\sigma)$ μ-a.s.

Proof. Let $Y \sim \mu$. Observe that the Lebesgue density of $\mu * \eta_\sigma$ agrees with $\mathbb{E}[\chi_\sigma(Y - \cdot)]$. We shall verify that, for any $x \in \mathrm{int}(\mathrm{spt}(\mu))$, $\mathbb{E}[\chi_\sigma(Y - \cdot)] > 0$ on $B(x, \sigma)$. Indeed, for any $y \in B(x, \sigma)$, there exists $0 < \epsilon_y \leq \sigma$ with $B(x, \epsilon_y) \subset B(y, \sigma)$. Since $\mu(B(x, \epsilon_y)) > 0$ and $\chi_\sigma(\cdot - y) > 0$ on $B(x, \epsilon_y) \subset B(y, \sigma)$, we have $\mathbb{E}[\chi_\sigma(Y - y)] \geq \mathbb{E}[\chi_\sigma(Y - y)\mathbb{1}_{B(x,\epsilon_y)}(Y)] > 0$.

Now, for any $f \in C_0^\infty + \mathbb{R}$ and $x \in \mathrm{spt}(\mu)$, we have

$$\begin{aligned}
|(f * \chi_\sigma)(x)| &\leq \int_{B(x,\sigma)} |f(y)\chi_\sigma(x - y)|\, dy \\
&= \int_{B(x,\sigma)} |f(y)| \frac{\chi_\sigma(x - y)}{\mathbb{E}[\chi_\sigma(Y - y)]} d(\mu * \eta_\sigma)(y) \\
&\leq \|f\|_{L^2(B(x,\sigma),\mu*\eta_\sigma)} \left\| \frac{\chi_\sigma(x - \cdot)}{\mathbb{E}[\chi_\sigma(Y - \cdot)]} \right\|_{L^2(B(x,\sigma),\mu*\eta_\sigma)},
\end{aligned}$$

where $\|\cdot\|_{L^2(B(x,\sigma),\mu*\eta_\sigma)}$ is the $L^2(\mu * \eta_\sigma)$-norm restricted to $B(x, \sigma)$.

Observe that, since $(\delta_x - \mu) * \eta_\sigma$ has total mass zero, for $\bar{f} = f - (\mu * \eta_\sigma)(f)$,

$$\left|((\delta_x - \mu) * \eta_\sigma)f\right| = \left|(\bar{f} * \chi_\sigma)(x)\right| \leq \|\bar{f}\|_{L^2(B(x,\sigma),\mu*\eta_\sigma)} \left\| \frac{\chi_\sigma(x - \cdot)}{\mathbb{E}[\chi_\sigma(Y - \cdot)]} \right\|_{L^2(B(x,\sigma),\mu*\eta_\sigma)}.$$

Since $\mu * \eta_\sigma$ satisfies the 2-Poincaré inequality and $B(x, \sigma) \subset \mathrm{spt}(\mu * \eta_\sigma)$,

$$\|\bar{f}\|_{L^2(B(x,\sigma),\mu*\eta_\sigma)} \leq \|\bar{f}\|_{L^2(\mu*\eta_\sigma)} \leq C_{\mu,\sigma}\|f\|_{\dot{H}^{1,2}(\mu*\eta_\sigma)}$$

for some constant $C_{\mu,\sigma} > 0$ depending only on μ and σ.

As for the other term, let $\Gamma = \{y : \mathbb{E}[\chi_\sigma(Y - y)] > 0\}$, and observe that, for $X \sim \mu$

independent of Y,

$$\begin{aligned}\int_\Gamma \frac{\mathbb{E}[\chi_\sigma^2(X-y)]}{\mathbb{E}[\chi_\sigma(X-y)]}dy &= \int_\Gamma \frac{\mathbb{E}[\chi_\sigma^2(X-y)\left(\mathbb{1}_{\{\chi_\sigma(X-\cdot)<1\}}(y)+\mathbb{1}_{\{\chi_\sigma(X-\cdot)\geq 1\}}(y)\right)]}{\mathbb{E}[\chi_\sigma(X-y)\left(\mathbb{1}_{\{\chi_\sigma(Y-\cdot)<1\}}(y)+\mathbb{1}_{\{\chi_\sigma(Y-\cdot)\geq 1\}}(y)\right)]}dy \\ &\leq \int_\Gamma \frac{\mathbb{E}[\chi_\sigma(X-y)\mathbb{1}_{\{\chi_\sigma(X-\cdot)<1\}}(y)]+\|\chi_\sigma^2\|_\infty \mu(\{\chi_\sigma(\cdot-y)\geq 1\})}{\mathbb{E}[\chi_\sigma(Y-y)\mathbb{1}_{\{\chi_\sigma(Y-\cdot)<1\}}(y)]+\mu(\{\chi_\sigma(\cdot-y)\geq 1\})}dy \\ &\leq (1\vee\|\chi_\sigma^2\|_\infty)\lambda(\Gamma).\end{aligned}$$

Applying the Fubini theorem and Jensen's inequality yields

$$\begin{aligned}\int_\Gamma \frac{\mathbb{E}[\chi_\sigma^2(X-y)]}{\mathbb{E}[\chi_\sigma(Y-y)]}dy &= \mathbb{E}\left[\left\|\frac{\chi_\sigma(X-\cdot)}{\mathbb{E}[\chi_\sigma(Y-\cdot)]}\right\|^2_{L^2(\Gamma,\mu*\eta_\sigma)}\right] \\ &\geq \left(\mathbb{E}\left[\left\|\frac{\chi_\sigma(X-\cdot)}{\mathbb{E}[\chi_\sigma(Y-\cdot)]}\right\|_{L^2(\Gamma,\mu*\eta_\sigma)}\right]\right)^2 \\ &\geq \left(\mathbb{E}\left[\left\|\frac{\chi_\sigma(X-\cdot)}{\mathbb{E}[\chi_\sigma(Y-\cdot)]}\right\|_{L^2(B(x,\sigma),\mu*\eta_\sigma)}\right]\right)^2,\end{aligned}$$

where the last inequality follows as $B(x,\sigma)\subset\Gamma$ for every $x\in\mathrm{spt}(\mu)$. Conclude that

$$\mathbb{E}\left[\left\|\frac{\chi_\sigma(X-\cdot)}{\mathbb{E}[\chi_\sigma(Y-\cdot)]}\right\|_{L^2(B(x,\sigma),\mu*\eta_\sigma)}\right]\leq\sqrt{(1\vee\|\chi_\sigma^2\|_\infty)\lambda(\mathcal{X}_\sigma)}.$$

We have shown that

$$\|(\delta_X-\mu)*\eta_\sigma\|_{\dot{H}^{-1,2}(\mu*\eta_\sigma)}\leq C_{\mu,\sigma}\left\|\frac{\chi_\sigma(X-\cdot)}{\mathbb{E}[\chi_\sigma(Y-\cdot)]}\right\|_{L^2(B(x,\sigma),\mu*\eta_\sigma)},$$

and the right-hand side is μ-a.s. finite as its expectation is finite, so $(\delta_X-\mu)*\eta_\sigma\in\dot{H}^{-1,2}(\mu*\eta_\sigma)$ μ-a.s. □

Lemma 59. For $\mu\in\mathcal{P}(\mathcal{X})$ such that $\mu*\eta_\sigma$ satisfies a 2-Poincaré inequality, $\sqrt{n}(\hat{\mu}_n-\mu)*\eta_\sigma\xrightarrow{d}\mathbb{G}_\mu$ in $\dot{H}^{-1,2}(\mu*\eta_\sigma)$, where $\mathbb{G}_\mu=\left(\mathbb{G}_\mu(f)\right)_{f\in\dot{H}^{1,2}(\mu*\eta_\sigma)}$ is a centered Gaussian process with paths in $\dot{H}^{-1,2}(\mu*\eta_\sigma)$.

Proof. By Lemma 58, $Z_i:=(\delta_{X_i}-\mu)*\eta_\sigma$ are i.i.d. mean-zero $\dot{H}^{-1,2}(\mu*\eta_\sigma)$-valued random variables. By Lemma 5.1 in [105], $\dot{H}^{-1,2}(\mu*\eta_\sigma)$ is isometrically isomorphic to a closed subspace of $L^2(\mu * \eta_\sigma; \mathbb{R}^d)$ and hence a separable Hilbert space. Further, from the proof of Lemma 58,

$$\mathbb{E}\left[\|Z_i\|^2_{\dot{H}^{-1,2}(\mu*\eta_\sigma)}\right] \le C^2_{\mu,\sigma}(1 \vee \|\chi^2_\sigma\|_\infty)\lambda(\mathcal{X}_\sigma) < \infty.$$

Conclude from Example 1.8.5 in [201] that $\sqrt{n}(\hat{\mu}_n - \mu) * \eta_\sigma = n^{-1/2}\sum_{i=1}^n Z_i$ satisfies a CLT in $\dot{H}^{-1,2}(\mu * \eta_\sigma)$. $\square$

Lemma 60. For $\mu \in \mathcal{P}(\mathcal{X})$, the map

$$\Phi : (h_1, h_2) \in \Xi_\mu \times \Xi_\mu \subset \dot{H}^{-1,2}(\mu * \eta_\sigma) \times \dot{H}^{-1,2}(\mu * \eta_\sigma) \mapsto \mathsf{W}_2(\mu * \eta_\sigma + h_1, \mu * \eta_\sigma + h_2),$$

is Hadamard directionally differentiable at $(0,0)$ with $\Phi'_{(0,0)}(h_1, h_2) = \|h_1 - h_2\|_{\dot{H}^{-1,2}(\mu*\eta_\sigma)}$, where

$$\Xi_\mu = \dot{H}^{-1,2}(\mu * \eta_\sigma) \cap \{h = (\rho - \mu) * \eta_\sigma : \rho \in \mathcal{P}(\mathcal{X}), \mathrm{spt}(\rho) \subset \mathrm{spt}(\mu)\}.$$

Remark 26. Any finite Borel signed measure on $\mathbb{R}^d$ with total mass 0 and finite $\|\cdot\|_{H^{-1,2}(\mu*\eta_\sigma)}$ norm corresponds to an element of $H^{-1,2}(\mu * \eta_\sigma)$, and this correspondence is one-to-one (as C_0^∞ is right enough). Thus, the above map Φ is well-defined.

The proof of Lemma 60 follows from that of Proposition 3.3 in [105] with only minor modifications and thus is omitted for brevity.

Proof of Theorem 5. Let $(T_{n,1}, T_{n,2}) := ((\hat{\mu}_n - \mu) * \eta_\sigma, (\hat{\nu}_n - \nu) * \eta_\sigma)$. Then, by independence of $T_{n,1}$ and $T_{n,2}$, and Lemma 59,

$$\sqrt{n}(T_{n,1}, T_{n,2}) \xrightarrow{d} (\mathbb{G}_\mu, \mathbb{G}'_\mu) \text{ in } \dot{H}^{-1,2}(\mu * \eta_\sigma) \times \dot{H}^{-1,2}(\mu * \eta_\sigma).$$

Observe that $\Xi_\mu \times \Xi_\mu$ in Lemma 60 is convex, so that the tangent cone $\mathcal{T}_{\Xi_\mu \times \Xi_\mu}(0,0)$ agrees with the closure of $\left\{(h_1, h_2)/t : (h_1, h_2) \in \Xi_\mu \times \Xi_\mu, t > 0\right\}$ in $\dot{H}^{-1,2}(\mu * \eta_\sigma) \times \dot{H}^{-1,2}(\mu * \eta_\sigma)$. Thus, $\sqrt{n}(T_{n,1}, T_{n,2}) \in \mathcal{T}_{\Xi_\mu \times \Xi_\mu}(0,0)$ and $(\mathbb{G}_\mu, \mathbb{G}'_\mu) \in \mathcal{T}_{\Xi_\mu \times \Xi_\mu}(0,0)$ by the portmanteau theorem.

Now, apply the extended functional delta method, Lemma 1, to conclude that

$$\begin{aligned}\sqrt{n}\mathsf{W}_p^{\sigma}(\hat{\mu}_n, \hat{\nu}_n) &= \sqrt{n}\left(\Phi\left(T_{n,1}, T_{n,2}\right) - \Phi\left(0,0\right)\right)\\ &\stackrel{d}{\to} \Phi'_{(0,0)}(\mathbb{G}_\mu, \mathbb{G}'_\mu) = \|\mathbb{G}_\mu - \mathbb{G}'_\mu\|_{\dot{H}^{-1,2}(\mu*\eta_\sigma)}.\end{aligned}$$

The one-sample result is obtained analogously. □

A.3.6 Proof of Proposition 7

Let $Z_i = (\delta_{X_i} - \mu) * \eta_\sigma$ be i.i.d. $\dot{H}^{-1,2}(\mu * \eta_\sigma)$-valued random variables appearing in the proof of Lemma 54. By Proposition 2.1 of [105],

$$\mathsf{W}_2^{\sigma}(\hat{\mu}_n, \mu) \leq 2\|(\hat{\mu}_n - \mu) * \eta_\sigma\|_{\dot{H}^{-1,2}(\mu*\eta_\sigma)} = 2\left\|n^{-1}\sum_{i=1}^n Z_i\right\|_{\dot{H}^{-1,2}(\mu*\eta_\sigma)}.$$

Since $\dot{H}^{-1,2}(\mu * \eta_\sigma)$ is a separable Hilbert space, the expectation on the right-hand side is bounded as

$$2n^{-1/2}\sqrt{\mathbb{E}\left[\|Z_1\|^2_{\dot{H}^{-1,2}(\mu*\eta_\sigma)}\right]}.$$

From the proof of Lemma 58,

$$\mathbb{E}\left[\|Z_1\|^2_{\dot{H}^{-1,2}(\mu*\eta_\sigma)}\right] \leq C^2_{\mu,\sigma}\mathbb{E}\left[\left\|\frac{\chi_\sigma(X-\cdot)}{\mathbb{E}[\chi_\sigma(Y-\cdot)]}\right\|^2_{L^2(\mu*\eta_\sigma,\Gamma)}\right] \leq C^2_{\mu,\sigma}(1 \vee \|\chi^2_\sigma\|_\infty)\lambda(\mathcal{X}_\sigma).$$

This completes the proof. □

A.3.7 Proofs of Theorem 6 and Corollary 4

Recall $\mathcal{F}_\sigma = \{f * \chi_\sigma : f \in \mathsf{Lip}_{1,0}\}$.

Lemma 61. The function class $\mathcal{F}_\sigma$ is μ-Donsker for any $\mu \in \mathcal{P}(\mathcal{X})$.

Proof. Pick any $f \in \mathsf{Lip}_{1,0}$. Since $|f(y)| = |f(y) - f(0)| \le \|y\|$, $\|y\| \le \|x\| + \sigma$ for $y \in B(x, \sigma)$, and χ_σ is supported on $B(0, \sigma)$, we have

$$|(f * \chi_\sigma)(x)| \le \int_{B(x,\sigma)} \|y\| \chi_\sigma(x-y)dy \le \int_{B(x,\sigma)} (\|x\| + \sigma)\chi_\sigma(x-y)dy = \|x\| + \sigma.$$

Next, we study the derivatives. For any multi-index $k = (k_1, \ldots, k_d) \in \mathbb{N}_0^d$, let $\partial^k = \partial_1^{k_1} \cdots \partial_d^{k_d}$ denote the differential operator and set $\bar{k} := \sum_{j=1}^d k_j$. Observe that, for any $x, x' \in \mathbb{R}^d$,

$$\begin{aligned}\left|\partial^k(f * \chi_\sigma)(x) - \partial^k(f * \chi_\sigma)(x')\right| &\le \int_{B(0,\sigma)} |f(x-y) - f(x'-y)| \left|\partial^k \chi_\sigma(y)\right| dy \\ &\le \|\partial^k \chi_\sigma\|_\infty \lambda(B(0,\sigma))\|x - x'\|.\end{aligned}$$

Since χ_σ is smooth and compactly supported, for any $s \in \mathbb{N}$, there exists a constant C_s independent of f such that $\|\partial^k(f * \chi_\sigma)\|_\infty \le C_s$ for any multi-index with $1 \le \bar{k} \le s$.

Pick sufficiently large $R > 0$ such that $\mathcal{X} \subset B(0, R)$. Let $\mathcal{F}_\sigma|_{B(0,R)} = \{f|_{B(0,R)} : f \in \mathcal{F}_\sigma\}$. Then, by Theorem 2.7.1 in [201], we have

$$\log N(\mathcal{F}_\sigma|_{B(0,R)}, \|\cdot\|_\infty, \epsilon) = O(\epsilon^{-d/s}), \ \epsilon \downarrow 0 \tag{A.11}$$

for any $s \in \mathbb{N}$. Choosing $s = \lfloor d/2 \rfloor + 1 > d/2$, we see that the function class $\mathcal{F}_\sigma$ is μ-Donsker from Theorem 2.5.6 in [201]. □

Proof of Theorem 6. Recall that $\mathsf{W}_1^\sigma(\mu, \nu) = \sup_{f \in \mathcal{F}_\sigma}(\mu - \nu)(f)$ by the Kantorovich-Rubinstein duality. Since $\mathcal{F}_\sigma$ is Donsker w.r.t. μ and ν, the conclusion of the theorem follows from a similar argument to the proof of Theorem 3 Parts (iii) and (iv). We omit the details for brevity. □

Proof of Corollary 4. Recall the expression $\mathsf{W}_1^\sigma(\hat{\mu}_n, \mu) = \sup_{f \in \mathcal{F}_\sigma}(\hat{\mu}_n - \mu)(f)$. As μ is compactly supported, the desired result follows by the bracketing entropy bound (A.11) combined with Theorem 2.14.2 in [201]. □

A.4 Computational Aspects of Smooth Wasserstein Distances

We take here a first step towards addressing computational aspects for smooth Wasserstein distances, leveraging the framework of [197] for computing the W_2 distance between measures with smooth compactly supported densities that are bounded from above and below. We first outline the approach of [197], then present a truncation argument that enables applying it for W_2^σ, and lastly describe some practical limitations of the original method.

A.4.1 Summary of the method

Throughout, let $c(x, y) = \|x - y\|^2/2$ and $\mu_1, \mu_2 \in \mathcal{P}(\mathbb{R}^d)$ be such that $A_i := \mathrm{int}(\mathrm{spt}(\mu_i))$ is bounded and convex for $i = 1, 2$. Assume in addition that μ_i has (Lebesgue) density f_i which is bounded away from zero and infinity on A_i with Lipschitz derivatives up to order $m > d$ and that A_i is smooth and uniformly convex (see e.g. [95, p.339]) for $i = 1, 2$.[2] These technical conditions guarantee that the OT potentials $\varphi_1^\star, \varphi_2^\star$ solving the alternate dual,

$$\frac{1}{2}\mathsf{W}_2^2(\mu_1, \mu_2) = \sup_{\substack{(\varphi_1,\varphi_2)\in L^1(\mu_1)\times L^2(\mu_2)\\ \varphi_1\oplus\varphi_2\leq c}} \left[\int_{A_1} \varphi_1 d\mu_1 + \int_{A_2} \varphi_2 d\mu_2\right], \tag{A.12}$$

are elements of certain Sobolev spaces.

The first insight of [197] is to exploit the reproducing kernel Hilbert spaces (RKHS) of such Sobolev spaces (cf. [212], [197, Proposition 4]) so as to facilate computations via the kernel trick [178] and a representer theorem [197, Lemma 13]. Next, (A.12) is replaced by an equivalent equality-constrained problem [197, Equation 2]. This is

[2]Assumption 1 in [197] omits smoothness and uniform convexity of A_i. However, in the absence of these conditions it is unclear if their claimed boundary regularity of OT maps and potentials hold (cf. Theorem 3.3 in [58]).

motivated by the observation that equality constraints can be subsampled (i.e., enforced only on a set Z_ℓ of cardinality ℓ) at the cost of incurring an error of $O(h_\ell^{m+1-d})$ provided $h_\ell \lesssim m^{-2}$, where $h_\ell := \sup_{x \in A_1, y \in A_2} \inf_{(\tilde{x}_\ell, \tilde{y}_\ell) \in Z_\ell} \|(x, y) - (\tilde{x}_\ell, \tilde{y}_\ell)\|$ [149, Theorem 2.12].

Finally, the objective of the subsampled and equality-constrained version of (A.12) is regularized to guarantee the uniqueness of its solution. The dual of this problem is a tractable ℓ-dimensional optimization problem whose solution can be used to compute the estimate $\widetilde{\mathsf{W}}_2^2(\mu_1, \mu_2)$ of $\mathsf{W}_2^2(\mu_1, \mu_2)$. The error of that estimate is $O(\lambda_1 + \lambda_2 + (\gamma + h_\ell^{m+1-d})^2/\lambda_2)$, provided that $h_\ell \lesssim m^{-2}$ and $h_\ell^{m+1-d} \lesssim \lambda_1$. Here (λ_1, λ_2) are positive hyperparameters introduced in the regularization procedure, and γ is related to the error of approximate integration (see Theorem 9 in [197]). A variant of Newton's method is shown to solve the dual problem within a tolerance of ϵ with time complexity $O(C + E\ell + \ell^{3.5} \log(\ell/\epsilon))$ and memory complexity $O(\ell^2)$, where C and E are costs of evaluating certain functions which may depend on ℓ [197, Equation 7].

The followup work [196] provides a variant of this method tailored for approximate computation of OT maps for the cost $c(x, y) = \langle x, y \rangle$ given samples from μ and ν. The improvement of this approach upon the method of [197] is twofold: it provides a heuristic by which to choose the hyperparameters and it proposes an alternative regularization method so as to leverage optimization methods which scale better with ℓ than the Newton method.

Remark 27 (Smoothness versus subsampling)**.** The value of h_ℓ^{m+1-d} plays an important role in the proposed method. While it may appear negligible provided m is sufficiently large and $h_\ell < 1$, the theoretical guarantees of [197] are contingent on assuming $h_\ell \lesssim m^{-2}$. Indeed, this condition enables controlling the amount by which solutions of the subsampled equality constraint violate the original constraint [149, Theorem 2.12]. Hence, utilizing higher regularity of densities comes at the cost of increasing the number of subsamples which in turn implies a greater computational cost in terms of time and memory.

A.4.2 Truncation of measures and application of the algorithm

We first observe that the above described method is not directly applicable to W_2^σ, since the density of $\mu * \eta_\sigma$ decays to zero on $\mathrm{spt}(\mu * \eta_\sigma)$, which violates the required lower-boundedness. To rectify this issue, we truncate the measures of interest and quantify the OT gap between the original measure and its truncated version.

Let $\mu \in \mathcal{P}(\mathbb{R}^d)$ be arbitrary and $A \subset \mathbb{R}^d$ be Borel measurable with $\mu(A) > 0$. The truncation of μ to A, denoted $\mu|_A$, is the conditional probability measure $\mu(\cdot|A)$. The next result quantifies the error (in W_p) of approximating a measure by its truncation.

Proposition 13 (Truncation error)**.** Let $\mu \in \mathcal{P}_{p+1}(\mathbb{R}^d)$, and $A \subset \mathbb{R}^d$ be a bounded Borel set with $\mu(A) > 0$. The following hold.

(i) If $1 < p < \infty$ and μ has a (c_1, c_2)-regular density f_μ [3], then

$$\mathsf{W}_p^p(\mu|_A, \mu) \leq C\left(\mathbb{E}_\mu[\|X\|] + \mathbb{E}_\mu[\|X\|^{p+1}] + (1-2\mu(A))(\mathbb{E}_{\mu|_A}[\|Y\|] + \mathbb{E}_{\mu|_A}[\|Y\|^{p+1}])\right),$$

where C depends only on $p, d, c_1, c_2, \mathrm{diam}(A)$, and a lower bound on $f_\mu(0)$.

(ii) If $1 \leq p < \infty$ and μ has compact support, then

$$\mathsf{W}_p^p(\mu|_A, \mu) \leq \left(\frac{1}{\mu(A)} - 1\right) \mathrm{diam}\,(\mathrm{spt}(\mu))^p.$$

These bounds exhibit the expected behaviour in the sense that the larger $\mu(A)$ is, the closer $\mu|_A$ is to μ under W_p, and if $\mu(A) = 1$ the error is zero. Both results follow

[3] f_μ is a (c_1, c_2)-regular density if $\|\nabla \log f_\mu\| \leq c_1\|\cdot\| + c_2$ for $c_1 > 0$ and $c_2 \geq 0$.

from estimates for OT potentials between probability measures adhering to the above assumptions; see Section A.4.3.

Proposition 13 implies that, at the cost of introducing an approximation error, kernel smoothed measures can be truncated so as to match the assumptions of Section A.4.1. The following result summarizes the overall accuracy and complexity guarantees obtained by combining the algorithm from [197] with Proposition 13.

Proposition 14 (Computational guarantees)**.** Let $\mu_1, \mu_2 \in \mathcal{P}(\mathbb{R}^d)$ be compactly supported, and $B_1, B_2 \subset \mathbb{R}^d$ be closed sets with $\mu_i * \eta_\sigma(B_i) > 0$, such that $A_i := \mathrm{int}(B_i)$ is smooth and uniformly convex for $i = 1, 2$. Further, assume that the density of $\mu_i * \eta_\sigma$ is bounded away from 0 on B_i for $i = 1, 2$. Then, $\widetilde{\mathsf{W}}_2\left(\mu_1 * \eta_\sigma|_{A_1}, \mu_2 * \eta_\sigma|_{A_2}\right)$ described in Section A.4.1 can be computed up to accuracy ϵ in $O(C + E\ell + \ell^{3.5}\log(\ell/\epsilon))$ time using $O(\ell^2)$ memory and, for any choice of $m \in \mathbb{N}$, this estimate approximates $\mathsf{W}_2(\mu_1 * \eta_\sigma, \mu_2 * \eta_\sigma)$ within error

$$O\left((\mu_1(A_1)^{-1} - 1)^{1/2} + (\mu_2(A_2)^{-1} - 1)^{1/2} + \left(\lambda_1 + \lambda_2 + (\gamma + h_\ell^{m+1-d})^2/\lambda_2\right)^{1/2}\right)$$

provided $h_\ell^{m+1-d} \lesssim \lambda_1, 0 < \lambda_2$, and $h_\ell \lesssim m^{-2}$.

Proposition 14 can be adapted to the case where $\mu_1 * \eta_\sigma$, $\mu_2 * \eta_\sigma$, or both have regular densities by replacing the truncation error for the compactly supported case with the one from Proposition 13(i). For instance, if the densities of μ_1 and μ_2 are regular, then so are those of the convolved measures (cf. Proposition 3 in [162]).

Remark 28 (Truncation to smooth uniformly convex sets)**.** In Proposition 14, if $\mathrm{int}\left(\mathrm{spt}(\mu_i * \eta_\sigma)\right)$ is convex, then it can be approximated from the inside by a sequence of open smooth uniformly convex sets converging to $\mathrm{int}\left(\mathrm{spt}(\mu_i * \eta_\sigma)\right)$ in the Hausdorff distance [128]. In particular, if $\mathrm{spt}(\mu_i)$ is convex, then so is $\mathrm{int}\left(\mathrm{spt}(\mu_i * \eta_\sigma)\right)$.

A.4.3 Proof of Proposition 13

The following result is a minor adaptation of Lemma 5.3 from [105] using the fact that distributions with regular densities are equivalent to the Lebesgue measure [162, Section 2].

Lemma 62. Let $1 < p < \infty$. Assume that ν is β-sub-Weibull for $\beta \in (0, 2]$[4], and that $\mu \in \mathcal{P}_p(\mathbb{R}^d)$ has a (c_1, c_2)-regular density f_μ. Let φ be an optimal transport potential from μ to ν for W_p. Then there exists a constant C depending only on p, d, β, c_1, c_2, an upper bound on $\||Y\||_{\psi_\beta}$ for $Y \sim \nu$, and a lower bound on $f_\mu(0)$, such that

$$|\varphi(x) - \varphi(0)| \le C\big(1 + \|x\|^{2p/\beta}\big)\|x\|, \quad \forall x \in \mathbb{R}^d. \tag{A.13}$$

Proof of Proposition 13. **Part (i).** We note that Lemma 62 is applicable in the transport from μ to $\mu|_A$, as μ is (c_1, c_2)-regular and $\mu|_A$ is compactly supported.

Let φ be an OT potential from μ to $\mu|_A$ for W_p with $\varphi(0) = 0$ which satisfies (A.13). Then,

$$\begin{aligned}
\mathsf{W}_p^p(\mu, \mu|_A) &= \int_{\mathbb{R}^d} \varphi d\mu + \int_{\mathbb{R}^d} \varphi^c d\mu|_A \\
&= \int_{\mathbb{R}^d} \varphi d\mu - \int_{\mathbb{R}^d} \varphi d\mu|_A + \int_{\mathbb{R}^d} \varphi d\mu|_A + \int_{\mathbb{R}^d} \varphi^c d\mu|_A \\
&\le \int_{\mathbb{R}^d} \varphi d\mu - \int_{\mathbb{R}^d} \varphi d\mu|_A,
\end{aligned}$$

where the last inequality is because $\int_{\mathbb{R}^d} \varphi d\mu|_A + \int_{\mathbb{R}^d} \varphi^c d\mu|_A \le \mathsf{W}_p^p(\mu|_A, \mu|_A) = 0$. It remains

[4] For $\beta \in (0, 2]$, $\nu \in \mathcal{P}(\mathbb{R}^d)$ is *β-sub-Weibull* if $\||Y\||_{\psi_\beta} := \inf\{C > 0 : \mathbb{E}[\psi_\beta(|Y|/C)] \le 1\} < \infty$ for $Y \sim \nu$, where $\psi_\beta(t) = e^{t^\beta} - 1$ for $t \ge 0$.

to analyze the last term of the above display,

$$
\begin{aligned}
&\int_{\mathbb{R}^d} \varphi d\mu - \int_{\mathbb{R}^d} \varphi d\mu|_A = \int_{A^c} \varphi d\mu + \left(1 - \frac{1}{\mu(A)}\right) \int_A \varphi d\mu \\
&\quad \le \int_{A^c} C(\|x\| + \|x\|^{p+1}) d\mu(x) + \left(\frac{1}{\mu(A)} - 1\right) \int_A C(\|x\| + \|x\|^{p+1}) d\mu(x) \\
&\quad = \int_{\mathbb{R}^d} C(\|x\| + \|x\|^{p+1}) d\mu(x) + \left(\frac{1}{\mu(A)} - 2\right) \int_A C(\|x\| + \|x\|^{p+1}) d\mu(x) \\
&\quad = C\left(\mathbb{E}_\mu[\|X\| + \mathbb{E}_\mu[\|X\|^{p+1}] + (1 - 2\mu(A))(\mathbb{E}_{\mu|_A}[\|Y\|] + \mathbb{E}_{\mu|_A}[\|Y\|^{p+1}])\right).
\end{aligned}
$$

Part (ii). Let φ be an OT potential from $\mu|_A$ to μ for W_p with $0 \le \varphi \le \mathrm{diam}(\mathrm{spt}(\mu))^p$; cf. Remark 1.13 of [207]. Then,

$$
\begin{aligned}
\mathsf{W}_p^p(\mu|_A, \mu) &= \int_{\mathcal{X}} \varphi d\mu|_A + \int_{\mathcal{X}} \varphi^c d\mu \\
&= \int_{\mathcal{X}} \varphi d\mu|_A - \int_{\mathcal{X}} \varphi d\mu + \int_{\mathbb{R}^d} \varphi d\mu + \int_{\mathbb{R}^d} \varphi^c d\mu \\
&\le \int_{\mathbb{R}^d} \varphi d\mu|_A - \int_{\mathbb{R}^d} \varphi d\mu,
\end{aligned}
$$

where the last inequality is because $\int_{\mathbb{R}^d} \varphi d\mu + \int_{\mathbb{R}^d} \varphi^c d\mu \le \mathsf{W}_p^p(\mu, \mu) = 0$. Since φ is nonnegative,

$$
\int_{\mathcal{X}} \varphi d\mu|_A - \int_{\mathcal{X}} \varphi d\mu \le \left(\frac{1}{\mu(A)} - 1\right) \int_{\mathcal{X}} \varphi d\mu \le \left(\frac{1}{\mu(A)} - 1\right) \mathrm{diam}(\mathrm{spt}(\mu))^p.
$$

This completes the proof. □

A.4.4 Proof of Proposition 14

The result follows by noting that

$$
\left| \mathsf{W}_2^\sigma(\mu_1, \mu_1) - \widetilde{\mathsf{W}}_2(\mu_1 * \eta_\sigma|_{A_1}, \mu_2 * \eta_\sigma|_{A_2}) \right| \le E_1 + E_2,
$$

where

$$
\begin{aligned}
E_1 &= \left|\mathsf{W}_2^\sigma(\mu_1,\mu_2) - \mathsf{W}_2(\mu_1 * \eta_\sigma|_{A_1}, \mu_2 * \eta_\sigma|_{A_2})\right|, \\
&\le (\mu_1(A_1)^{-1} - 1)^{1/2}\operatorname{diam}(\operatorname{spt}(\mu_1)) + (\mu_2(A_2)^{-1} - 1)^{1/2}\operatorname{diam}(\operatorname{spt}(\mu_2)),
\end{aligned}
$$

where the upper bound is due to Proposition 13(ii), and

$$
\begin{aligned}
E_2 &= \left|\mathsf{W}_2(\mu_1 * \eta_\sigma|_{A_1}, \mu_2 * \eta_\sigma|_{A_2}) - \widetilde{\mathsf{W}}_2(\mu_1 * \eta_\sigma|_{A_1}, \mu_2 * \eta_\sigma|_{A_2})\right|, \\
&\le \left|\mathsf{W}_2^2(\mu_1 * \eta_\sigma|_{A_1}, \mu_2 * \eta_\sigma|_{A_2}) - \widetilde{\mathsf{W}}_2^2(\mu_1 * \eta_\sigma|_{A_1}, \mu_2 * \eta_\sigma|_{A_2})\right|^{1/2}.
\end{aligned}
$$

Conclude by applying the error of approximating W_2^2 by $\widetilde{\mathsf{W}}_2^2$ from Section A.4.1. □

A.5 Proof for Section 2.6

A.5.1 Proof of Theorem 7

We will mostly focus on the one-sample case and prove the two-sample case as an extension of the proof for the one-sample case. We start from recalling certain regularity properties of optimal EOT potentials between sub-Gaussian distributions. Recall that, for a multi-index $k = (k_1, \dots, k_d) \in \mathbb{N}_0^d$, $\partial^k = \partial_1^{k_1} \cdots \partial_d^{k_d}$ denotes the differential operator and $\bar{k} := \sum_{j=1}^d k_j$. We say that μ is σ^2-sub-Gaussian for $\sigma > 0$ if $\mathbb{E}_\mu[e^{\|X\|^2/(2d\sigma^2)}] \le 2$. Observe that if μ is σ^2-sub-Gaussian, then μ is $\bar{\sigma}^2$-sub-Gaussian for any $\bar{\sigma} > \sigma$. Also μ is sub-Gaussian (i.e., $\|\|X\|\|_{\psi_2} < \infty$ for $X \sim \mu$) if and only if μ is σ^2-sub-Gaussian for some $\sigma > 0$. The following directly follows from Proposition 1 in [142] and its proof.

Lemma 63 (Regularity of optimal EOT potentials [142]). Let $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$ be σ^2-sub-Gaussian. Then, there exist a version of optimal EOT potentials (φ, ψ) for (μ, ν) such that the following hold: for any positive integer $s \ge 2$, there exists a constant $C_{s,d,\sigma} < \infty$ that depends only on s, d, σ such that

$$\begin{aligned}
&\varphi(x) \le \frac{1}{2}(\|x\| + \sqrt{2d}\sigma)^2, \quad \forall x \in \mathbb{R}^d, \\
&|\partial^k \varphi(x)| \le C_{s,d,\sigma}(1 + \|x\|^s), \quad \forall x \in \mathbb{R}^d,\ \forall k \in \mathbb{N}_0^k \text{ with } \bar{k} \le s, \\
&\psi(y) \le \frac{1}{2}(\|y\| + \sqrt{2d}\sigma)^2, \quad \forall y \in \mathbb{R}^d, \\
&|\partial^k \psi(y)| \le C_{s,d,\sigma}(1 + \|y\|^s), \quad \forall y \in \mathbb{R}^d,\ \forall k \in \mathbb{N}_0^k \text{ with } \bar{k} \le s.
\end{aligned} \tag{A.14}$$

Further, the optimality condition (2.13) holds for every $(x, y) \in \mathbb{R}^d \times \mathbb{R}^d$.

In what follows, for sub-Gaussian distributions, we always choose versions of optimal EOT potentials that satisfy (2.13) for every $(x, y) \in \mathbb{R}^d \times \mathbb{R}^d$; with this convention, optimal EOT potentials are unique up to additive constants.

With the estimates (A.14) for optimal EOT potentials in mind, let

$$\mathcal{F}_\sigma = \{f \in C^s(\mathbb{R}^d) : |\partial^k f(x)| \le C_{s,d,\sigma}(1 + \|x\|^s), \forall x \in \mathbb{R}^d, \forall k \in \mathbb{N}_0^k \text{ with } \bar{k} \le s\} \tag{A.15}$$

for $s = \max\{\lfloor d/2 \rfloor + 1, 2\}$. We may assume without loss of generality that $C_{s,d,\sigma}$ is increasing in σ, so that $\mathcal{F}_\sigma \subset \mathcal{F}_{\bar{\sigma}}$ for any $\bar{\sigma} > \sigma$. We first establish Lipschitz continuity of the EOT objective w.r.t. $\|\cdot\|_{\infty,\mathcal{F}_\sigma}$ and characterize its (directional) derivative.

Lemma 64. Let $\mu_0, \mu_1, \nu \in \mathcal{P}(\mathbb{R}^d)$ be σ^2-sub-Gaussian. Then,

$$|\mathsf{S}(\mu_1, \nu) - \mathsf{S}(\mu_0, \nu)| \le \|\mu_0 - \mu_1\|_{\infty,\mathcal{F}_\sigma}. \tag{A.16}$$

Furthermore, we have

$$\lim_{t \downarrow 0} \frac{\mathsf{S}(\mu_0 + t(\mu_1 - \mu_0), \nu) - \mathsf{S}(\mu_0, \nu)}{t} = \int \varphi_0 d(\mu_1 - \mu_0) = (\mu_1 - \mu_0)(\varphi_0), \tag{A.17}$$

where (φ_0, ψ_0) are optimal EOT potentials for (μ_0, ν).

Proof. We first prove (A.17). Let $\mu_t = \mu_0 + t(\mu_1 - \mu_0) = (1 - t)\mu_0 + t\mu_1$ for $t \in [0, 1]$ and let (φ_t, ψ_t) be optimal EOT potentials for (μ_t, ν). Since μ_t is σ^2-sub-Gaussian, we may choose (φ_t, ψ_t) in such a way that they satisfy (2.13) for every $(x, y) \in \mathbb{R}^d \times \mathbb{R}^d$ with (μ, ν) replaced with (μ_t, ν) and (A.14) for $s = 2$. Observe that

$$\begin{aligned}
\mathsf{S}(\mu_t, \nu) &= \int \varphi_t d\mu_t + \int \psi_t d\nu \geq \int \varphi_0 d\mu_t + \int \psi_0 d\nu - \int e^{\varphi_0 \oplus \psi_0 - c} d\mu_t \otimes \nu + 1 \\
&= \int \varphi_0 d\mu_t + \int \psi_0 d\nu = \int \psi_0 d\mu_0 + \int \psi_0 d\nu + t \int \varphi_0 d(\mu_1 - \mu_0) \\
&= \mathsf{S}(\mu_0, \nu) + t \int \varphi_0 d(\mu_1 - \mu_0),
\end{aligned} \tag{A.18}$$

where the third equality uses the fact that $\int e^{\varphi_0(x) + \psi_0(y) - c(x,y)} d\nu(y) = 1$ for all $x \in \mathbb{R}^d$. Thus,

$$\liminf_{t \downarrow 0} \frac{\mathsf{S}(\mu_t, \nu) - \mathsf{S}(\mu_0, \nu)}{t} \geq \int \varphi_0 d(\mu_1 - \mu_0).$$

Likewise, we have

$$\begin{aligned}
\mathsf{S}(\mu_t, \nu) &= \int \varphi_t d\mu_t + \int \psi_t d\nu = \int \varphi_t d\mu_0 + \int \psi_t d\nu + t \int \varphi_t d(\mu_1 - \mu_0) \\
&\leq \int \varphi_0 d\mu_0 + \int \psi_0 d\nu + \int e^{\varphi_t \oplus \psi_t - c} d(\mu_0 \otimes \nu) - 1 + t \int \varphi_t d(\mu_1 - \mu_0) \\
&= \int \psi_0 d\mu_0 + \int \psi_0 d\nu + t \int \varphi_t d(\mu_1 - \mu_0) \\
&= \mathsf{S}(\mu_0, \nu) + t \int \varphi_t d(\mu_1 - \mu_0),
\end{aligned} \tag{A.19}$$

where the penultimate equality is because $\int e^{\varphi_t(x) + \psi_t(y) - c(x,y)} d\nu(y) = 1$ for all $x \in \mathbb{R}^d$. It suffices to show that for any sequence $t_n \downarrow 0$,

$$\lim_{n \to \infty} \int \varphi_{t_n} d(\mu_1 - \mu_0) = \int \varphi_0 d(\mu_1 - \mu_0). \tag{A.20}$$

Pick any subsequence n' of n. From (A.14) and the Ascoli-Arzela theorem, there exists a further subsequence n'' such that $\varphi_{t_{n''}} \to \varphi$ and $\psi_{t_{n''}} \to \psi$ locally uniformly for some (continuous) functions φ, ψ. Again, from (A.14) and the dominated convergence theorem, (φ, ψ) satisfies (2.13) for every $(x, y) \in \mathbb{R}^d \times \mathbb{R}^d$, so that they are optimal EOT potentials for (μ_0, ν). We shall now verify that $\varphi(x) = \varphi_0(x) + a$ for every $x \in \mathbb{R}^d$ for some constant $a \in \mathbb{R}$. To see this, by uniqueness of optimal EOT potentials, $\psi(y) = \psi_0(y) - a$ for ν-almost every $y \in \mathbb{R}^d$ for some constant $a \in \mathbb{R}$. We then have

$$\varphi(x) = -\log \int_{\mathbb{R}^d} e^{\psi(y) - c(x,y)} d\nu(y) = -\log \int_{\mathbb{R}^d} e^{\psi_0(y) - c(x,y)} d\nu(y) + a = \varphi_0(x) + a$$

for every $x \in \mathbb{R}^d$. Conclude that, by (A.14) and the dominated convergence theorem,

$$\lim_{n''\to\infty} \int \varphi_{t_{n''}} d(\mu_1 - \mu_0) = \int \varphi d(\mu_1 - \mu_0) = \int \varphi_0 d(\mu_1 - \mu_0).$$

Since the limit does not depend on the choice of subsequence, we have proved (A.20), completing the proof of (A.17).

Next, we prove (A.16). From the inequalities (A.18) and (A.19), we have

$$|\mathsf{S}(\mu_1, \nu) - \mathsf{S}(\mu_0, \nu)| \leq \int \varphi_0 d(\mu_1 - \mu_0) \bigvee \int \varphi_1 d(\mu_1 - \mu_0).$$

In view of the estimates (A.14), we see that the right-hand side can be bounded by $\|\mu_1 - \mu_0\|_{\infty, \mathcal{F}_\sigma}$. □

Second, we will verify that the function class $\mathcal{F}_\sigma$ is μ-Donsker.

Lemma 65. If μ is sub-Gaussian, then $\mathcal{F}_\sigma$ is μ-Donsker.

The proof relies on the the following lemma, which is a simple application of Theorem 1.1 in [199].

Lemma 66 (Lemma 8 in [152])**.** Let $\mathcal{F} \subset C^s(\mathbb{R}^d)$ be a function class where s is a positive integer with $s > d/2$, and let $\{\mathcal{X}_j\}_{j=1}^\infty$ be a cover of $\mathbb{R}^d$ consisting of nonempty bounded convex sets with bounded diameter. Set $M_j = \sup_{f\in\mathcal{F}} \|f\|_{C^s(\mathcal{X}_j)}$ with $\|f\|_{C^s(\mathcal{X}_j)} = \max_{\bar{k}\leq s} \sup_{x\in \mathrm{int}(\mathcal{X}_j)} |\partial^k f(x)|$. If $\sum_{j=1}^\infty M_j \mu(\mathcal{X}_j)^{1/2} < \infty$, then $\mathcal{F}$ is μ-Donsker.

Proof of Lemma 65. We construct a cover $\{\mathcal{X}_j\}_{j=1}^\infty$ and verify the conditions of Lemma 66 for the function class $\mathcal{F}_\sigma$. Let $B_r = \{x : \|x\| \leq r\}$. For $r = 2, 3, \ldots$, let $\{x_1^{(r)}, \ldots, x_{N_r}^{(r)}\}$ be a minimal 1-net of $B_r \setminus B_{r-1}$. Set $x_1^{(1)} = 0$ with $N_1 = 1$. From a simple volumetric argument, we see that $N_r = O(r^{d-1})$. Set $\mathcal{X}_j = \{x : \|x - x_j^{(r)}\| \leq 1\}$ for $j = \sum_{\ell=1}^{r-1} N_\ell + 1, \ldots, \sum_{\ell=1}^{r} N_\ell$. By construction, $\{\mathcal{X}_j\}_{j=1}^\infty$ forms a cover of $\mathbb{R}^d$ with diameter 2. Further, by construction, for

$M_j := \sup_{f\in\mathcal{F}_\sigma} \|f\|_{C^s(\mathcal{X}_j)}$ with $s = \max\{\lfloor d/2\rfloor+1, 2\}$, we have $\max_{\sum_{\ell=1}^{r-1} N_\ell+1\le j\le\sum_{\ell=1}^{r} N_\ell} M_j \lesssim r^s$. The function class $\mathcal{F}_\sigma$ is μ-Donsker if $\sum_{r=1}^{\infty} r^{s+d-1}\mu(B_{r-1}^c)^{1/2} < \infty$, which holds as μ is sub-Gaussian. $\square$

We are ready to prove Theorem 7.

Proof of Theorem 7. The proof for the bootstrap consistency for each case of (i) and (ii) is analogous to that of $\underline{\mathsf{W}}_p$ in Theorem 1, so we only prove the CLT results.

Part (i). Suppose μ is σ^2-sub-Gaussian. Observe that, for any constant $\bar{\sigma} > \sigma$, $\hat{\mu}_n$ is $\bar{\sigma}^2$-sub-Gaussian with probability approaching one. Indeed, by the law of large numbers,

$$\mathbb{E}_{\hat{\mu}_n}[e^{\|X\|^2/(2d\bar{\sigma}^2)}] = \frac{1}{n}\sum_{i=1}^{n} e^{\|X_i\|^2/(2d\bar{\sigma}^2)} \to \mathbb{E}_\mu[e^{\|X\|^2/(2d\bar{\sigma}^2)}] \le 2^{\sigma^2/\bar{\sigma}^2} < 2$$

a.s., so $\mathbb{E}_{\hat{\mu}_n}[e^{\|X\|^2/(2d\bar{\sigma}^2)}] \le 2$ with probability approaching one.

With this in mind, for any fixed $\bar{\sigma} > \sigma$, we apply Proposition 1 with $\mathcal{P}_0 = \{\rho \in \mathcal{P}(\mathbb{R}^d) : \rho$ is $\bar{\sigma}^2$-sub-Gaussian$\}$, $\mathcal{F} = \mathcal{F}_{\bar{\sigma}}$, $F(x) = C_{s,d,\bar{\sigma}}(1 + \|x\|^s)$ with $s = \max\{\lfloor d/2\rfloor + 1, 2\}$, and $\delta(\rho) = \mathsf{S}(\rho, \nu)$. We have already verified Conditions (a)–(c) in Proposition 1 with $\delta'_\mu(m) = m(\varphi)$. Conclude that

$$\sqrt{n}(\mathsf{S}(\hat{\mu}_n, \nu) - \mathsf{S}(\mu, \nu)) = \sqrt{n}(\delta(\hat{\mu}_n) - \delta(\mu)) \xrightarrow{d} \delta'_\mu(G_\mu) = G_\mu(\varphi) \sim N(0, \mathrm{Var}_\mu(\varphi)).$$

As the derivative δ'_μ is the point evaluation at φ, the fact that $\mathrm{Var}_\mu(\varphi)$ coincides with the semiparametric efficiency bound follows directly from Proposition 2 (note: since μ is σ^2-sub-Gaussian and $\bar{\sigma} > \sigma$, for any bounded μ-mean zero function h, $(1 + th)\mu$ is $\bar{\sigma}^2$-sub-Gaussian for sufficiently small $t > 0$).

Part (ii). Suppose μ, ν are σ^2-sub-Gaussian. Set $s = \max\{\lfloor d/2\rfloor + 1, 2\}$. Let

$$\mathcal{F}_\sigma^\oplus = \{\varphi \oplus \psi : (\varphi, \psi) \text{ satisfies (A.14)}\}.$$

We will show that, if $\mu_i, \nu_i, i = 0, 1$ are σ^2-sub-Gaussian, then

$$|\mathsf{S}(\mu_1, \nu_1) - \mathsf{S}(\mu_0, \nu_0)| \leq \|\mu_1 \otimes \nu_1 - \mu_0 \otimes \nu_0\|_{\infty, \mathcal{F}_\sigma^\oplus} \tag{A.21}$$

Let $(\varphi_{i,j}, \psi_{i,j})$ be optimal EOT potentials for (μ_i, ν_j) satisfying (A.14). Then, from (A.18), we have

$$\begin{aligned}
\mathsf{S}(\mu_1, \nu_1) - \mathsf{S}(\mu_1, \nu_1) &= \mathsf{S}(\mu_1, \nu_1) - \mathsf{S}(\mu_0, \nu_1) + \mathsf{S}(\mu_0, \nu_1) - \mathsf{S}(\mu_0, \nu_0) \\
&\geq \int \varphi_{0,1} d(\mu_1 - \mu_0) + \int \psi_{0,0} d(\nu_1 - \nu_0) \\
&= \int (\varphi_{0,1} \oplus \psi_{0,0}) d(\mu_1 \otimes \nu_1 - \mu_0 \otimes \nu_0).
\end{aligned}$$

Likewise, using (A.19), we see that

$$\mathsf{S}(\mu_1, \nu_1) - \mathsf{S}(\mu_1, \nu_1) \leq \int (\varphi_{1,1} \oplus \psi_{0,1}) d(\mu_1 \otimes \nu_1 - \mu_0 \otimes \nu_0).$$

Since all $(\varphi_{i,j}, \psi_{i,j})$ satisfy (A.14), we obtain (A.21).

Further, arguing as in the proof of Lemma 64, we have

$$\begin{aligned}
&\lim_{t \downarrow 0} \frac{\mathsf{S}(\mu_0 + t(\mu_1 - \mu_0), \nu_0 + t(\nu_1 - \nu_0)) - \mathsf{S}(\mu_0, \nu_0)}{t} \\
&= \int (\varphi_{0,0} \oplus \psi_{0,0}) d(\mu_1 \otimes \nu_1 - \mu_0 \otimes \nu_0).
\end{aligned}$$

We shall verify that

$$\sqrt{n}(\hat{\mu}_n \otimes \hat{\nu}_n - \mu \otimes \nu) \xrightarrow{d} G_{\mu \otimes \nu} \quad \text{in} \quad \ell^\infty(\mathcal{F}_\sigma^\oplus)$$

for some tight Gaussian process $G_{\mu \otimes \nu}$ in $\ell^\infty(\mathcal{F}_\sigma^\oplus)$. But this follows from the same argument as in the proof of Theorem 1 Part (ii); we omit the details for brevity.

Pick and fix any $\bar{\sigma} > \sigma$. Then $\hat{\mu}_n$ and $\hat{\nu}_n$ are $\bar{\sigma}^2$-sub-Gaussian with probability approaching one. We now apply Proposition 1 with $\mathcal{P}_0 = \{\rho_1 \otimes \rho_2 \in \mathcal{P}(\mathbb{R}^{2d}) : \rho_1, \rho_2 \text{ are } \bar{\sigma}^2\text{-sub-Gaussian}\}$, $\mathcal{F} = \mathcal{F}_{\bar{\sigma}}^\oplus, F(x, y) = C_{s,d,\bar{\sigma}}(2 + \|x\|^s + \|y\|^s)$ with $s = \max\{\lfloor d/2 \rfloor + 1, 2\}$, and $\delta(\rho_1 \otimes \rho_2) = \mathsf{S}(\rho_1, \rho_2)$. As in the proof of Theorem 1 Part (ii), we may identify $(1-t)(\mu_0 \otimes \nu_0) + t(\mu_1 \otimes \nu)$ and $((1-t)\mu_0 + t\mu_1) \otimes ((1-t)\nu_0 + t\nu_1)$ as elements of $\ell^\infty(\mathcal{F})$, and the class $\mathcal{P}_0$ is convex as a subset of $\ell^\infty(\mathcal{F})$. We have already verified Conditions (a)–(c) in Proposition 1 with $\delta'_{\mu\otimes\nu}(m) = m(\varphi \oplus \psi)$ (see also Remark 2). Conclude that

$$\begin{aligned}\sqrt{n}(\mathsf{S}(\hat{\mu}_n, \hat{\nu}_n) - \mathsf{S}(\mu, \nu)) &= \sqrt{n}(\delta(\hat{\mu}_n \otimes \hat{\nu}_n) - \delta(\mu \otimes \nu)) \\ &\stackrel{d}{\to} G_{\mu\otimes\nu}(\varphi \oplus \psi) \sim N(0, \mathrm{Var}_\mu(\varphi) + \mathrm{Var}_\nu(\psi)).\end{aligned}$$

Finally, the fact that $\mathrm{Var}_\mu(\varphi) + \mathrm{Var}_\nu(\psi)$ coincides with the semiparametric efficiency bound follows directly from Corollary 2. $\square$

APPENDIX B

SUPPLEMENT TO CHAPTER 3

B.1 Variational Formulation of Entropic Gromov-Wasserstein Distances

The GW variational form (3.8) adapts to the entropic case, for any $\varepsilon \geq 0$, as follows (see Theorem 1 in [223]),

$$\mathsf{D}_\varepsilon(\mu_0,\mu_1)^2 = \mathsf{D}_\varepsilon(\bar{\mu}_0,\bar{\mu}_1)^2 = \mathsf{S}_1(\bar{\mu}_0,\bar{\mu}_1) + \mathsf{S}_{2,\varepsilon}(\bar{\mu}_0,\bar{\mu}_1), \tag{B.1}$$

where $\mathsf{S}_1(\mu_0,\mu_1)$ is defined as in (3.8), while $\mathsf{S}_{2,\varepsilon}(\mu_0,\mu_1)$ is obtained by replacing $\mathsf{OT}_{\mathbf{A}}(\mu_0,\mu_1)$ in $\mathsf{S}_2(\mu_0,\mu_1)$ with $\mathsf{OT}_{\mathbf{A},\varepsilon}(\mu_0,\mu_1) \coloneqq \mathsf{OT}_{c_{\mathbf{A}},\varepsilon}(\mu_0,\mu_1)$, with the same cost function $c_{\mathbf{A}}(x,y) = -4\|x\|^2\|y\|^2 - 32x^\intercal\mathbf{A}y$. Again, S_1 is a constant whereas $\mathsf{S}_{2,\varepsilon}$ is a minimization problem with objective function,

$$\Phi_{(\mu_0,\mu_1),\varepsilon} : \mathbf{A} \in \mathbb{R}^{d_0\times d_1} \mapsto 32\|\mathbf{A}\|_{\mathrm{F}}^2 + \mathsf{OT}_{\mathbf{A},\varepsilon}(\mu_0,\mu_1), \tag{B.2}$$

and, if $\pi^\star$ is optimal for (3.10), then $\mathbf{A}^\star = \frac{1}{2}\int xy^\intercal d\pi^\star(x,y)$ is optimal for (B.2) provided that μ_0,μ_1 are centered (see the proof of Theorem 1 in [223]). The implications of Theorem 8 carry over to the entropic case as follows.

Theorem 23 (On minimizers of (B.2))**.** Fix $\varepsilon > 0$. If $(\mu_0,\mu_1) \in \mathcal{P}_{4,\sigma}(\mathbb{R}^{d_0}) \times \mathcal{P}_{4,\sigma}(\mathbb{R}^{d_1})$ are 4-sub-Weibull with parameter $\sigma^2 > 0$, then

1. $\Phi_{(\mu_0,\mu_1),\varepsilon}$ is locally Lipschitz continuous and coercive. Moreover, $\Phi_{(\mu_0,\mu_1),\varepsilon}$ is Fréchet differentiable at $\mathbf{A} \in \mathbb{R}^{d_0\times d_1}$ with $\left(D\Phi_{(\mu_0,\mu_1),\varepsilon}\right)_{[\mathbf{A}]}(\mathbf{B}) = 64\langle \mathbf{A} - \frac{1}{2}\int xy^\intercal d\pi_{\mathbf{A}}(x,y), \mathbf{B}\rangle_{\mathrm{F}}$, where $\pi_{\mathbf{A}}$ is the unique optimal coupling for $\mathsf{OT}_{\mathbf{A},\varepsilon}(\mu_0,\mu_1)$.

2. If $\mathbf{A}^\star$ minimizes (3.9), then $2\mathbf{A}^\star = \int xy^\intercal d\pi^\star(x,y) \in B_{\mathrm{F}}(\sqrt{M_2(\mu_0)M_2(\mu_1)})$ for the EOT plan $\pi^\star$ for $\mathsf{OT}_{\mathbf{A}^\star,\varepsilon}(\mu_0,\mu_1)$. If μ_0,μ_1 are centered, then $\pi^\star$ solves (3.7).

Compared with Theorem 8, Theorem 23 holds under the weaker assumption of 4-sub-Weibull measures with the advantage that $\Phi_{(\mu_0,\mu_1),\varepsilon}$ is guaranteed to be continuously differentiable. These improvements are obtained by leveraging the strong structural properties of solutions to the EOT problem, see Section B.1.1 for details.

B.1.1 Proof of Theorem 23

Lemma 67. $\Phi_{(\mu_0,\mu_1),\varepsilon}$ is Fréchet differentiable at any $\mathbf{A} \in \mathbb{R}^{d_0\times d_1}$ with

$$\left(D\Phi_{(\mu_0,\mu_1),\varepsilon}\right)_{[\mathbf{A}]}(\mathbf{B}) = 64\langle \mathbf{A} - \frac{1}{2}\int xy^\intercal d\pi_{\mathbf{A}}(x,y), \mathbf{B}\rangle_{\mathrm{F}},$$

where $\pi_{\mathbf{A}}$ is the unique optimal coupling for $\mathsf{OT}_{\mathbf{A},\varepsilon}(\mu_0,\mu_1)$.

Proof. As noted in the proof of Lemma 4, $\|\cdot\|_{\mathrm{F}}^2$ is Fréchet differentiable at $\mathbf{A}$ with derivative $2\langle \mathbf{A},\cdot\rangle_{\mathrm{F}}$. Now, for any $\mathbf{H} \in \mathbb{R}^{d_0\times d_1}$, we have that

$$\mathsf{OT}_{\mathbf{A}+\mathbf{H}}(\mu_0,\mu_1) - \mathsf{OT}_{\mathbf{A}}(\mu_0,\mu_1) \le \int c_{\mathbf{A}+\mathbf{H}} d\pi_{\mathbf{A}} - \int c_{\mathbf{A}} d\pi_{\mathbf{A}} = -32\int x^\intercal \mathbf{H} y d\pi_{\mathbf{A}}(x,y),$$

for the unique EOT plan $\pi_{\mathbf{A}}$ for $\mathsf{OT}_{\mathbf{A},\varepsilon}(\mu_0,\mu_1)$. Similarly,

$$\mathsf{OT}_{\mathbf{A}+\mathbf{H}}(\mu_0,\mu_1) - \mathsf{OT}_{\mathbf{A}}(\mu_0,\mu_1) \ge -32\int x^\intercal \mathbf{H} y d\pi_{\mathbf{A}+\mathbf{H}}(x,y),$$

for the optimal coupling $\pi_{\mathbf{A}+\mathbf{H}}$ for $\mathsf{OT}_{\mathbf{A}+\mathbf{H},\varepsilon}(\mu_0,\mu_1)$. Proceeding as in the proof of Lemma 4, it suffices to show that

$$\left\|\int xy^\intercal d\pi_{\mathbf{A}+\mathbf{H}}(x,y) - \int xy^\intercal d\pi_{\mathbf{A}}(x,y)\right\|_{\mathrm{F}} \to 0$$

as $\mathbf{H} \downarrow 0$.

To this end, consider an arbitrary sequence $\mathbf{H}_n$ converging to 0 and let $(\varphi_{0,n}, \varphi_{1,n})$ be EOT potentials for $\mathsf{OT}_{\mathbf{A}+\mathbf{H}_n,\varepsilon}(\mu_0, \mu_1)$. By Lemma 4 in [223], we may choose the EOT potentials as to satisfy the estimates

$$\varphi_{i,n} \leq K(1 + \|\cdot\|^2), \quad -\varphi_{i,n} \leq K(1 + \|\cdot\|^4), \quad i \in \{0, 1\},$$

for all n sufficiently large, where K is a constant that depends only on $\|\mathbf{A}\|_\infty, \sigma, d_0, d_1, \varepsilon$ (though the cited result holds for $\varepsilon = 1$, the proof of Lemma 26 shows how to extend the result to arbitrary $\varepsilon > 0$). Applying the Arzelà-Ascoli theorem, for any subsequence n' of n there exists a further subsequence n'' along which $(\varphi_{0,n}, \varphi_{1,n})$ converges to a pair of continuous functions (φ_0, φ_1) uniformly on compact sets (in particular pointwise). Note also that $c_{\mathbf{A}+\mathbf{H}_{n''}} \to c_{\mathbf{A}}$ pointwise. It follows from the 4-sub-Weibull condition and the dominated convergence theorem that

$$\begin{aligned}
e^{-\frac{\varphi_{0,n''}(x)}{\varepsilon}} &= \int e^{\frac{\varphi_{1,n''} - c_{\mathbf{A}+\mathbf{H}_{n''}}(x,\cdot)}{\varepsilon}} d\mu_1 \to \int e^{\frac{\varphi_1 - c_{\mathbf{A}}(x,\cdot)}{\varepsilon}} d\mu_1 = e^{-\frac{\varphi_0(x)}{\varepsilon}}, \\
e^{-\frac{\varphi_{1,n''}(y)}{\varepsilon}} &= \int e^{\frac{\varphi_{0,n''} - c_{\mathbf{A}+\mathbf{H}_{n''}}(\cdot,y)}{\varepsilon}} d\mu_0 \to \int e^{\frac{\varphi_0 - c_{\mathbf{A}}(\cdot,y)}{\varepsilon}} d\mu_0 = e^{-\frac{\varphi_1(y)}{\varepsilon}},
\end{aligned}$$

for $\mu_0 \otimes \mu_1$-a.e. (x, y) so that (φ_0, φ_1) solve the Schrödinger system for $\mathsf{OT}_{\mathbf{A},\varepsilon}(\mu_0, \mu_1)$. Given the characterization of the optimal coupling in terms of the EOT potentials from Section 3.2.3, we obtain that

$$\begin{aligned}
\int xy^\intercal d\pi_{\mathbf{A}+\mathbf{H}_{n''}}(x, y) &= \int xy^\intercal e^{\frac{\varphi_{0,n''}(x) + \varphi_{1,n''}(y) - c_{\mathbf{A}+\mathbf{H}_{n''}}(x,y)}{\varepsilon}} d\mu_0 \otimes \mu_1(x, y), \\
&\to \int xy^\intercal e^{\frac{\varphi_0(x) + \varphi_1(y) - c_{\mathbf{A}}(x,y)}{\varepsilon}} d\mu_0 \otimes \mu_1(x, y), \\
&= \int xy^\intercal d\pi_{\mathbf{A}}(x, y),
\end{aligned}$$

by applying the dominated convergence theorem; recalling the EOT potential estimates and the sub-Weibull assumption. Given that the limit is independent of the choice of original subsequence, we have that $\int xy^\intercal d\pi_{\mathbf{A}+\mathbf{H}}(x, y) \to \int xy^\intercal d\pi_{\mathbf{A}}(x, y)$ as $\mathbf{H} \to 0$ proving the claim. $\square$

Proof of Theorem 23. Given Lemma 67, the remainder of the proof of Theorem 23 (1) follows the same lines as the proof of Lemma 6, noting that the proof of that result requires only that μ_0 and μ_1 have finite fourth moments.

Given the coercivity of $\Phi_{(\mu_0,\mu_1),\varepsilon}$ and the fact that it is continuously differentiable, its minimizers must be stationary points. □

B.2 Convergence of minimizers of Φ

Let $\mathcal{X}_0$ and $\mathcal{X}_1$ be compact subsets of $\mathbb{R}^{d_0}$ and $\mathbb{R}^{d_1}$. We prove the following general result.

Proposition 15. Let $(\nu_{0,n})_{n\in\mathbb{N}} \subset \mathcal{P}(\mathcal{X}_0)$ and $(\nu_{1,n})_{n\in\mathbb{N}} \subset \mathcal{P}(\mathcal{X}_1)$ converge weakly to ν_0 and ν_1 respectively. Then, if $\mathbf{A}_n \in \operatorname{argmin}_{\mathbb{R}^{d_0\times d_1}} \Phi_{(\nu_{0,n},\nu_{1,n})}$, any limit point of $(\mathbf{A}_n)_{n\in\mathbb{N}}$ minimizes $\Phi_{(\nu_0,\nu_1)}$.

Proof. By Theorem 8, the minimizers of all relevant problems are contained in $B_{\mathrm{F}}(M)$ for $M = \frac{1}{2}\|\mathcal{X}_0\|_\infty\|\mathcal{X}_1\|_\infty$. Consider the functions

$$F_n(\mathbf{A}) = \begin{cases} \Phi_{(\nu_{0,n},\nu_{1,n})}(\mathbf{A}), & \text{if } \mathbf{A} \in B_{\mathrm{F}}(M), \\ +\infty, & \text{otherwise,} \end{cases}$$

$$F(\mathbf{A}) = \begin{cases} \Phi_{(\nu_0,\nu_1)}(\mathbf{A}), & \text{if } \mathbf{A} \in B_{\mathrm{F}}(M), \\ +\infty, & \text{otherwise,} \end{cases}$$

and note that, upon restriction to $B_{\mathrm{F}}(M)$, they are Lipschitz continuous with a shared Lipschitz constant (see the proof of Lemma 6). As $\mathsf{OT}_{\mathbf{A}}(\nu_{0,n}, \nu_{1,n})$ converges to $\mathsf{OT}_{\mathbf{A}}(\nu_0, \nu_1)$ for any $\mathbf{A} \in B_{\mathrm{F}}(M)$ (see Theorem 5.20 in [208]), it follows from Proposition 7.15 in [168] and Corollary 1.8.6 in [26] that F_n epi-converges to F relative to $B_{\mathrm{F}}(M)$. Explicitly, this means that $\liminf_{n\in\mathbb{N}} F_n(\mathbf{A}_n) \geq F(\mathbf{A})$ for every sequence $B_{\mathrm{F}}(M) \supset \mathbf{A}_n \to \mathbf{A}$ and

$\limsup_{n\in\mathbb{N}} F_n(\mathbf{A}_n') \leq F(\mathbf{A})$ for some sequence $B_{\mathrm{F}}(M) \supset \mathbf{A}_n' \to \mathbf{A}$, for any choice of $\mathbf{A} \in B_{\mathrm{F}}(M)$. As $B_{\mathrm{F}}(M)$ is closed, both of these conditions extend verbatim with $\mathbb{R}^{d_0\times d_1}$ in place of $B_{\mathrm{F}}(M)$. With this, we may apply Theorem 7.33 in [168] to obtain that any limit point of a sequence $\mathbf{A}_n \in \operatorname{argmin}_{\mathbb{R}^{d_0\times d_1}} F_n$ minimizes F as desired. □

As each of the empirical measures $\hat{\rho}_{k,ij,n}$ converge weakly to $\hat{\rho}_{k,ij}$ with probability 1 for $k \in \{0,1\}$ and $i,j \in [N]$, $i < j$ (see e.g. Theorem 3 in [203]), it readily follows that $\mu_{0,n}$ and $\mu_{1,n}$ converge weakly to μ_0 and μ_1 with probability 1.

B.3 Additional experiments

B.3.1 Initialization of the local solver for distributions on binary graphs

This experiment illustrates the value obtained at each of the matrices $(\mathbf{A}_T)_{T\in\mathcal{T}}$ when applying the exhaustive search procedure described in Section 3.6.3.

First, we fix isomorphic distributions, ν_0, ν_1, on the set of all binary graphs with 7 vertices, see Figure B.1 for the precise edge probabilities and the choice of permutation. Next, we sample $n \in \{10, 25, 50, 500, 5000\}$ graphs according to these distributions to construct the empirical estimators $\mu_{0,n}$ and $\mu_{1,n}$. Finally, we run Algorithm 2 at each matrix $(\mathbf{A}_T)_{T\in\mathcal{T}}$ and store the resulting value, denoted Λ_T, for each of of the 5040 permutations in $\mathcal{T}$ and for each of the pairs $(\mu_{0,n}, \mu_{1,n})$ as well as population distributions (μ_0, μ_1). These values are compiled in Figure B.2.

We highlight that, for these experiments, $\operatorname{argmin}_{T\in\mathcal{T}} \Lambda_T = \{\bar{T}\}$, where $\bar{T}$ coincides

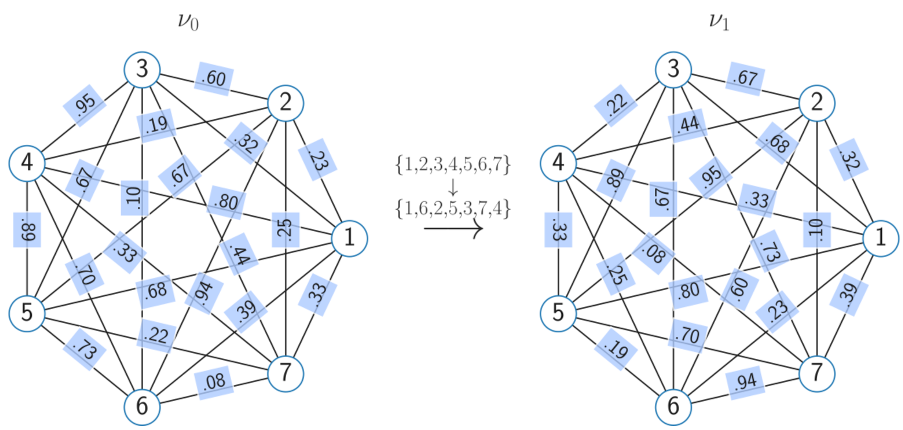


Figure B.1: Underlying distributions for this experiment. First, edge probabilities for ν_0 were sampled from the uniform distribution on $[0, 1]$, then a random permutation was chosen to generate ν_1 from ν_0 as depicted. Edge probabilities are presented with two significant figures for ease of reading.

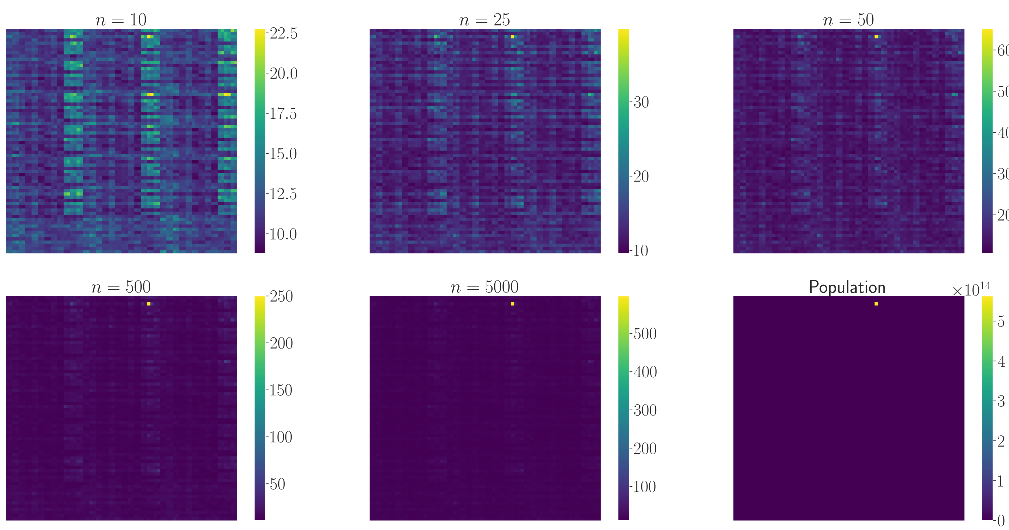


Figure B.2: Each image corresponds to performing the described experiment with different numbers of samples or on the true population distributions ν_0 and ν_1 as illustrated by the title. Each pixel of the 70×72 images above is the inverse of the value obtained starting from one of the 5040 possible initializations. Apart from the experiment with $n = 10$, the images contain a unique yellow pixel, indicating that the associated permutation results in a value significantly closer to the true value of 0 than the other permutations and, indeed, corresponds to the permutation relating ν_0 and ν_1.

with the true permutation, $T^\star$, even when only 10 samples are used. On the other hand, the relaxed graph matching approach, Algorithm 3 fails to recover $T^\star$ for $n = 10$, but succeeds otherwise.

The ratio $\frac{\min_{T\in\mathcal{T}\setminus\{\bar{T}\}}\Lambda_T}{\Lambda_{\bar{T}}}$ was computed for each experiment and is of the order $1.05, 1.30, 3.63, 4.49, 9.61$, and 8.47×10^{12} for $n = 10, 25, 50, 500, 5000$, and the exact distributions respectively. This illustrates that there is a substantial gap between the smallest and next smallest value (i.e. $\Lambda_{\bar{T}}$ is significantly smaller than Λ_T for every $T \neq \bar{T}$). Finally, as ν_0 and ν_1 are isomorphic, $\mathsf{D}(\mu_{0,n}, \mu_{1,n})$ converges to 0 as $n\to\infty$, this is reflected by the values obtained for $\Lambda_{\bar{T}}$ as n varies.

Comparison between warm-start methods

As evidenced by Figure 3.2, using relaxed graph matching to approximate the GW distance between distributions on graphs appears to result in inflated values when the number of samples is low. It is thus of interest to explore if the exhaustive search approach fares better. To this end, we repeat the procedure described in Section 3.7.1 with binary graphs on 6 nodes so that both methods can be implemented. The distributions for this experiment are displayed in Figure B.3.

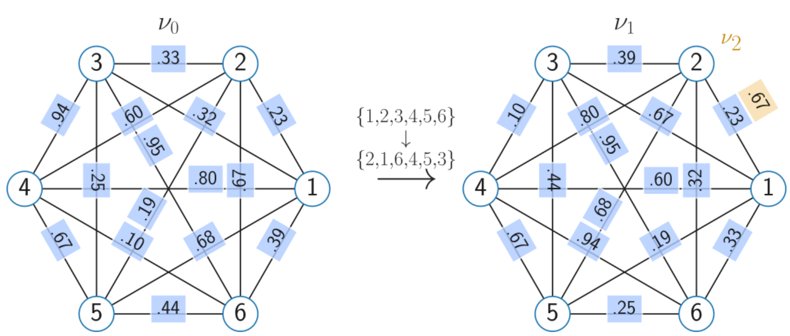


Figure B.3: Distributions for Section B.3.1. ν_0 is generated by sampling $(p_{0,ij})_{\substack{i,j\in[N]\\ i<j}}$ from the uniform distribution on $[0,1]$. ν_1 is obtained from ν_0 by a randomly chosen permutation. The modified probability for ν_2 is highlighted in orange. Edge probabilities are presented with two significant figures for ease of reading.

It can be seen from Figure B.4 that, as expected, the exhaustive search results in a better approximation of the GW distance than the relaxed graph matching approach. This

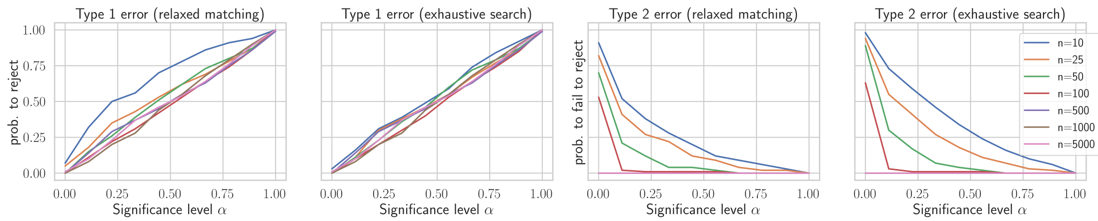


Figure B.4: Comparison of the type 1 and type 2 error between the warm-start method based on relaxed graph matching and the exhaustive search for the experiment described in Section B.3.1.

is particularly evident when $n = 10$ or 25 as reflected by the behaviors of the type 1 and type 2 errors.

B.4 The delta method

Our approach to proving limit distributions for GW distances is based on the functional delta method [182, 74, 169, 79]. To introduce this approach, we treat probability measures $(\nu_0, \nu_1) \in \mathfrak{P} \subset \mathcal{P}(\mathbb{R}^{d_0}) \times \mathcal{P}(\mathbb{R}^{d_1})$ as functionals on some function space $\mathcal{F}^{\oplus} = \mathcal{F}_0 \oplus \mathcal{F}_1 := \{f_0 \oplus f_1 : f_0 \in \mathcal{F}_0, f_1 \in \mathcal{F}_1\}$, where $\mathcal{F}_i$ is a class of Borel measurable functions on $\mathbb{R}^{d_i}$ for $i \in \{0, 1\}$. Here, we require $\mathfrak{P}$ and $\mathcal{F}^{\oplus}$ to be compatible in the sense that

$$\|\nu_0 \otimes \nu_1\|_{\infty, \mathcal{F}^{\oplus}} := \sup_{(f_0, f_1) \in \mathcal{F}_0 \times \mathcal{F}_1} |\nu_0 \otimes \nu_1(f_0 \oplus f_1)| < \infty, \quad \forall (\nu_0, \nu_1) \in \mathfrak{P},$$

whereby $\nu_0 \otimes \nu_1$ can be identified with an element of $\ell^\infty(\mathcal{F}^{\oplus})$.

With these preliminaries, we may specialize Proposition 1 in [106] as follows.

Proposition 16 (Proposition 1 in [106])**.** For $i \in \{0, 1\}$, let $\mathcal{F}_i$ be a class of Borel measurable functions on $\mathbb{R}^{d_i}$ with a finite envelope F_i and set $\mathcal{F}^{\oplus} = \mathcal{F}_0 \oplus \mathcal{F}_1$. Fix $(\mu_0, \mu_1) \in \mathcal{P}(\mathbb{R}^{d_0}) \times \mathcal{P}(\mathbb{R}^{d_1})$ and let L be a functional on a subset $\mathfrak{P}$ of $\mathcal{P}(\mathbb{R}^{d_0}) \times \mathcal{P}(\mathbb{R}^{d_1})$, where $\mathfrak{P}$ can be identified with a convex subset of $\ell^\infty(\mathcal{F}^{\oplus})$ and $\int F_0 d\nu_0 + \int F_1 d\nu_1 < \infty$ for every $(\nu_0, \nu_1) \in \mathfrak{P}$. Suppose, furthermore, that

(a) $(\mu_{0,n}, \mu_{1,n}) : \Omega \to \mathfrak{P}$, $n \in \mathbb{N}$, are pairs of random probability measures with values in $\mathfrak{P}$ such that $\sqrt{n}(\mu_{0,n} - \mu_0) \otimes \sqrt{n}(\mu_{1,n} - \mu_1) \xrightarrow{d} G_{\mu_0 \otimes \mu_1}$ in $\ell^\infty(\mathcal{F}^\oplus)$, where $G_{\mu_0 \otimes \mu_1}$ is a tight random variable in $\ell^\infty(\mathcal{F}^\oplus)$.

(b) there exists a finite constant C for which $|\mathsf{L}(\nu_0, \nu_1) - \mathsf{L}(\rho_0, \rho_1)| \leq C \|\nu_0 \otimes \nu_1 - \rho_0 \otimes \rho_1\|_{\infty, \mathcal{F}^\oplus}$ for any pairs $(\nu_0, \nu_1), (\rho_0, \rho_1) \in \mathfrak{P}$.

(c) for any pair $(\nu_0, \nu_1) \in \mathfrak{P}$, the limit $\lim_{t \downarrow 0} t^{-1}(\mathsf{L}(\mu_0 + t(\nu_0 - \mu_0), \mu_1 + t(\nu_1 - \mu_1)) - \mathsf{L}(\mu_0, \mu_1)) =: \mathsf{L}'_{(\mu_0, \mu_1)}(\nu_0 - \mu_0, \nu_1 - \mu_1)$ exists.

Then, $\sqrt{n}(\mathsf{L}(\mu_{0,n}, \mu_{1,n}) - \mathsf{L}(\mu_0, \mu_1)) \xrightarrow{d} \mathsf{L}'_{(\mu_0, \mu_1)}(G_{\mu_0 \otimes \mu_1})$.

APPENDIX C

SUPPLEMENT TO CHAPTER 4

C.1 Sharpness of variance bound from Corollary 7

Let $\mu_0 = \frac{1}{2}(\delta_0 + \delta_a)$ and $\mu_1 = \frac{1}{2}(\delta_0 + \delta_b)$ for $a \in \mathbb{R}^{d_0}$ and $b \in \mathbb{R}^{d_1}$. In this case, any coupling $\pi \in \Pi(\mu_0, \mu_1)$ is of the form $\pi_{00}\delta_{(0,0)} + \pi_{0b}\delta_{(0,b)} + \pi_{a0}\delta_{(a,0)} + \pi_{ab}\delta_{(a,b)}$ with the constraint that $\pi_{00} = \pi_{ab}$ and $\pi_{0b} = \pi_{a0} = \frac{1}{2} - \pi_{ab}$. For any $\boldsymbol{A} \in \mathcal{D}_M$, $\mathsf{OT}_{\boldsymbol{A},\varepsilon}(\mu_0, \mu_1)$ is given by

$$\begin{aligned} &\inf_{\pi \in \Pi(\mu_0,\mu_1)} \left\{ \int -4\|x\|^2\|y\|^2 - 32x^\intercal \boldsymbol{A} y \, d\pi(x,y) + \varepsilon \mathsf{D}_{\mathsf{KL}}(\pi \| \mu_0 \otimes \mu_1) \right\} \\ &= \inf_{\pi_{ab} \in (0,1/2)} \left\{ -\pi_{ab}(4\|a\|^2\|b\|^2 + 32a^\intercal \boldsymbol{A} b) + 2\varepsilon\pi_{ab}\log(4\pi_{ab}) + (1-2\pi_{ab})\,\varepsilon\log(2-4\pi_{ab}) \right\}, \end{aligned}$$

the objective is a sum of convex functions and the first-order optimality condition reads

$$4\|a\|^2\|b\|^2 + 32a^\intercal \boldsymbol{A} b = 2\varepsilon\log(4\pi_{ab}) - 2\varepsilon\log(2-4\pi_{ab}) \iff \pi_{ab} = \frac{e^z}{2\,(1+e^z)},$$

for $z = \left(2\|a\|^2\|b\|^2 + 16a^\intercal \boldsymbol{A} b\right)/\varepsilon$. Let $\pi^\star$ be the EOT coupling for $\mathsf{OT}_{\boldsymbol{A},\varepsilon}(\mu_0, \mu_1)$. For any $\boldsymbol{C} \in \mathbb{R}^{d_0 \times d_1}$,

$$\mathrm{Var}_{\pi^\star}[X^\intercal \boldsymbol{C} Y] = \pi^\star_{ab}(1-\pi^\star_{ab})(a^\intercal \boldsymbol{C} b)^2 \le \pi^\star_{ab}(1-\pi^\star_{ab})\|\boldsymbol{C}\|_F^2\|a\|^2\|b\|^2,$$

with equality for $\boldsymbol{C} = Cab^\intercal$ with $C \in \mathbb{R}$. Hence,

$$\sup_{\|\boldsymbol{C}\|_F=1} \{\mathrm{Var}_{\pi^\star}[X^\intercal \boldsymbol{C} Y]\} = \pi^\star_{ab}(1-\pi^\star_{ab})\|a\|^2\|b\|^2,$$

which can be made arbitrarily close to $\frac{1}{4}\|a\|^2\|b\|^2$ for fixed a, b by choosing $\boldsymbol{A} \in \mathcal{D}_M$ and $\varepsilon > 0$ as to make z sufficiently large. On the other hand, $\sqrt{M_4(\mu_0)M_4(\mu_1)} = \frac{1}{2}\|a\|^2\|b\|^2$, such that the variance bound in Corollary 7 is tight up to a constant factor.

C.2 Sinkhorn's Algorithm as an inexact oracle

Given $\mu_0 = \sum_{i=1}^{N_0} a_i \delta_{x^{(i)}} \in \mathcal{P}(\mathbb{R}^{d_0})$ and $\mu_1 = \sum_{j=1}^{N_1} b_j \delta_{y^{(j)}} \in \mathcal{P}(\mathbb{R}^{d_1})$, let a, b denote the corresponding (positive) probability vectors. Fix an underlying cost function $c : \mathbb{R}^{d_0} \times \mathbb{R}^{d_1} \to \mathbb{R}$ and $\varepsilon > 0$, and let $\boldsymbol{K} \in \mathbb{R}^{N_0 \times N_1}$ with $\boldsymbol{K}_{ij} = e^{-\frac{c\left(x^{(i)}, y^{(j)}\right)}{\varepsilon}}$. Consider the standard implementation of Sinkhorn's algorithm (cf. e.g. [53, 82]).

Algorithm 7 Sinkhorn Algorithm

Fix a threshold γ and a maximum iteration number $k_{\max}$.
$u_0 \leftarrow \mathbb{1}_{N_0}/N_0$
$k \leftarrow 1$
repeat
    $v_k \leftarrow b/(\boldsymbol{K}^\top u_{k-1})$
    $u_k \leftarrow a/(\boldsymbol{K} v_k)$
    $\boldsymbol{\Pi}^k \leftarrow \mathrm{diag}(u_k)\boldsymbol{K}\mathrm{diag}(v_k)$
    $k \leftarrow k + 1$
until $\|(\boldsymbol{\Pi}^k)^\top \mathbb{1}_{N_0} - b\|_2 < \gamma$ or $k > k_{\max}$
return $\boldsymbol{\Pi}^k$

In Algorithm 7 and the remainder of this section, the division of two vectors is understood componentwise. The stopping condition is based only on one of the marginal constraints, as $\boldsymbol{\Pi}^k \mathbb{1}_{N_1} = a$ by construction.

The following definitions enable describing the convergence properties of Algorithm 7; we follow the approach of [84] with only minor modifications. Let $\mathbb{R}^d_+$ denote the set of vectors with positive entries and, for $x, y \in \mathbb{R}^d_+$, let

$$\mathsf{d}_H(x, y) = \log \max_{1 \le i,j \le d} \frac{x_i y_j}{y_i x_j},$$

denote Hilbert's projective metric[1] on $\mathbb{R}^d_+$. By definition,

$$\mathsf{d}_H(x, y) = \mathsf{d}_H(x/y, \mathbb{1}_d), \tag{C.1}$$

[1] $\mathsf{d}_H(x, y) = 0$ if and only if $x = \alpha y$ for $\alpha > 0$, d_H is symmetric and satisfies the triangle inequality.

for any $x, y \in \mathbb{R}^d_+$ and, setting $x = e^w, y = e^z$ componentwise,

$$
\begin{aligned}
\mathsf{d}_H(x, y) &= \log \max_{1\le i,j\le d} e^{w_i+z_j-w_j-z_i}, \\
&= \max_{1\le i,j\le d} w_i + z_j - w_j - z_i, \\
&= \max_{1\le i\le d}(\log x_i - \log y_i) - \min_{1\le i\le d}(\log x_i - \log y_i), \\
&= \max_{1\le i\le d} \log\left(\frac{x_i}{y_i}\right) - \min_{1\le i\le d} \log\left(\frac{x_i}{y_i}\right).
\end{aligned}
\tag{C.2}
$$

It was proved in [19, 173] that multiplication with a positive matrix is a strict contraction w.r.t. d_H. Precisely,

$$
\mathsf{d}_H(\boldsymbol{A}x, \boldsymbol{A}y) \le \lambda(\boldsymbol{A})\mathsf{d}_H(x, y), \tag{C.3}
$$

for any $\boldsymbol{A} \in \mathbb{R}^{d'\times d}_+$ and $x, y \in \mathbb{R}^d_+$, where

$$
\lambda(\boldsymbol{A}) = \frac{\sqrt{\eta(\boldsymbol{A})} - 1}{\sqrt{\eta(\boldsymbol{A})} + 1} < 1, \quad \eta(\boldsymbol{A}) = \max_{\substack{1\le i,j\le d' \\ 1\le k,l\le d}} \frac{\boldsymbol{A}_{ik}\boldsymbol{A}_{jl}}{\boldsymbol{A}_{jk}\boldsymbol{A}_{il}}.
$$

Of note is that $\lambda(\boldsymbol{A}) = \lambda(\boldsymbol{A}^\intercal)$. Let

$$
E = \{\boldsymbol{A} \in \mathbb{R}^{N_0\times N_1}_+ : \boldsymbol{A} = \mathrm{diag}(x)\boldsymbol{K}\mathrm{diag}(y) \text{ for some } x \in \mathbb{R}^{N_0}_+, y \in \mathbb{R}^{N_1}_+\},
$$

and observe that if $\boldsymbol{A}, \boldsymbol{B} \in E$, there exists $x_{\boldsymbol{A},\boldsymbol{B}} \in \mathbb{R}^{N_0}_+, y_{\boldsymbol{A},\boldsymbol{B}} \in \mathbb{R}^{N_1}_+$ for which $\boldsymbol{A} = \mathrm{diag}(x_{\boldsymbol{A},\boldsymbol{B}})\boldsymbol{B}\mathrm{diag}(y_{\boldsymbol{A},\boldsymbol{B}})$. In this setting, let $\mathsf{d} : E \times E \mapsto [0, \infty)$ be such that

$$
\mathsf{d}(\boldsymbol{A}, \boldsymbol{B}) = \mathsf{d}_H(x_{\boldsymbol{A},\boldsymbol{B}}, \mathbb{1}_{N_0}) + \mathsf{d}_H(y_{\boldsymbol{A},\boldsymbol{B}}, \mathbb{1}_{N_1}),
$$

then d is a metric on E. As the EOT coupling $\boldsymbol{\Pi}^\star$ satisfies

$$
\frac{\boldsymbol{\Pi}^\star_{ij}}{a_i b_j} = e^{\frac{\varphi\left(x^{(i)}\right)+\psi\left(y^{(j)}\right)-c\left(x^{(i)},y^{(j)}\right)}{\varepsilon}},
$$

where (φ, ψ) is any pair of EOT potentials, $\boldsymbol{\Pi}^\star = \mathrm{diag}(u^\star)\boldsymbol{K}\mathrm{diag}(v^\star) \in E$ for

$$
u^\star_i = a_i e^{\frac{\varphi\left(x^{(i)}\right)}{\varepsilon}}, \quad v^\star_j = b_j e^{\frac{\psi\left(y^{(j)}\right)}{\varepsilon}}.
$$

Note that $u^\star = a/(\boldsymbol{K}v^\star)$ and $v^\star = b/(\boldsymbol{K}^\intercal u^\star)$.

In the sequel, we analyze the convergence of $\boldsymbol{\Pi}^k$ to $\boldsymbol{\Pi}^\star$ under d. The following result translates bounds on $\mathsf{d}(\boldsymbol{\Pi}^k, \boldsymbol{\Pi}^\star)$ to bounds on $\|\boldsymbol{\Pi}^k - \boldsymbol{\Pi}^\star\|_\infty$.

Lemma 68. Fix $\delta > 0$. If $\mathsf{d}(\mathbf{\Pi}^k, \mathbf{\Pi}^\star) \le \delta$, it follows that $\|\mathbf{\Pi}^k - \mathbf{\Pi}^\star\|_\infty \le e^\delta - 1$.

Proof. By Lemma 3 in [84], whenever $\mathsf{d}(\mathbf{\Pi}^k, \mathbf{\Pi}^\star) \le \delta$ it holds that

$$e^{-\delta} - 1 \le \frac{\mathbf{\Pi}^\star_{ij}}{\mathbf{\Pi}^k_{ij}} - 1 \le e^\delta - 1,$$

for every $1 \le i \le N_0, 1 \le j \le N_1$. As such,

$$|\mathbf{\Pi}^\star_{ij} - \mathbf{\Pi}^k_{ij}| \le \mathbf{\Pi}^k_{ij}\left((1 - e^{-\delta}) \vee (e^\delta - 1)\right) \le (1 - e^{-\delta}) \vee (e^\delta - 1) = e^\delta - 1,$$

yielding $\|\mathbf{\Pi}^\star - \mathbf{\Pi}^k\|_\infty \le e^\delta - 1$. □

The number of iterations required to achieve $\mathsf{d}(\mathbf{\Pi}^k, \mathbf{\Pi}^\star) \le \delta$ can be bounded as follows.

Proposition 17. Let $\mathbf{\Pi}^k$ be given by Algorithm 7 and fix $\delta > 0$. Then, if

$$k \ge 1 + \frac{1}{2\log(\lambda(\boldsymbol{K}))}\log\left(\frac{\delta(1 - \lambda(\boldsymbol{K}))}{\mathsf{d}_H((\mathbf{\Pi}^1)^\intercal \mathbb{1}_{N_0}, b)}\right),$$

it follows that $\mathsf{d}(\mathbf{\Pi}^k, \mathbf{\Pi}^\star) \le \delta$.

The proof of Proposition 17 follows immediately from Lemma 70 ahead. We first prove an auxiliary lemma.

Lemma 69. For $k \ge 1$, the iterates u_k, v_k of Algorithm 7 satisfy

$$\mathsf{d}_H(u_{k+1}, u^\star) \le \lambda(\boldsymbol{K})^2 \mathsf{d}_H(u_k, u^\star), \quad \mathsf{d}_H(v_{k+1}, v^\star) \le \lambda(\boldsymbol{K})^2 \mathsf{d}_H(v_k, v^\star),$$
$$\mathsf{d}_H(u_k, u^\star) \le \frac{\mathsf{d}_H(u_k, u_{k+1})}{1 - \lambda(\boldsymbol{K})^2}, \quad \mathsf{d}_H(v_k, v^\star) \le \frac{\mathsf{d}_H(v_k, v_{k+1})}{1 - \lambda(\boldsymbol{K})^2}.$$

Proof. To prove the first claim, we have by (C.3) that

$$\mathsf{d}_H(u_{k+1}, u^\star) = \mathsf{d}_H\left(a/(\boldsymbol{K}v_{k+1}), a/(\boldsymbol{K}v^\star)\right) = \mathsf{d}_H(\boldsymbol{K}v^\star, \boldsymbol{K}v_{k+1}) \le \lambda(\boldsymbol{K})\mathsf{d}_H(v_{k+1}, v^\star).$$

It follows similarly that $\mathsf{d}_H(v_{k+1}, v^\star) \leq \lambda(\boldsymbol{K})\mathsf{d}_H(u_k, u^\star)$. Combining these bounds,

$$\mathsf{d}_H(u_{k+1}, u^\star) \leq \lambda(\boldsymbol{K})^2\mathsf{d}_H(u_k, u^\star), \quad \mathsf{d}_H(v_{k+1}, v^\star) \leq \lambda(\boldsymbol{K})^2\mathsf{d}_H(v_k, v^\star),$$

which proves the first claim. Applying the triangle inequality yields

$$\mathsf{d}_H(u_k, u^\star) \leq \mathsf{d}_H(u_{k+1}, u^\star) + \mathsf{d}_H(u_k, u_{k+1}) \leq \lambda(\boldsymbol{K})^2\mathsf{d}_H(u_k, u^\star) + \mathsf{d}_H(u_k, u_{k+1}),$$

such that $(1 - \lambda(\boldsymbol{K})^2)\mathsf{d}_H(u_k, u^\star) \leq \mathsf{d}_H(u_k, u_{k+1})$ which proves the second claim; the same argument holds for the iterates v_k. □

Lemma 70 translates the bound from Lemma 69 to a bound on $\mathsf{d}(\Pi^k, \Pi^\star)$. We introduce the notation $\odot$ to denote the componentwise product of vectors.

Lemma 70. For $k \geq 2$, $\mathsf{d}(\boldsymbol{\Pi}^k, \boldsymbol{\Pi}^\star) \leq \frac{\lambda(\boldsymbol{K})^{2(k-1)}}{1-\lambda(\boldsymbol{K})}\mathsf{d}_H((\boldsymbol{\Pi}^1)^\intercal \mathbb{1}_{N_0}, b)$.

Proof. As $\boldsymbol{\Pi}^k = \text{diag}(u_k)\boldsymbol{K}\text{diag}(v_k)$ and $\boldsymbol{\Pi}^\star = \text{diag}(u^\star)\boldsymbol{K}\text{diag}(v^\star)$,

$$\begin{aligned}\mathsf{d}(\boldsymbol{\Pi}^k, \boldsymbol{\Pi}^\star) &= \mathsf{d}_H(u_k, u^\star) + \mathsf{d}_H(v_k, v^\star) \\ &\leq \lambda(\boldsymbol{K})^{2(k-1)}(1 + \lambda(K))d_H(v_1, v^\star) \\ &\leq \frac{\lambda(\boldsymbol{K})^{2(k-1)}}{1 - \lambda(\boldsymbol{K})} d_H(v_1, v_2),\end{aligned}$$

where both inequalities follow from Lemma 69 and its proof. Finally,

$$\mathsf{d}_H(v_1, v_2) = \mathsf{d}_H\left(v_1, \frac{b}{\boldsymbol{K}^\intercal u_1}\right) = \mathsf{d}_H\left(v_1 \odot \boldsymbol{K}^\intercal u_1, b\right) = \mathsf{d}_H((\boldsymbol{\Pi}^1)^\intercal \mathbb{1}_{N_0}, b).$$

□

Now, we demonstrate why the termination condition based on the 2-norm endows us with a δ'-oracle approximation and provide a bound on the number of iterations required to achieve it. Theorem 1 in [77] proves that there exists $\bar{k} \leq 1 + \frac{R}{\gamma}$ satisfying

$$\|u_{\bar{k}} \odot \boldsymbol{K} v_{\bar{k}+1} - a\|_1 + \|(\boldsymbol{\Pi}^k)^\intercal \mathbb{1}_{N_0} - b\|_1 \leq \gamma,$$

for $R = -2\log\left(e^{-\|C\|_\infty/\varepsilon}\min_{\substack{1\le i\le N_0\\1\le j\le N_1}} a_i \wedge b_j\right)$. This gives a bound on the maximal number of iterations to achieve the 2-norm termination condition via the standard inequality $\|\cdot\|_2 \le \|\cdot\|_1$. We clarify that the analysis in [77] is for a slightly different implementation of Sinkhorn's algorithm. First, running Algorithm 7 is tantamount to running their algorithm with reversed marginals. Next, one iteration of Algorithm 7 corresponds to two iterations in their analysis. Finally, their approach uses $\mathbb{1}_{N_0}$ rather than $\mathbb{1}_{N_0}/N_0$ for the initialization. This difference is innocuous for achieving the termination condition, as the iterates are identical up to multiplying v_k by N_0 and dividing u_k by N_0; $\mathbf{\Pi}^k$ is invariant this operation.

We now bound d_H in terms of the Euclidean distance with the aim of controlling $\mathsf{d}(\mathbf{\Pi}^k, \mathbf{\Pi}^\star)$ by $\|(\mathbf{\Pi}^k)^\top \mathbb{1}_{N_0} - b\|_2$.

Lemma 71. Let $r, s \in \mathbb{R}^d_+$ be arbitrary, then

$$\mathsf{d}_H(s, r) \le (r_{i^\star}^{-1} + s_{i_\star}^{-1})\|r - s\|_2,$$

where $i^* \in \operatorname{argmax}_{1\le i\le d} \frac{s_i}{r_i}$ and $i_* \in \operatorname{argmin}_{1\le i\le d} \frac{s_i}{r_i}$.

Proof. We have by (C.2) that

$$\mathsf{d}_H(s, r) = \max_{1\le i\le N_0} \log\left(\frac{s_i}{r_i}\right) - \min_{1\le i\le N_0} \log\left(\frac{s_i}{r_i}\right).$$

Observe that $1 - \frac{r_i}{s_i} \le \log\left(\frac{s_i}{r_i}\right) \le \frac{s_i}{r_i} - 1$, as follows the inequalities $\frac{x}{1+x} \le \log(1 + x) \le x$ for $x > -1$. Whence,

$$\begin{aligned}\mathsf{d}_H(s, r) &\le r_{i^\star}^{-1}\,(s_{i^\star} - r_{i^\star}) - s_{i_\star}^{-1}\,(s_{i_\star} - r_{i_\star})\\ &\le \left(r_{i^\star}^{-1} + s_{i_\star}^{-1}\right)\|s - r\|_2\,.\end{aligned}$$

∎

Note that the bound in Lemma 71 is symmetric in the sense that interchanging s and r does not modify the constant. This can be seen by noting that $\operatorname{argmax}_{1\le i\le d} \frac{s_i}{r_i} =$

$\operatorname{argmin}_{1\le i\le d}\frac{r_i}{s_i}$ and $\operatorname{argmin}_{1\le i\le d}\frac{s_i}{r_i} = \operatorname{argmax}_{1\le i\le d}\frac{r_i}{s_i}$. By combining Lemmas 70 and 71 we arrive at the desired result.

Proposition 18. Let $\underline{b} = \min_{1\le i\le N_1} b_i$ and set $0 < \gamma < \underline{b}$. Then, for $k \ge 1$,

$$\mathsf{d}(\mathbf{\Pi}^k, \mathbf{\Pi}^\star) \le \frac{\left((\mathbf{\Pi}^k)^\intercal \mathbb{1}_{N_0}\right)_{i_\star}^{-1} + b_{i^\star}^{-1}}{1-\lambda(\boldsymbol{K})}\|(\mathbf{\Pi}^k)^\intercal \mathbb{1}_{N_0} - b\|$$

where $i^* \in \operatorname{argmax}_{1\le i\le N_1}\frac{\left((\mathbf{\Pi}^k)^\intercal \mathbb{1}_{N_0}\right)_i}{b_i}$ and $i_* \in \operatorname{argmin}_{1\le i\le N_1}\frac{\left((\mathbf{\Pi}^1)^\intercal \mathbb{1}_{N_0}\right)_i}{b_i}$. Further, there exists $\bar{k} \le 1 + \frac{R}{\gamma}$ for which $\|\mathbf{\Pi}^{\bar{k}\intercal}\mathbb{1}_{N_0} - b\| \le \gamma$ and $\mathsf{d}(\mathbf{\Pi}^{\bar{k}}, \mathbf{\Pi}^\star) \le \frac{(\underline{b}-\gamma)^{-1}+\underline{b}^{-1}}{1-\lambda(\boldsymbol{K})}\gamma$. In particular, setting

$$\gamma = \bar{\alpha}\underline{b} \text{ for } \bar{\alpha} = \frac{\delta(1-\lambda(\boldsymbol{K})) + 2 - \sqrt{\delta^2(1-\lambda(\boldsymbol{K}))^2 + 4}}{2} \in (0,1),$$

it holds that $\mathsf{d}(\mathbf{\Pi}^k, \mathbf{\Pi}^\star) = \delta$.

Proof. From the proof of Lemma 70,

$$\mathsf{d}(\mathbf{\Pi}^k, \mathbf{\Pi}^\star) \le (1+\lambda(\boldsymbol{K}))\mathsf{d}_H(v_k, v^\star) \le \frac{\mathsf{d}_H(v_k, v_{k+1})}{1-\lambda(\boldsymbol{K})} = \frac{\mathsf{d}_H((\mathbf{\Pi}^k)^\intercal \mathbb{1}_{N_0}, b)}{1-\lambda(\boldsymbol{K})},$$

where the final inequality and equality stem from Lemma 69 and its proof. Applying Lemma 71 proves the first claim.

As for the second claim, it is clear from the discussion preceding Lemma 71 that $\|(\mathbf{\Pi}^{\bar{k}})^\intercal \mathbb{1}_{N_0} - b\| \le \gamma$ for some $\bar{k} \le 1 + \frac{R}{\gamma}$. Now, let $w = (\mathbf{\Pi}^{\bar{k}})^\intercal \mathbb{1}_{N_0}$ and observe that $\|w-b\|_\infty \le \|w-b\| \le \gamma$ such that $b_i - \gamma \le w_i \le b_i + \gamma$ for $i = 1,\dots,N_1$. Hence $w_i^{-1} \le (b_i-\gamma)^{-1} \le (\underline{b}-\gamma)^{-1}$ as $\gamma < \underline{b}$. Applying this bound to the previous inequality proves the claim.

For the final claim, observe that $\bar{\alpha}$ solves the equation $\frac{2\alpha-\alpha^2}{1-\alpha} = \delta(1-\lambda(\boldsymbol{K}))$ for $\alpha \in (0,1)$ (indeed, $\delta(1-\lambda(\boldsymbol{K})) < \sqrt{\delta^2(1-\lambda(\boldsymbol{K}))^2+4} < \delta(1-\lambda(\boldsymbol{K}))+2$). Setting $\gamma = \bar{\alpha}\underline{b}$ in the previously derived bound on $\mathsf{d}(\mathbf{\Pi}^k, \mathbf{\Pi}^\star)$ gives

$$\mathsf{d}(\mathbf{\Pi}^k, \mathbf{\Pi}^\star) \le \frac{(1-\bar{\alpha})^{-1}+1}{1-\lambda(\boldsymbol{K})}\bar{\alpha} = \frac{2\bar{\alpha}-\bar{\alpha}^2}{(1-\bar{\alpha})(1-\lambda(\boldsymbol{K}))} = \delta.$$

$\square$

The proof of Proposition 11 follows by combining Propositions 17 and 18; the maximal number of iterations for Algorithm 7 to output a matrix $\tilde{\mathbf{\Pi}}$ satisfying $\mathsf{d}(\tilde{\mathbf{\Pi}}, \mathbf{\Pi}^\star) \leq \delta$ is

$$\begin{aligned}\tilde{k} = \min\Bigg\{ & 1 + \frac{1}{2\log(\lambda(\boldsymbol{K}))} \log\left(\frac{\delta(1-\lambda(\boldsymbol{K}))}{\mathsf{d}_H((\mathbf{\Pi}^1)^\intercal \mathbb{1}_{N_0}, b)}\right), \\ & 1 - \frac{4\underline{b}^{-1}}{\delta(1-\lambda(\boldsymbol{K})) + 2 - \sqrt{\delta^2(1-\lambda(\boldsymbol{K}))^2 + 4}} \log\left(e^{-\|\boldsymbol{C}\|_\infty/\varepsilon} \min_{\substack{1\leq i\leq N_0 \\ 1\leq j\leq N_1}} a_i \wedge b_j\right)\Bigg\},\end{aligned} \tag{C.4}$$

which corresponds to an $(e^\delta - 1)$-oracle for the entropic transport plan in light of Lemma 68.

C.3 Convergence of Algorithm 6

In what follows, we slightly adapt the proof of Theorem 2 in [89] to conform to the inexact setting. We first clarify that they treat the composite problem

$$\inf_{x\in\mathbb{R}^d} f(x) + g(x) + Q(x),$$

where f is L'-smooth and non-convex, g is L''-smooth and convex, and Q is non-smooth and convex with a bounded domain. Hence $f + g$ is $L = L' + L''$ smooth and possibly non-convex.

Our problem conforms to this setting (up to vectorization) with $f = \mathsf{OT}_{(\cdot),\varepsilon}(\mu_0, \mu_1)$, $g = 32\|\cdot\|_F^2$, and $Q = \mathcal{I}_{\mathcal{D}_M}$, the indicator function of the set $\mathcal{D}_M$, defined by

$$\mathcal{I}_{\mathcal{D}_M}(\boldsymbol{A}) = \begin{cases} 0, & \text{if } \boldsymbol{A} \in \mathcal{D}_M, \\ +\infty, & \text{otherwise.}\end{cases}$$

When Φ is convex, we set $f = 0$ and $g = \Phi$ hence $L' = 0$, $L = L''$.

As Φ is L-smooth, by Lemma 5 in [89],

$$\Phi(\boldsymbol{B}_k) \leq \Phi(\boldsymbol{A}_k) + \operatorname{tr}\left(D\Phi_{[\boldsymbol{A}_k]}^{\intercal}(\boldsymbol{B}_k - \boldsymbol{A}_k)\right) + \frac{L}{2}\|\boldsymbol{B}_k - \boldsymbol{A}_k\|_F^2, \tag{C.5}$$

and for any $\boldsymbol{H} \in \mathbb{R}^{d_0 \times d_1}$, letting L' denote the Lipschitz constant of $\mathsf{OT}_{(\cdot),\varepsilon}(\mu_0, \mu_1)$, the same result yields

$$\begin{aligned}
&\Phi(\boldsymbol{A}_k) - ((1-\tau_k)\Phi(\boldsymbol{B}_{k-1}) + \tau_k\Phi(\boldsymbol{H})) \\
&= \tau_k\left(\Phi(\boldsymbol{A}_k) - \Phi(\boldsymbol{H})\right) + (1-\tau_k)\left(\Phi(\boldsymbol{A}_k) - \Phi(\boldsymbol{B}_{k-1})\right) \\
&\leq \tau_k\left(\operatorname{tr}\left(D\Phi_{[\boldsymbol{A}_k]}^{\intercal}(\boldsymbol{A}_k - \boldsymbol{H})\right) + \frac{L'}{2}\|\boldsymbol{H} - \boldsymbol{A}_k\|_F^2\right) \\
&+ (1-\tau_k)\left(\operatorname{tr}\left(D\Phi_{[\boldsymbol{A}_k]}^{\intercal}(\boldsymbol{A}_k - \boldsymbol{B}_{k-1})\right) + \frac{L'}{2}\|\boldsymbol{B}_{k-1} - \boldsymbol{A}_k\|_F^2\right) \\
&= \operatorname{tr}\left(D\Phi_{[\boldsymbol{A}_k]}^{\intercal}(\boldsymbol{A}_k - \tau_k\boldsymbol{H} - (1-\tau_k)\boldsymbol{B}_{k-1})\right) \\
&+ \frac{L'\tau_k}{2}\|\boldsymbol{H} - \boldsymbol{A}_k\|_F^2 + \frac{L'(1-\tau_k)}{2}\underbrace{\|\boldsymbol{B}_k - \boldsymbol{A}_k\|_F^2}_{\tau_k^2\|\boldsymbol{B}_{k-1} - \boldsymbol{C}_{k-1}\|_F^2},
\end{aligned} \tag{C.6}$$

recalling the update $\boldsymbol{A}_k = \tau_k\boldsymbol{C}_{k-1} + (1-\tau_k)\boldsymbol{B}_{k-1}$.

Denote the subdifferential of $\mathcal{I}_{\mathcal{D}_M}$ at $\boldsymbol{A} \in \mathbb{R}^{d_0 \times d_1}$ by

$$\partial\mathcal{I}_{\mathcal{D}_M}(\boldsymbol{A}) := \left\{\boldsymbol{P} \in \mathbb{R}^{d_0 \times d_1} : \mathcal{I}_{\mathcal{D}_M}(\boldsymbol{X}) - \mathcal{I}_{\mathcal{D}_M}(\boldsymbol{A}) \geq \operatorname{tr}(\boldsymbol{P}^{\intercal}(\boldsymbol{X} - \boldsymbol{A})),\ \text{for every } \boldsymbol{X} \in \mathbb{R}^{d_0 \times d_1}\right\}.$$

As $\boldsymbol{C}_k$ is optimal for the problem $\operatorname{argmin}_{\boldsymbol{V} \in \mathbb{R}^{d_0 \times d_1}}\left\{\frac{1}{2\gamma_k}\|\boldsymbol{V} - (\boldsymbol{C}_{k-1} - \gamma_k\boldsymbol{G}_k)\|_F^2 + \mathcal{I}_{\mathcal{D}_M}(\boldsymbol{V})\right\}$, there exists $\boldsymbol{P} \in \partial\mathcal{I}_{\mathcal{D}_M}(\boldsymbol{C}_k)$ for which $\boldsymbol{G}_k + \boldsymbol{P} + \frac{1}{\gamma_k}(\boldsymbol{C}_k - \boldsymbol{C}_{k+1}) = 0$ (see Theorem 23.8, Theorem 25.1, and p. 264 in [167]). Thus, for any $\boldsymbol{U} \in \mathbb{R}^{d_0 \times d_1}$,

$$\begin{aligned}
\operatorname{tr}((\boldsymbol{G}_k + \boldsymbol{P})^{\intercal}(\boldsymbol{C}_k - \boldsymbol{U})) &= \frac{1}{\gamma_k}\operatorname{tr}((\boldsymbol{C}_k - \boldsymbol{C}_{k-1})^{\intercal}(\boldsymbol{U} - \boldsymbol{C}_k)) \\
&= \frac{1}{2\gamma_k}\left(\|\boldsymbol{C}_{k-1} - \boldsymbol{U}\|_F^2 - \|\boldsymbol{C}_k - \boldsymbol{U}\|_F^2 - \|\boldsymbol{C}_k - \boldsymbol{C}_{k-1}\|_F^2\right),
\end{aligned}$$

where the final line follows from some simple algebra. As $\boldsymbol{P} \in \partial\mathcal{I}_{\mathcal{D}_M}(\boldsymbol{C}_k)$, $\operatorname{tr}(\boldsymbol{P}^{\intercal}(\boldsymbol{C}_k - \boldsymbol{U})) \geq \mathcal{I}_{\mathcal{D}_M}(\boldsymbol{C}_k) - \mathcal{I}_{\mathcal{D}_M}(\boldsymbol{U}) = -\mathcal{I}_{\mathcal{D}_M}(\boldsymbol{U})$, whence

$$\operatorname{tr}\left(\boldsymbol{G}_k^{\intercal}(\boldsymbol{C}_k - \boldsymbol{U})\right) \leq \mathcal{I}_{\mathcal{D}_M}(\boldsymbol{U}) + \frac{1}{2\gamma_k}\left(\|\boldsymbol{C}_{k-1} - \boldsymbol{U}\|_F^2 - \|\boldsymbol{C}_k - \boldsymbol{U}\|_F^2 - \|\boldsymbol{C}_k - \boldsymbol{C}_{k-1}\|_F^2\right). \tag{C.7}$$

By the same steps applied to the other subproblem with $\boldsymbol{B}_k$ and $\boldsymbol{A}_k$ taking the place of $\boldsymbol{C}_k$ and $\boldsymbol{C}_{k-1}$ respectively,

$$\operatorname{tr}\left(\boldsymbol{G}_k^{\intercal}(\boldsymbol{B}_k-\boldsymbol{U})\right) \leq \mathcal{I}_{\mathcal{D}_M}(\boldsymbol{U}) + \frac{1}{2\beta_k}\left(\|\boldsymbol{A}_k-\boldsymbol{U}\|_F^2 - \|\boldsymbol{B}_k-\boldsymbol{U}\|_F^2 - \|\boldsymbol{B}_k-\boldsymbol{A}_k\|_F^2\right).$$

Setting $\boldsymbol{U} = \tau_k\boldsymbol{C}_k + (1-\tau_k)\boldsymbol{B}_{k-1} \in \mathcal{D}_M$ (by convexity) in the previous display, bounding $-\|\boldsymbol{B}_k-\boldsymbol{U}\|_F^2$ above by 0, and recalling that $\boldsymbol{A}_k = \tau_k\boldsymbol{C}_{k-1} + (1-\tau_k)\boldsymbol{B}_{k-1}$ such that $\boldsymbol{A}_k - \boldsymbol{U} = \tau_k(\boldsymbol{C}_{k-1}-\boldsymbol{C}_k)$,

$$\operatorname{tr}\left(\boldsymbol{G}_k^{\intercal}(\boldsymbol{B}_k-\tau_k\boldsymbol{C}_k+(1-\tau_k)\boldsymbol{B}_{k-1})\right) \leq \frac{1}{2\beta_k}\left(\tau_k^2\|\boldsymbol{C}_k-\boldsymbol{C}_{k-1}\|_F^2 - \|\boldsymbol{B}_k-\boldsymbol{A}_k\|_F^2\right).$$

Combining with (C.7) upon scaling by τ_k,

$$\begin{aligned}\operatorname{tr}\left(\boldsymbol{G}_k^{\intercal}(\boldsymbol{B}_k-\tau_k\boldsymbol{U}+(1-\tau_k)\boldsymbol{B}_{k-1})\right) \leq \tau_k\mathcal{I}_{\mathcal{D}_M}(\boldsymbol{U}) &+ \frac{1}{2\beta_k}\left(\tau_k^2\|\boldsymbol{C}_k-\boldsymbol{C}_{k-1}\|_F^2 - \|\boldsymbol{B}_k-\boldsymbol{A}_k\|_F^2\right),\\ &+ \frac{\tau_k}{2\gamma_k}\left(\|\boldsymbol{C}_{k-1}-\boldsymbol{U}\|_F^2 - \|\boldsymbol{C}_k-\boldsymbol{U}\|_F^2 - \|\boldsymbol{C}_k-\boldsymbol{C}_{k-1}\|_F^2\right),\end{aligned}$$

by the choice of $\tau_k, \beta_k, \gamma_k$, we have that $\frac{\tau_k^2}{\beta_k} - \frac{\tau_k}{\gamma_k} \leq 0$ such that

$$\begin{aligned}\operatorname{tr}\left(\boldsymbol{G}_k^{\intercal}(\boldsymbol{B}_k-\tau_k\boldsymbol{U}+(1-\tau_k)\boldsymbol{B}_{k-1})\right) \leq \tau_k\mathcal{I}_{\mathcal{D}_M}(\boldsymbol{U}) &+ \frac{\tau_k}{2\gamma_k}\left(\|\boldsymbol{C}_{k-1}-\boldsymbol{U}\|_F^2 - \|\boldsymbol{C}_k-\boldsymbol{U}\|_F^2\right)\\ &- \frac{1}{2\beta_k}\|\boldsymbol{B}_k-\boldsymbol{A}_k\|_F^2.\end{aligned}$$

Combining the equation above with (C.5) and (C.6) and setting $\boldsymbol{H} = \boldsymbol{U} \in \mathcal{D}_M$ (otherwise the bound is vacuous),

$$\begin{aligned}&\Phi(\boldsymbol{B}_k) - \Phi(\boldsymbol{H})\\ &\quad\leq (1-\tau_k)\left(\Phi(\boldsymbol{B}_{k-1})-\Phi(\boldsymbol{H})\right) + \operatorname{tr}\left(D\Phi_{[\boldsymbol{A}_k]}^{\intercal}(\boldsymbol{B}_k-\tau_k\boldsymbol{H}-(1-\tau_k)\boldsymbol{B}_{k-1})\right)\\ &\quad+ \frac{L'\tau_k}{2}\|\boldsymbol{H}-\boldsymbol{A}_k\|_F^2 + \frac{L'(1-\tau_k)}{2}\tau_k^2\|\boldsymbol{B}_{k-1}-\boldsymbol{C}_{k-1}\|_F^2 + \frac{L}{2}\|\boldsymbol{B}_k-\boldsymbol{A}_k\|_F^2\\ &\quad\leq (1-\tau_k)\left(\Phi(\boldsymbol{B}_{k-1})-\Phi(\boldsymbol{H})\right) + \delta' + \frac{\tau_k}{2\gamma_k}\left(\|\boldsymbol{C}_{k-1}-\boldsymbol{H}\|_F^2 - \|\boldsymbol{C}_k-\boldsymbol{H}\|_F^2\right)\\ &\quad+ \frac{L'\tau_k}{2}\|\boldsymbol{H}-\boldsymbol{A}_k\|_F^2 + \frac{L'(1-\tau_k)}{2}\tau_k^2\|\boldsymbol{B}_{k-1}-\boldsymbol{C}_{k-1}\|_F^2 + \left(\frac{L}{2}-\frac{1}{2\beta_k}\right)\|\boldsymbol{B}_k-\boldsymbol{A}_k\|_F^2,\end{aligned}$$

where the inequality follows from the δ-oracle which implies the bound (cf. (4.16))

$$\sup_{\boldsymbol{Y},\boldsymbol{Z}\in\mathcal{D}_M}\left\{\left|\operatorname{tr}\left(\boldsymbol{G}_k - D\Phi_{[\boldsymbol{A}_k]}\right)^\intercal(\boldsymbol{Y}-\boldsymbol{Z})\right)\right|\right\} \leq \delta',$$

observing that $\boldsymbol{B}_k, \tau_k\boldsymbol{H} + (1-\tau_k)\boldsymbol{B}_{k-1} \in \mathcal{D}_M$ by convexity ($\tau_k \in (0,1]$).

Applying Lemma 1 in [89] yields, for $A_i = \frac{2}{i(i+1)}$,

$$\begin{aligned}
\frac{\Phi(\boldsymbol{B}_k)-\Phi(\boldsymbol{H})}{A_k} &\leq \sum_{i=1}^{k} A_i^{-1}\left(\delta' + \frac{\tau_i}{2\gamma_i}\left(\|\boldsymbol{C}_{i-1}-\boldsymbol{H}\|_F^2 - \|\boldsymbol{C}_i-\boldsymbol{H}\|_F^2\right) + \frac{L'\tau_i}{2}\|\boldsymbol{H}-\boldsymbol{A}_i\|^2\right.\\
&\quad \left.+\frac{L'(1-\tau_i)}{2}\tau_i^2\|\boldsymbol{B}_{i-1}-\boldsymbol{C}_{i-1}\|_F^2 + \left(\frac{L}{2}-\frac{1}{2\beta_i}\right)\|\boldsymbol{B}_i-\boldsymbol{A}_i\|^2\right)\\
&\leq \frac{\|\boldsymbol{C}_0-\boldsymbol{H}\|_F^2}{2\gamma_1} + \sum_{i=1}^{k} A_i^{-1}\left(\delta' + \frac{L'\tau_i}{2}\|\boldsymbol{H}-\boldsymbol{A}_i\|^2\right.\\
&\quad \left.+\frac{L'(1-\tau_i)}{2}\tau_i^2\|\boldsymbol{B}_{i-1}-\boldsymbol{C}_{i-1}\|_F^2 + \left(\frac{L}{2}-\frac{1}{2\beta_i}\right)\|\boldsymbol{B}_i-\boldsymbol{A}_i\|^2\right).
\end{aligned}$$

By convexity of $\|\cdot\|_F^2$,

$$\begin{aligned}
&\|\boldsymbol{H}-\boldsymbol{A}_i\|_F^2 + \tau_i(1-\tau_i)\|\boldsymbol{B}_{i-1}-\boldsymbol{C}_{i-1}\|_F^2\\
&\leq 2\left(\|\boldsymbol{H}\|_F^2 + \|\boldsymbol{A}_i\|_F^2 + \tau_i(1-\tau_i)\left(\|\boldsymbol{B}_{i-1}\|_F^2 + \|\boldsymbol{C}_{i-1}\|_F^2\right)\right)\\
&\leq 2\left(\|\boldsymbol{H}\|_F^2 + (1-\tau_i)\|\boldsymbol{B}_{i-1}\|_F^2 + \tau_i\|\boldsymbol{C}_{i-1}\|_F^2 + \tau_i(1-\tau_i)\left(\|\boldsymbol{B}_{i-1}\|_F^2 + \|\boldsymbol{C}_{i-1}\|_F^2\right)\right)\\
&\leq 2\left(\|\boldsymbol{H}\|_F^2 + (1+\tau_i(1-\tau_i))\max_{\mathcal{D}_M}\|\cdot\|_F^2\right)\\
&\leq 2\left(\|\boldsymbol{H}\|_F^2 + \frac{5}{16}M^2\right),
\end{aligned}$$

observing that $\tau_i \in (0,1]$ hence $\tau_i(1-\tau_i) \leq \frac{1}{4}$. Thus, for $\boldsymbol{H} = \boldsymbol{B}^\star$, a global minimizer of Φ,

$$\begin{aligned}
\frac{\Phi(\boldsymbol{B}_k)-\Phi(\boldsymbol{B}^\star)}{A_k} + \sum_{i=1}^{k}\frac{1-L\beta_i}{2A_i\beta_i}\|\boldsymbol{B}_i-\boldsymbol{A}_i\|_F^2 &\leq \frac{\|\boldsymbol{C}_0-\boldsymbol{B}^\star\|_F^2}{2\gamma_1}\\
&\quad + \sum_{i=1}^{k} A_i^{-1}\left(\delta' + L'\tau_i\left(\|\boldsymbol{B}^\star\|_F^2 + \frac{5}{16}M^2\right)\right).
\end{aligned}$$

By construction, $\sum_{i=1}^{k} A_i^{-1} L' \tau_i = \frac{L'}{A_k}$, and $\Phi(\boldsymbol{B}_k) - \Phi(\boldsymbol{B}^\star) \geq 0$. It follows that

$$\begin{aligned}
&\min_{i=1}^{k} \left\|\beta_i^{-1} (\boldsymbol{B}_i - \boldsymbol{A}_i)\right\|_F^2 \\
&\leq 2\left(\sum_{i=1}^{k} \frac{\beta_i (1 - L\beta_i)}{A_i}\right)^{-1} \left(\frac{\|\boldsymbol{C}_0 - \boldsymbol{B}^\star\|_F^2}{2\gamma_1} + \sum_{i=1}^{k} A_i^{-1}\delta' + \frac{L'}{A_k}\left(\|\boldsymbol{B}^\star\|_F^2 + \frac{5}{16} M^2 d_0^2 d_1^2\right)\right).
\end{aligned}$$

As $\beta_i = \frac{L}{2}$, $\gamma_1 = \frac{1}{4L}$, and $A_i = \frac{2}{i(i+1)}$, $\sum_{i=1}^{k} \frac{\beta_i(1-L\beta_i)}{A_i} = \frac{1}{4L} \sum_{i=1}^{k} A_i^{-1} = \frac{k(k+1)(k+2)}{24L}$, so

$$\min_{i=1}^{k} \left\|\beta_i^{-1} (\boldsymbol{B}_i - \boldsymbol{A}_i)\right\|_F^2 \leq \frac{96L^2}{k(k+1)(k+2)} \|\boldsymbol{C}_0 - \boldsymbol{B}^\star\|_F^2 + 8L\delta' + \frac{24LL'}{N}\left(\|\boldsymbol{B}^\star\|_F^2 + \frac{5M^2}{16}\right).$$

This proves the claimed result in the non-convex setting.

In the convex regime, recall from the prior discussion that we may set $L' = 0$ in the previous display, proving the claim.

C.4 Additional Results

C.4.1 Proof of Lemma 33

The proof of Lemma 33 follows from the following lemma coupled with the chain rule for Fréchet differentiable maps.

Lemma 72. Let $\mu_i \in \mathcal{P}(\mathbb{R}^{d_i})$, for $i = 0, 1$, be compactly supported with $\mathrm{spt}(\mu_i) = S_i$. Then, the map $f \in C(S_0 \times S_1) \mapsto \left(\int e^{f(\cdot,y)} d\mu_1(y), \int e^{f(x,\cdot)} d\mu_0(x)\right) \in C(S_0) \times C(S_1)$ is smooth with first derivative at $f \in C(S_0 \times S_1)$ given by

$$h \in C(S_0 \times S_1) \mapsto \left(\int h(\cdot, y) e^{f(\cdot,y)} d\mu_1(y), \int h(x, \cdot) e^{f(x,\cdot)} d\mu_0(x)\right) \in C(S_0) \times C(S_1).$$

Proof. First, we show that the map $f \in C(S_0 \times S_1) \mapsto e^f \in C(S_0 \times S_1)$ is Fréchet differentiable with $D(e^{(\cdot)})_{[f]}(h) = he^f$. Fix $f \in C(S_0 \times S_1)$ and consider

$$\lim_{\substack{h\in C(S_0\times S_1)\\ \|h\|_{\infty,S_0\times S_1}\to 0}} \frac{\left\|e^{f+h} - e^f - he^f\right\|_{\infty,S_0\times S_1}}{\|h\|_{\infty,S_0\times S_1}} \le \|e^f\|_{\infty,S_0\times S_1} \lim_{\substack{h\in C(S_0\times S_1)\\ \|h\|_{\infty,S_0\times S_1}\to 0}} \frac{\left\|e^h - 1 - h\right\|_{\infty,S_0\times S_1}}{\|h\|_{\infty,S_0\times S_1}}.$$

Fix arbitrary $(x, y) \in S_0 \times S_1$. By a Taylor expansion,

$$e^{h(x,y)} = 1 + h(x, y) + \frac{1}{2}e^{\xi(x,y)}h^2(x, y),$$

where $|\xi(x, y)| \in [0, |h(x, y)|]$ i.e. $\|\xi\|_{\infty,S_0\times S_1} \le \|h\|_{\infty,S_0\times S_1}$. That is,

$$\lim_{\substack{h\in C(S_0\times S_1)\\ \|h\|_{\infty,S_0\times S_1}\to 0}} \frac{\left\|e^h - 1 - h\right\|_{\infty,S_0\times S_1}}{\|h\|_{\infty,S_0\times S_1}} \le \lim_{\substack{h\in C(S_0\times S_1)\\ \|h\|_{\infty,S_0\times S_1}\to 0}} \frac{1}{2}e^{\|\xi\|_{\infty,S_0\times S_1}}\|h\|_{\infty,S_0\times S_1} = 0.$$

On the other hand, the derivative of $f \in C(S_0 \times S_1) \mapsto \int f(x, y)d\mu_1(y) \in C(S_0)$ at any point is given by $h \in C(S_0 \times S_1) \mapsto \int h(x, y)d\mu_1(y) \in C(S_0)$. The claimed expression for the first derivative then follows by the chain rule. The derivatives of this map can be computed to arbitrary order inductively by the prior argument. □

Proof of Lemma 33. Observe that the map $(\boldsymbol{A}, \varphi_0, \varphi_1) \in \mathbb{R}^{d_0\times d_1} \times \mathfrak{E} \mapsto \varphi_0 \oplus \varphi_1 - c_{\boldsymbol{A}} \in C(S_0 \times S_1)$ is smooth with first derivative at $(\boldsymbol{A}, \varphi_0, \varphi_1) \in \mathbb{R}^{d_0\times d_1} \times \mathfrak{E}$ given by

$$(\boldsymbol{B}, h_0, h_1) \in \mathbb{R}^{d_0\times d_1} \times \mathfrak{E} \mapsto h_0 \oplus h_1 + 32x^\intercal \boldsymbol{B} y \in C(S_0 \times S_1).$$

The result then follows from Lemma 72 by applying the chain rule. □

C.4.2 Compactness of $\mathcal{L}$

Lemma 73 (Example 2 in [220])**.** Let $\varepsilon > 0$, $\mu_0 \in \mathcal{P}(\mathbb{R}^{d_0})$, $\mu_1 \in \mathcal{P}(\mathbb{R}^{d_1})$, and $\boldsymbol{A} \in \mathbb{R}^{d_0\times d_1}$ be arbitrary and let $(\varphi_0^{\boldsymbol{A}}, \varphi_1^{\boldsymbol{A}})$ be EOT potentials for $\mathsf{OT}_{\boldsymbol{A},\varepsilon}(\mu_0, \mu_1)$. Then, the map $\mathcal{L}$: $L^2(\mu_0) \times L^2(\mu_1) \mapsto L^2(\mu_0) \times L^2(\mu_1)$ defined by

$$\mathcal{L}(f_0, f_1) = \left(\int f_1(y)e^{\frac{\varphi_0^{\boldsymbol{A}}(x)+\varphi_1^{\boldsymbol{A}}(y)-c_{\boldsymbol{A}}(x,y)}{\varepsilon}}d\mu_1(y), \int f_0(x)e^{\frac{\varphi_0^{\boldsymbol{A}}(x)+\varphi_1^{\boldsymbol{A}}(y)-c_{\boldsymbol{A}}(x,y)}{\varepsilon}}d\mu_0(x)\right),$$

is compact.

Proof. For simplicity, we prove only that

$$\mathcal{L}_2 : f \in L^2(\mu_0) \mapsto \int f(x)\xi(x,\cdot)d\mu_0(x) \in L^2(\mu_1),$$

is a compact operator for $\xi : (x,y) \in \mathbb{R}^{d_0} \times \mathbb{R}^{d_1} \mapsto e^{\frac{\varphi_0^A(x)+\varphi_1^A(y)-c_A(x,y)}{\varepsilon}}$. For any $y \in \mathbb{R}^{d_1}$ and $f \in L^2(\mu_0)$, $|\mathcal{L}_2(f)(y)|^2 \leq \|f\|^2_{L^2(\mu_0)} \int |\xi(\cdot,y)|^2 d\mu_0$, as $\xi(\cdot,y)$ is bounded on $\mathrm{spt}(\mu_0)$ this operator is well-defined.

Let f_n be a bounded sequence in $L^2(\mu_0)$. By the Eberlein-Šmulian theorem [220, p. 141], up to passing to a subsequence, f_n converges weakly to $f \in L^2(\mu_0)$. For fixed $y \in \mathbb{R}^{d_1}$, $\xi(\cdot,y) \in L^2(\mu_0)$, hence $\mathcal{L}_2(f_n)(y) \to \mathcal{L}_2(f)(y)$ and it follows from the dominated convergence theorem that, for any $g \in L^2(\mu_1)$, $\int \mathcal{L}_2(f_n)g d\mu_1 \to \int \mathcal{L}_2(f)g d\mu_1$ such that $\mathcal{L}_2(f_n) \to \mathcal{L}_2(f)$ weakly in $L^2(\mu_1)$. Also, by dominated convergence,

$$\|\mathcal{L}_2(f_n)\|^2_{L^2(\mu_1)} = \int \mathcal{L}_2(f_n)^2 d\mu_1 \to \int \mathcal{L}_2(f)^2 d\mu_1 = \|\mathcal{L}_2(f)\|^2_{L^2(\mu_1)},$$

such that $\mathcal{L}_2(f_n) \to \mathcal{L}_2(f)$ strongly in $L^2(\mu_1)$. As f_n was an arbitrary bounded sequence in $L^2(\mu_0)$ and $\mathcal{L}_2(f_n) \to \mathcal{L}_2(f)$ strongly in $L^2(\mu_1)$ up to a subsequence, $\mathcal{L}_2$ is a compact operator. □

C.5 The Inner Product Cost

In this section, we summarize how the previous theory adapts when the inner product cost is used in place of the quadratic cost. Explicitly, the inner product EGW problem between $\mu_0 \in \mathcal{P}(\mathbb{R}^{d_0}), \mu_1 \in \mathcal{P}(\mathbb{R}^{d_1})$ is given by

$$\mathsf{F}_\varepsilon(\mu_0,\mu_1) = \inf_{\pi\in\Pi(\mu_0,\mu_1)} \int |\langle x,x'\rangle - \langle y,y'\rangle|^2 \, d\pi\otimes\pi(x,y,x',y') + \varepsilon\mathsf{D}_{\mathsf{KL}}(\pi\|\mu_0\otimes\mu_1).$$

Of note is that F_ε is invariant to orthogonal transformations, but not to translations. Following the arguments of [223], it is straightforward to show that

$$\mathsf{F}_\varepsilon(\mu_0,\mu_1) = \mathsf{F}^1(\mu_0,\mu_1) + \mathsf{F}^2_\varepsilon(\mu_0,\mu_1), \tag{C.8}$$

where

$$\begin{aligned}
\mathsf{F}^1(\mu_0,\mu_1) &= \int |\langle x,x'\rangle|^2 d\mu_0\otimes\mu_0(x,x') + \int |\langle y,y'\rangle|^2 d\mu_1\otimes\mu_1(y,y'),\\
\mathsf{F}^2_\varepsilon(\mu_0,\mu_1) &= \inf_{\pi\in\Pi(\mu_0,\mu_1)} -2\sum_{\substack{1\le i\le d_0\\ 1\le j\le d_1}}\left(\int x_i y_j\, d\pi(x,y)\right)^2 + \varepsilon\mathsf{D}_{\mathsf{KL}}(\pi\|\mu_0\otimes\mu_1),\\
&= \inf_{\mathbf{A}\in\mathcal{D}_M} 8\|\mathbf{A}\|_F^2 + \mathsf{IOT}_{\mathbf{A},\varepsilon}(\mu_0,\mu_1),
\end{aligned}$$

where $\mathcal{D}_M = \{\mathbf{A}\in\mathbb{R}^{d_0\times d_1} : \|\mathbf{A}\|_F \le M/2\}$ for any $M \ge \sqrt{M_2(\mu_0)M_2(\mu_1)}$ and $\mathsf{IOT}_{\mathbf{A},\varepsilon}(\mu_0,\mu_1)$ is the EOT problem with cost function $c_{\mathbf{A}} : (x,y)\in\mathbb{R}^{d_0}\times\mathbb{R}^{d_1} \mapsto -8x^\intercal\mathbf{A}y$. We remark that, in contrast to the quadratic case, μ_0 and μ_1 need not be centered for this decomposition to hold. Given the similarities between the variational forms for the quadratic and inner products, all of the results in the main text hold with only minor modifications to the constants and replacing $\mathsf{OT}_{\mathbf{A},\varepsilon}$ by $\mathsf{IOT}_{\mathbf{A},\varepsilon}$.

Corollary 10 (Stability of Inner Product EGW)**.** Assume that $\mu_0\in\mathcal{P}(\mathbb{R}^{d_0}), \mu_1\in\mathcal{P}(\mathbb{R}^{d_1})$ are compactly supported and define the map $\Psi : \mathbf{A}\in\mathbb{R}^{d_0\times d_1} \mapsto 8\|\mathbf{A}\|_F^2 + \mathsf{IOT}_{\mathbf{A},\varepsilon}(\mu_0,\mu_1)$. Then,

1. Ψ is smooth and coercive, and $D\Phi_{[\mathbf{A}]}(\mathbf{B}) = 16\mathrm{tr}(\mathbf{A}^\intercal\mathbf{B}) - 8\int x^\intercal\mathbf{B}y d\pi_{\mathbf{A}}(x,y)$, where $\pi_{\mathbf{A}}$ is the unique EOT coupling for $\mathsf{IOT}_{\mathbf{A},\varepsilon}(\mu_0,\mu_1)$. As such, all global minima of Ψ are critical points, i.e. satisfy $\frac{1}{2}\int xy^\intercal d\pi_{\mathbf{A}}(x,y) = \mathbf{A}\in\mathcal{D}_M$.
2. Ψ is weakly convex with parameter at most $8^2\varepsilon^{-1}\sqrt{M_4(\mu_0)M_4(\mu_1)} - 16$ and, if $\varepsilon/4 > \sqrt{M_4(\mu_0)M_4(\mu_1)}$, it is strictly convex and admits a unique minimizer.
3. Ψ is L-smooth on $\mathcal{D}_M$ with $L \le 16 \vee \left(8^2\varepsilon^{-1}\sqrt{M_4(\mu_0)M_4(\mu_1)} - 16\right)$.

4. $\mathbf{A}$ minimizes Ψ if and only if $\pi_{\mathbf{A}}$ is optimal for F_ε and satisfies $\mathbf{A} = \frac{1}{2}\int xy^\intercal d\pi_{\mathbf{A}}(x,y)$.

Using this result, Algorithm 5 and Algorithm 6 can be used to minimize Ψ by substituting $\tilde{D}\Phi$ by $\tilde{D}\Psi$ and setting the step size according to the value of L provided in the corollary.

We conclude this section with a numerical experiment. Namely, we empirically validate Theorem 2 in [223][2], which states that if $X_{0,1},\dots,X_{0,n}$ and $X_{1,1},\dots,X_{1,n}$ are independent and identically distributed samples from μ_0 and μ_1, respectively, and $\hat{\mu}_{0,n} \coloneqq \frac{1}{n}\sum_{i=1}^n \delta_{X_{0,i}}$, $\hat{\mu}_{1,n} \coloneqq \frac{1}{n}\sum_{i=1}^n \delta_{X_{1,i}}$ are the corresponding empirical distributions, then

$$\mathbb{E}\left[|\mathsf{F}_\varepsilon(\mu_0,\mu_1) - \mathsf{F}_\varepsilon(\hat{\mu}_{0,n},\hat{\mu}_{1,n})|\right] \le C_{d_0,d_1,\sigma,\varepsilon} n^{-1/2},$$

where $C_{d_0,d_1,\sigma,\varepsilon}$ is a constant which depends only on d_0, d_1, ε and σ, where σ is a shared sub-Gaussian constant of μ_0 and μ_1 (i.e. the smallest constant for which $\mathbb{E}_{\mu_0}[e^{\|X\|^2/2\sigma^2}] \vee \mathbb{E}_{\mu_1}[e^{\|X\|^2/2\sigma^2}] \le 2$) which is assumed to be finite.

To verify this result, we set $\mu_0 = N(0, \mathrm{diag}(0.1, 0.2, 0.3, 0.4))$, the mean zero normal distribution on $\mathbb{R}^4$ with covariance $\mathrm{diag}(0.1, 0.2, 0.3, 0.4)$, and $\mu_1 = N(0, \mathrm{diag}(0.1, 0.2, 0.3, 0.4, 0.5))$ defined similarly. To estimate $\mathsf{F}_\varepsilon(\hat{\mu}_{0,n},\hat{\mu}_{1,n})$, we draw n samples from μ_0 and μ_1, construct the empirical measures, and use Algorithm 5 to minimize the corresponding function Ψ. To guarantee that each such subproblem is strictly convex (so that Algorithm 5 is guaranteed to converge to an approximate minimizer), we set $\varepsilon = 12$ and numerically verify that this choice of ε satisfies the condition of point 2 in Corollary 10. The value of the population level quantity, $\mathsf{F}_\varepsilon(\mu_0,\mu_1)$, is computed to be 0.85 by applying Theorem 3.1 in [127] which, to our knowledge, is the only nontrivial case where an explicit formula has been derived for the EGW distance. To estimate $\mathbb{E}\left[|\mathsf{F}_\varepsilon(\mu_0,\mu_1) - \mathsf{F}_\varepsilon(\hat{\mu}_{0,n},\hat{\mu}_{1,n})|\right]$ for a fixed number of samples, we recompute $\mathsf{F}_\varepsilon(\hat{\mu}_{0,n},\hat{\mu}_{1,n})$

[2]Though the cited result holds for the quadratic cost, it can easily be adapted to the inner product cost.

based on n new samples from μ_0 and μ_1, and repeat this procedure 2000 times to estimate the average value. The results of this experiment are presented in Figure C.1.

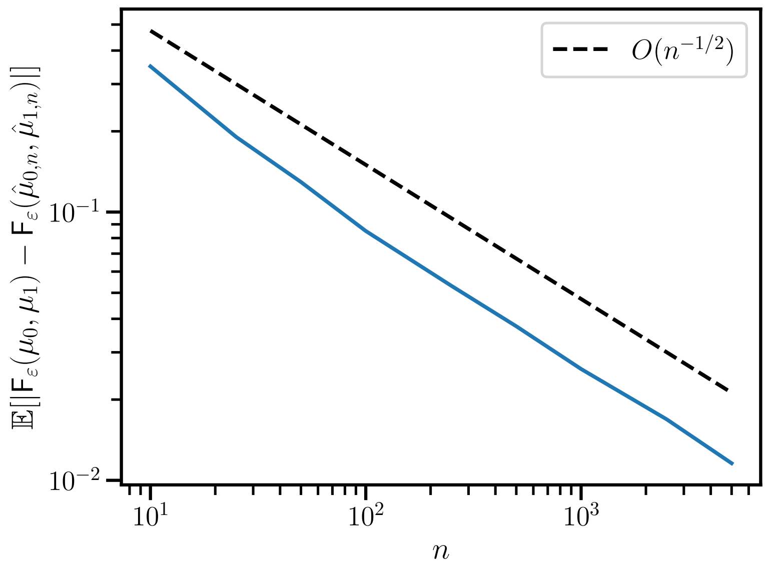


Figure C.1: We plot the expected gap between $\mathsf{F}_\varepsilon(\mu_0, \mu_1)$ and its empirical analogue as a function of the number of samples, $n \in \{10, 25, 50, 100, 250, 500, 1000, 2500, 5000\}$, based on 2000 repetitions on a log-log scale. It is seen that this quantity scales at the theoretical rate of $O(n^{-1/2})$.

BIBLIOGRAPHY

[1] Robert A. Adams and John J.F. Fournier. *Sobolev spaces*. Elsevier, 2003.

[2] Yonathan Aflalo, Alexander Bronstein, and Ron Kimmel. On convex relaxation of graph isomorphism. *Proceedings of the National Academy of Sciences*, 112(10):2942–2947, 2015.

[3] Charalambos D Aliprantis and Kim C Border. *Infinite Dimensional Analysis: A Hitchhiker's Guide*. Springer Science & Business Media, 2006.

[4] Jason Altschuler, Jonathan Weed, and Philippe Rigollet. Near-linear time approximation algorithms for optimal transport via Sinkhorn iteration. In *Proceedings of the 31st Annual Conference on Neural Information Processing Systems*, pages 1961–1971, 2017.

[5] David Alvarez-Melis and Tommi Jaakkola. Gromov-Wasserstein alignment of word embedding spaces. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1881–1890. Association for Computational Linguistics, 2018.

[6] Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient Flows: in Metric Spaces and in the Space of Probability Measures*. Springer Science & Business Media, 2008.

[7] Donald W.K. Andrews and David Pollard. An introduction to functional central limit theorems for dependent stochastic processes. *International Statistical Review/Revue Internationale de Statistique*, pages 119–132, 1994.

[8] Tom M. Apostol. *Mathematical analysis*. Addison-Wesley, 5 edition, 1974.

[9] Miguel Angel Arcones and Bin Yu. Central limit theorems for empirical and U-processes of stationary mixing sequences. *Journal of Theoretical Probability*, 7(1):47–71, 1994.

[10] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *Proceedings of the 34th International Conference on Machine Learning*, pages 214–223, 2017.

[11] Jongsig Bae and Shlomo Levental. Uniform CLT for Markov chains and its invariance principle: a martingale approach. *Journal of Theoretical Probability*, 8(3):549–570, 1995.

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

[13] Stephen R. Becker, Emmanuel J. Candès, and Michael C. Grant. Templates for convex cone problems with applications to sparse signal recovery. *Mathematical Programming Computation*, 3:165–218, 2011.

[14] 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.

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

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

[17] Peter J. Bickel, Friedrich. Götze, and Willem R. van Zwet. Resampling fewer than n observations: Gains, losses, and remedies for losses. In Sara van de Geer and Marten Wegkamp, editors, *Selected Works of Willem van Zwet*, pages 267–297. Springer New York, 2012.

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

[19] Garrett Birkhoff. Extensions of Jentzsch's theorem. *Transactions of the American Mathematical Society*, 85(1):219–227, 1957.

[20] Jose Blanchet and Karthyek Murthy. Quantifying distributional model risk via optimal transport. *Mathematics of Operations Research*, 44(2):565–600, 2019.

[21] Adam Block, Zeyu Jia, Yury Polyanskiy, and Alexander Rakhlin. Intrinsic dimension estimation. *arXiv preprint arXiv:2106.04018*, 2021.

[22] Adam Block, Zeyu Jia, Yury Polyanskiy, and Alexander Rakhlin. Rate of convergence of the smoothed empirical Wasserstein distance. *arXiv preprint arXiv:2205.02128*, 2022.

[23] Andrew J. Blumberg, Mathieu Carriere, Michael A. Mandell, Raul Rabadan,

and Soledad Villar. MREC: a fast and versatile framework for aligning and matching point clouds with applications to single cell molecular data. *arXiv preprint arXiv:2001.01666*, 2020.

[24] Sergey G. Bobkov. Isoperimetric and analytic inequalities for log-concave probability measures. *The Annals of Probability*, 27(4):1903–1921, 1999.

[25] Sergey G. Bobkov and Michel Ledoux. *One-dimensional empirical measures, order statistics, and Kantorovich transport distances*. American Mathematical Society, 2019.

[26] Vladimir I Bogachev and Oleg G Smolyanov. *Real and functional analysis*. Springer Science & Business Media, 2020.

[27] J. Frédéric Bonnans and Alexander Shapiro. *Perturbation analysis of optimization problems*. Springer Science & Business Media, 2013.

[28] Nicolas Bonnotte. *Unidimensional and evolution methods for optimal transportation*. PhD thesis, Paris 11, 2013.

[29] Andrea Braides. *Local minimization, variational evolution and* Γ*-convergence*. Springer, 2014.

[30] Haim Brézis. *Functional analysis, Sobolev spaces and partial differential equations*. Springer Science & Business Media, 2011.

[31] Peter Bühlmann. The blockwise bootstrap for general empirical processes of stationary sequences. *Stochastic Processes and their Applications*, 58(2):247–265, 1995.

[32] Charlotte Bunne, David Alvarez-Melis, Andreas Krause, and Stefanie Jegelka. Learning generative models across incomparable spaces. In *Proceedings of the 36th International Conference on Machine Learning*, pages 851–861, 2019.

[33] Tomasz Byczkowski. Gaussian measures on L_p spaces $0 \leq p < \infty$. *Studia Mathematica*, 59:249–261, 1977.

[34] Kai Cao, Yiguang Hong, and Lin Wan. Manifold alignment for heterogeneous single-cell multi-omics data integration using pamona. *Bioinformatics*, 38(1):211–219, 2022.

[35] Javier Cárcamo, Antonio Cuevas, and Luis-Alberto Rodríguez. Directional differentiability for supremum-type functionals: statistical applications. *Bernoulli*, 26(3):2143–2175, 2020.

[36] Guillaume Carlier, Victor Chernozhukov, and Alfred Galichon. Vector quantile regression: an optimal transport approach. *The Annals of Statistics*, 44(3):1165–1192, 2016.

[37] Guillaume Carlier, Vincent Duval, Gabriel Peyré, and Bernhard Schmitzer. Convergence of entropic schemes for optimal transport and gradient flows. *SIAM Journal on Mathematical Analysis*, 49(2):1385–1418, 2017.

[38] Guillaume Carlier and Maxime Laborde. A differential approach to the multi-marginal Schrödinger system. *SIAM Journal on Mathematical Analysis*, 52(1):709–717, 2020.

[39] Sayak Chatterjee, Dibyendu Saha, Soham Dan, and Bhaswar B. Bhattacharya. Two-sample tests for inhomogeneous random graphs in L_r norm: Optimality and asymptotics. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics*, pages 6903–6911, 2023.

[40] Hong-Bin Chen and Jonathan Niles-Weed. Asymptotics of smoothed wasserstein distances. *Potential Analysis*, pages 1–25, 2021.

[41] Junyu Chen, Binh Nguyen, Shang Hui Koh, and Yong Sheng Soh. Semidefinite relaxations of the Gromov-Wasserstein distance. In *Proceedings of the 38th Annual Conference on Neural Information Processing Systems*, 2024.

[42] Li Chen, Jie Zhou, and Lizhen Lin. Hypothesis testing for populations of networks. *Communications in Statistics-Theory and Methods*, 52(11):3661–3684, 2023.

[43] Liqun Chen, Zhe Gan, Yu Cheng, Linjie Li, Lawrence Carin, and Jingjing Liu. Graph optimal transport for cross-domain alignment. In *Proceedings of the 37th International Conference on Machine Learning*, pages 1542–1553, 2020.

[44] Shuo Chen, Bin Shi, and Ya-xiang Yuan. Gradient norm minimization of Nesterov acceleration: $o(1/k^3)$. *arXiv preprint arXiv:2209.08862*, 2022.

[45] Yao Chen, Qingyi Gao, and Xiao Wang. Inferential Wasserstein generative adversarial networks. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), to appear*, 2021.

[46] Victor Chernozhukov, Alfred Galichon, Marc Hallin, and Marc Henry. Monge–kantorovich depth, quantiles, ranks and signs. *The Annals of Statistics*, 45(1):223–256, 2017.

[47] Victor Chernozhukov, Sokbae Lee, and Adam M. Rosen. Intersection bounds: estimation and inference. *Econometrica*, 81(2):667–737, 2013.

[48] Francis H. Clarke, Yuri S. Ledyaev, Ronald J. Stern, and Peter R. Wolenski. *Nonsmooth analysis and control theory*. Springer Science & Business Media, 2008.

[49] Frank H. Clarke. Generalized gradients and applications. *Transactions of the American Mathematical Society*, 205:247–262, 1975.

[50] Frank H. Clarke. *Optimization and nonsmooth analysis*. SIAM, 1990.

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

[52] Nicolas Courty, Rémi Flamary, Devis Tuia, and Alain Rakotomamonjy. Optimal transport for domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(9):1853–1865, 2016.

[53] Marco Cuturi. Sinkhorn distances: lightspeed computation of optimal transport. In *Proceedings of the 26th Annual Conference on Neural Information Processing Systems*, pages 2292–2300, 2013.

[54] George B. Dantzig. Maximization of a linear function of variables subject to linear inequalities. *Activity analysis of production and allocation*, 13:339–347, 1951.

[55] Alexandre d'Aspremont. Smooth optimization with approximate gradient. *SIAM Journal on Optimization*, 19(3):1171–1183, 2008.

[56] Laurent Davezies, Xavier d'Haultfœuille, and Yannick Guyonvarch. Empirical process results for exchangeable arrays. *The Annals of Statistics*, 49(2):845–862, 2021.

[57] Yu A. Davydov, Mikhail A. Lifschits, and N. V. Smorodina. *Local Properties of Distributions of Stochastic Functionals*. Translation of Mathematical Monographs. American Mathematical Society, 1998.

[58] Guido De Philippis and Alessio Figalli. The monge–ampère equation and its link to optimal transportation. *Bulletin of the American Mathematical Society*, 51(4):527–580, 2014.

[59] Eustasio del Barrio, Evarist Giné, and Carlos Matrán. Central limit theorems for the Wasserstein distance between the empirical and the true distributions. *The Annals of Probability*, 27(2):1009–1071, 1999.

[60] Eustasio del Barrio, Evarist Giné, and Frederic Utzet. Asymptotics for L_2 functionals of the empirical quantile process, with applications to tests of fit based on weighted Wasserstein distances. *Bernoulli*, 11(1):131–189, 2005.

[61] Eustasio del Barrio, Alberto González-Sanz, and Jean-Michel Loubes. Central limit theorems for general transportation costs. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 60(2):847 – 873, 2024.

[62] Eustasio Del Barrio, Alberto González Sanz, and Jean-Michel Loubes. Central limit theorems for semi-discrete Wasserstein distances. *Bernoulli*, 30(1):554–580, 2024.

[63] Eustasio del Barrio, Paula Gordaliza, and Jean-Michel Loubes. A central limit theorem for L^p transportation cost on the real line with application to fairness assessment in machine learning. *Information and Inference: A Journal of the IMA*, 8(4):817–849, 2019.

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

[65] Eustasio del Barrio, Alberto González Sanz, Jean-Michel Loubes, and Jonathan Niles-Weed. An improved central limit theorem and fast convergence rates for entropic transportation costs. *SIAM Journal on Mathematics of Data Science*, 5(3):639–669, 2023.

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

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

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

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

[70] Olivier Devolder, François Glineur, and Yurii Nesterov. First-order methods of smooth convex optimization with inexact oracle. *Mathematical Programming*, 146:37–75, 2014.

[71] Paul Doukhan, Pascal Massart, and Emmanuel Rio. Invariance principles for absolutely regular empirical processes. *Annales de l'IHP Probabilités et statistiques*, 31(2):393–427, 1995.

[72] Richard M. Dudley. *Real Analysis and Probability*. Cambridge University Press, 2002.

[73] Richard M. Dudley. *Uniform central limit theorems*. Cambridge university press, 2014.

[74] Lutz Dümbgen. On nondifferentiable functions and the bootstrap. *Probability Theory and Related Fields*, 95:125–140, 1993.

[75] 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.

[76] Pavel Dvurechensky. Gradient method with inexact oracle for composite nonconvex optimization. *arXiv preprint arXiv:1703.09180*, 2017.

[77] Pavel Dvurechensky, Alexander Gasnikov, and Alexey Kroshnin. Computational optimal transport: complexity by accelerated gradient descent is better than by Sinkhorn's algorithm. In *Proceedings of the 37th International Conference on Machine Learning*, pages 1367–1376, 2018.

[78] Stephan Eckstein and Marcel Nutz. Quantitative stability of regularized optimal transport and convergence of Sinkhorn's algorithm. *SIAM Journal on Mathematical Analysis*, 54(6):5922–5948, 2022.

[79] Zheng Fang and Andres Santos. Inference on directionally differentiable functions. *The Review of Economic Studies*, 86:377–412, 2019.

[80] Herbert Federer. *Geometric measure theory*. Springer Science & Business Media, 2014.

[81] Jean Feydy, Thibault Séjourné, François-Xavier Vialard, Shun-ichi Amari, Alain Trouvé, and Gabriel Peyré. Interpolating between optimal transport and MMD using Sinkhorn divergences. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*, pages 2681–2690. PMLR, 2019.

[82] 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.

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

[84] Joel Franklin and Jens Lorenz. On the scaling of multidimensional matrices. *Linear Algebra and its applications*, 114:717–735, 1989.

[85] Tomas Gal and Harvey J Greenberg. *Advances in sensitivity analysis and parametric programming*. Springer Science & Business Media, 2012.

[86] Jared Lee Gearhart, Kristin Lynn Adair, Justin David Durfee, Katherine A Jones, Nathaniel Martin, and Richard Joseph Detry. Comparison of open-source linear programming solvers. Technical report, Sandia National Lab.(SNL-NM), Albuquerque, NM (United States), 2013.

[87] Aude Genevay, Lénaic Chizat, Francis Bach, Marco Cuturi, and Gabriel Peyré. Sample complexity of Sinkhorn divergences. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*, pages 1574–1583, 2019.

[88] Aude Genevay, Gabriel Peyré, and Marco Cuturi. Learning generative models with Sinkhorn divergences. In *Proceedings of the 21st International Conference on Artificial Intelligence and Statistics*, pages 1608–1617, 2018.

[89] Saeed Ghadimi and Guanghui Lan. Accelerated gradient methods for nonconvex nonlinear and stochastic programming. *Mathematical Programming*, 156(1-2):59–99, 2016.

[90] Promit Ghosal, Marcel Nutz, and Espen Bernton. Stability of entropic optimal transport and Schrödinger bridges. *Journal of Functional Analysis*, 283(9):109622, 2022.

[91] Promit Ghosal and Bodhisattva Sen. Multivariate ranks and quantiles using optimal transport: Consistency, rates and nonparametric testing. *The Annals of Statistics*, 50(2):1012 – 1037, 2022.

[92] Debarghya Ghoshdastidar, Maurilio Gutzeit, Alexandra Carpentier, and Ulrike von Luxburg. Two-sample tests for large random graphs using network statistics. In *Proceedings of the 2017 Conference on Learning Theory*, pages 954–977. PMLR, 2017.

[93] Debarghya Ghoshdastidar, Maurilio Gutzeit, Alexandra Carpentier, and Ulrike Von Luxburg. Two-sample hypothesis testing for inhomogeneous random graphs. *The Annals of Statistics*, 48(4):2208–2229, 2020.

[94] Debarghya Ghoshdastidar and Ulrike Von Luxburg. Practical methods for graph two-sample testing. In *Proceedings of the 32nd Annual Conference on Neural Information Processing Systems*, 2018.

[95] David Gilbarg and Neil S. Trudinger. *Elliptic partial differential equations of second order*, volume 224. Springer Science & Business Media, 2015.

[96] Evarist Giné and Richard Nickl. *Mathematical foundations of infinite-dimensional statistical models*. Cambridge University Press, 2021.

[97] Evarist Giné and Joel Zinn. Bootstrapping general empirical measures. *The Annals of Probability*, pages 851–869, 1990.

[98] Cedric E. Ginestet, Jun Li, Prakash Balachandran, Steven Rosenberg, and Eric D. Kolaczyk. Hypothesis testing for network data in functional neuroimaging. *The Annals of Applied Statistics*, pages 725–750, 2017.

[99] María Jesús Gisbert, María Josefa Cánovas, Juan Parra, and Fco. Javier Toledo. Lipschitz modulus of the optimal value in linear programming. *Journal of Optimization Theory and Applications*, 182(1):133–152, Jul 2019.

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

[101] Ziv Goldfeld, Kristjan Greenewald, and Kengo Kato. Asymptotic guarantees for generative modeling based on the smooth Wasserstein distance. In *Proceedings of the 34th annual Conference on Neural Information Processing Systems*, pages 2527–2539, 2020.

[102] Ziv Goldfeld, Kristjan Greenewald, Jonathan Niles-Weed, and Yury Polyanskiy. Convergence of smoothed empirical measures with applications to entropy estimation. *IEEE Transactions on Information Theory*, 66(7):4368–4391, 2020.

[103] Ziv Goldfeld and Kengo Kato. Limit distribution for smooth total variation and χ^2-divergence in high dimensions. In *Proceedings of the 2020 IEEE International Symposium on Information Theory*, 2020.

[104] Ziv Goldfeld, Kengo Kato, Sloan Nietert, and Gabriel Rioux. Limit distribution theory for smooth p-Wasserstein distances. *The Annals of Applied Probability*, 34(2):2447 – 2487, 2024.

[105] Ziv Goldfeld, Kengo Kato, Gabriel Rioux, and Ritwik Sadhu. Limit theorems for entropic optimal transport maps and Sinkhorn divergence. *Electronic Journal of Statistics*, 18(1):980–1041, 2024.

[106] 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.

[107] Paul J. Goulart and Yuwen Chen. Clarabel: An interior-point solver for conic programs with quadratic objectives. *arXiv preprint arXiv:2405.12762*, 2024.

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

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

[110] Michel Groppe and Shayan Hundrieser. Lower complexity adaptation for empirical

entropic optimal transport. *Journal of Machine Learning Research*, 25(344):1–55, 2024.

[111] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron Courville. Improved training of Wasserstein GANs. In *Proceedings of the International Conference on Neural Information Processing Systems (NeurIPS-2017)*, pages 5769–5779, 2017.

[112] Adityanand Guntuboyina and Bodhisattva Sen. L1 covering numbers for uniformly bounded convex functions. In *Proceedings of the 2012 Conference on Learning Theory*, 2012.

[113] Marc Hallin, Eustasio del Barrio, Juan Cuesta-Albertos, and Carlos Matrán. Distribution and quantile functions, ranks and signs in dimension d: A measure transportation approach. *The Annals of Statistics*, 49(2):1139 – 1165, 2021.

[114] Fang Han, Zhen Miao, and Yandi Shen. Nonparametric mixture MLEs under Gaussian-smoothed optimal transport distance. *IEEE Transactions on Information Theory*, 69(12):7823–7835, 2023.

[115] Jean-Baptiste Hiriart-Urruty and Claude Lemaréchal. *Fundamentals of convex analysis*. Springer Science & Business Media, 2004.

[116] Qi Huangfu and J.A. Julian Hall. Parallelizing the dual revised simplex method. *Mathematical Programming Computation*, 10(1):119–142, 2018.

[117] Shayan Hundrieser, Marcel Klatt, Axel Munk, and Thomas Staudt. A unifying approach to distributional limits for empirical optimal transport. *Bernoulli*, 30(4):2846 – 2877, 2024.

[118] Lars Hörmander. *The analysis of linear partial differential operators I: Distribution theory and Fourier analysis*. Springer Science & Business Media, 2003.

[119] Jiashun Jin, Zheng Tracy Ke, Shengming Luo, and Yucong Ma. Optimal network pairwise comparison. *Journal of the American Statistical Association*, pages 1–25, 2024.

[120] Leonid V. Kantorovich. On the translocation of masses. In *Doklady Akademii Nauk USSR*, volume 37, pages 199–201, 1942.

[121] Narendra Karmarkar. A new polynomial-time algorithm for linear programming.

In *Proceedings of the 16th annual ACM symposium on Theory of computing*, pages 302–311, 1984.

[122] Jun Kitagawa, Quentin Mérigot, and Boris Thibert. Convergence of a Newton algorithm for semi-discrete optimal transport. *Journal of the European Mathematical Society*, 21(9):2603–2651, 2019.

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

[124] Diethard Klatte. On the Lipschitz behavior of optimal solutions in parametric problems of quadratic optimization and linear complementarity. *Optimization*, 16(6):819–831, 1985.

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

[126] Soheil Kolouri, Phillip E. Pope, Charles E. Martin, and Gustavo K. Rohde. Sliced Wasserstein auto-encoders. In *Proceedings of the 7th International Conference on Learning Representations*, 2019.

[127] Khang Le, Dung Q. Le, Huy Nguyen, Dat Do, Tung Pham, and Nhat Ho. Entropic Gromov-Wasserstein between Gaussian distributions. In *Proceedings of the 39th International Conference on Machine Learning*, pages 12164–12203, 2022.

[128] Nam Q. Le. The eigenvalue problem for the Monge-Ampère operator on general bounded convex domains. *arXiv preprint arXiv:1701.05165*, 2017.

[129] Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces: isoperimetry and processes*. Springer Science & Business Media, 1991.

[130] Christian Léonard. A survey of the Schrödinger problem and some of its connections with optimal transport. *Discrete & Continuous Dynamical Systems*, 34(4):1533, 2014.

[131] Shlomo Levental. Uniform limit theorems for Harris recurrent Markov chains. *Probability theory and related fields*, 80(1):101–118, 1988.

[132] Peihua Li, Qilong Wang, and Lei Zhang. A novel earth mover's distance methodol-

ogy for image matching with gaussian mixture models. In *Proceedings of the 2013 IEEE International Conference on Computer Vision*, pages 1689–1696, 2013.

[133] Yezheng Li and Hongzhe Li. Two-sample test of community memberships of weighted stochastic block models. *arXiv preprint arXiv:1811.12593*, 2018.

[134] Tianyi Lin, Zeyu Zheng, Elynn Y. Chen, Marco Cuturi, and Michael I. Jordan. On projection robust optimal transport: Sample complexity and model misspecification. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics*, pages 262–270, 2021.

[135] László Lovász and Santosh Vempala. The geometry of logconcave functions and sampling algorithms. *Random Structures & Algorithms*, 30(3):307–358, 2007.

[136] Paul Malliavin. *Integration and probability*. Springer Science & Business Media, 1995.

[137] Tudor Manole, Sivaraman Balakrishnan, Jonathan Niles-Weed, and Larry Wasserman. Plugin estimation of smooth optimal transport maps. *The Annals of Statistics*, 52(3):966 – 998, 2024.

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

[139] Tudor Manole and Jonathan Niles-Weed. Sharp convergence rates for empirical optimal transport with smooth costs. *The Annals of Applied Probability*, 34(1B):1108 – 1135, 2024.

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

[141] Facundo Mémoli. Gromov-Wasserstein distances and the metric approach to object matching. *Foundations of Computational Mathematics*, 11(4):417–487, 2011.

[142] Gonzalo Mena and Jonathan Niles-Weed. Statistical bounds for entropic optimal transport: sample complexity and the central limit theorem. In *Proceedings of the 33rd annual Conference on Neural Information Processing Systems*, 2019.

[143] Toshio Mikami. Regularity of Schrodinger's functional equation and mean field pdes for h-path processes. *Osaka Journal of Mathematics*, 56(4):831–842, 2019.

[144] Toshio Mikami. Regularity of Schrödinger's functional equation in the weak topology and moment measures. *Journal of the Mathematical Society of Japan*, 73(1):99–123, 2021.

[145] Emanuel Milman. On the role of convexity in isoperimetry, spectral gap and concentration. *Inventiones mathematicae*, 177(1):1–43, 2009.

[146] Gaspard Monge. Mémoire sur la théorie des déblais et des remblais. *Histoire de l'Académie Royale des Sciences de Paris*, 1781.

[147] Kimia Nadjahi, Alain Durmus, Lénaïc Chizat, Soheil Kolouri, Shahin Shahrampour, and Umut Simsekli. Statistical and topological properties of sliced probability divergences. pages 20802–20812, 2020.

[148] Kimia Nadjahi, Alain Durmus, Umut Simsekli, and Roland Badeau. Asymptotic guarantees for learning generative models with the sliced-wasserstein distance. In *Proceedings of the 32nd Annual Conference on Neural Information Processing Systems*, 2019.

[149] Francis Narcowich, Joseph Ward, and Holger Wendland. Sobolev bounds on functions with scattered zeros, with applications to radial basis function surface fitting. *Mathematics of Computation*, 74(250):743–763, 2005.

[150] Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*. Springer Science & Business Media, 2003.

[151] Yurii Nesterov. Gradient methods for minimizing composite functions. *Mathematical programming*, 140(1):125–161, 2013.

[152] Sloan Nietert, Ziv Goldfeld, and Kengo Kato. Smooth p-Wasserstein distance: Structure, empirical approximation, and statistical applications. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8172–8183, 2021.

[153] Jonathan Niles-Weed and Philippe Rigollet. Estimation of Wasserstein distances in the Spiked Transport Model. *Bernoulli*, 28(4):2663 – 2688, 2022.

[154] Yoichi Nishiyama. Weak convergence of some classes of martingales with jumps. *The Annals of Probability*, 28(2):685–712, 2000.

[155] Marcel Nutz. Introduction to entropic optimal transport. *Lecture notes, Columbia University*, 2021.

[156] Marcel Nutz and Johannes Wiesel. Entropic optimal transport: Convergence of potentials. *Probability Theory and Related Fields*, pages 1–24, 2021.

[157] Marcel Nutz and Johannes Wiesel. Stability of Schrödinger potentials and convergence of Sinkhorn's algorithm. *The Annals of Probability*, 51(2):699–722, 2023.

[158] Masashi Okamoto. Distinctness of the eigenvalues of a quadratic form in a multivariate sample. *The Annals of Statistics*, 1(4):763 – 765, 1973.

[159] Hermina Petric Maretic, Mireille El Gheche, Giovanni Chierchia, and Pascal Frossard. Got: an optimal transport framework for graph comparison. In *Proceedings of the 33rd Annual Conference in Neural Information Processing Systems*, 2019.

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

[161] Dimitris N Politis and Joseph P Romano. Large sample confidence regions based on subsamples under minimal assumptions. *The Annals of Statistics*, pages 2031–2050, 1994.

[162] Yury Polyanskiy and Yihong Wu. Wasserstein continuity of entropy and outer bounds for interference channels. *IEEE Transactions on Information Theory*, 62(7):3992–4002, 2016.

[163] Julien Rabin, Gabriel Peyré, Julie Delon, and Marc Bernot. Wasserstein barycenter and its application to texture mixing. In *Proceedings of the 3rd International Conference on Scale Space and Variational Methods in Computer Vision*, pages 435–446. Springer, 2011.

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

[165] Gabriel Rioux, Ziv Goldfeld, and Kengo Kato. Limit Laws for Gromov-Wasserstein Alignment with Applications to Testing Graph Isomorphisms. *Submitted to the Annals of Statistics*, 2024. arXiv preprint arXiv:2410.18006.

[166] Gabriel Rioux, Apoorva Nitsure, Mattia Rigotti, Kristjan Greenewald, and Youssef Mroueh. Multivariate stochastic dominance via optimal transport and applications to models benchmarking. In *Proceedings of the Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[167] Ralph Tyrrell Rockafellar. *Convex analysis*. Princeton university press, 1997.

[168] Ralph Tyrrell Rockafellar and Roger J-B Wets. *Variational analysis*. Springer Science & Business Media, 2009.

[169] Werner Römisch. Delta method, infinite dimensional. In *Encyclopedia of Statistical Sciences*. Wiley, 2004.

[170] Yossi Rubner, Carlo Tomasi, and Leonidas J. Guibas. The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision*, 40(2):99–121, 2000.

[171] Alessandro Rudi, Ulysse Marteau-Ferey, and Francis Bach. Finding global minima via kernel approximations. *Mathematical Programming*, 209(1):703–784, 2025.

[172] Ritwik Sadhu, Ziv Goldfeld, and Kengo Kato. Limit distribution theory for the smooth 1-Wasserstein distance with applications. *arXiv preprint arXiv:2107.13494*, 2021.

[173] Hans Samelson. On the Perron-Frobenius theorem. *Michigan Mathematical Journal*, 4(1):57 – 59, 1957.

[174] Roman Sandler and Michael Lindenbaum. Nonnegative matrix factorization with earth mover's distance metric for image analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(8):1590–1602, 2011.

[175] Filippo Santambrogio. *Optimal transport for applied mathematicians*. Springer Science & Business Media, 2015.

[176] Adrien Saumard and Jon A Wellner. Log-concavity and strong log-concavity: a review. *Statistics surveys*, 8:45, 2014.

[177] Meyer Scetbon, Gabriel Peyré, and Marco Cuturi. Linear-time Gromov-Wasserstein distances using low rank couplings and costs. In *Proceedings of the 39th International Conference on Machine Learning*, pages 19347–19365, 2022.

[178] Bernhard Schölkopf, Alexander Smola, and Klaus-Robert Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural computation*, 10(5):1299–1319, 1998.

[179] Erwin Schrödinger. Über die umkehrung der naturgesetze. *Akad. Wiss. Berlin. Phys. Math.*, 144:144–153, 1931.

[180] Thibault Séjourné, François-Xavier Vialard, and Gabriel Peyré. The unbalanced Gromov-Wasserstein distance: conic formulation and relaxation. *Proceedings of the 35th Annual Conference on Neural Information Processing Systems*, 34:8766–8779, 2021.

[181] Alexander Shapiro. On concepts of directional differentiability. *Journal of Optimization Theory and Applications*, 66:477–487, 1990.

[182] Alexander Shapiro. Asymptotic analysis of stochastic programs. *Annals of Operations Research*, 30:169–186, 1991.

[183] Bin Shi, Simon S. Du, Michael I. Jordan, and Weijie J. Su. Understanding the acceleration phenomenon via high-resolution differential equations. *Mathematical Programming*, pages 1–70, 2021.

[184] Richard Sinkhorn. Diagonal equivalence to matrices with prescribed row and column sums. *The American Mathematical Monthly*, 74(4):402–405, 1967.

[185] Justin Solomon, Fernando De Goes, Gabriel Peyré, Marco Cuturi, Adrian Butscher, Andy Nguyen, Tao Du, and Leonidas Guibas. Convolutional Wasserstein distances: Efficient optimal transportation on geometric domains. *ACM Transactions on Graphics*, 34(4):66, 2015.

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

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

[188] Daniel W. Stroock. *Probability theory: an analytic view*. Cambridge University Press, 2010.

[189] Karl-Theodor Sturm. *The space of spaces: curvature bounds and gradient flows*

on the space of metric measure spaces, volume 290. American Mathematical Society, 2023.

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

[191] Minh Tang, Avanti Athreya, Daniel L. Sussman, Vince Lyzinski, Youngser Park, and Carey E. Priebe. A semiparametric two-sample hypothesis testing problem for random graphs. *Journal of Computational and Graphical Statistics*, 26(2):344–354, 2017.

[192] Minh Tang, Avanti Athreya, Daniel L. Sussman, Vince Lyzinski, and Carey E. Priebe. A nonparametric two-sample hypothesis testing problem for random graphs. *Bernoulli*, 23(3):1599 – 1630, 2017.

[193] Ilya Tolstikhin, Olivier Bousquet, Sylvain Gelly, and Bernhard Schölkopf. Wasserstein auto-encoders. In *In Proceedings of the 6th International Conference on Learning Representations*, 2018.

[194] William Torous, Florian Gunsilius, and Philippe Rigollet. An optimal transport approach to causal inference. *arXiv preprint arXiv:2108.05858*, 2021.

[195] Paul Tseng. On accelerated proximal gradient methods for convex-concave optimization. `https://www.mit.edu/~dimitrib/PTseng/papers/apgm.pdf`, 2008.

[196] Adrien Vacher, Boris Muzellec, Francis Bach, François-Xavier Vialard, and Alessandro Rudi. Optimal estimation of smooth transport maps with kernel sos. *SIAM Journal on Mathematics of Data Science*, 6(2):311–342, 2024.

[197] Adrien Vacher, Boris Muzellec, Alessandro Rudi, Francis Bach, and Francois-Xavier Vialard. A dimension-free computational upper-bound for smooth optimal transport estimation. In *Proceedings of the 34th Annual Conference on Learning Theory*, pages 4143–4173, 2021.

[198] Aad W van der Vaart. Efficiency and hadamard differentiability. *Scandinavian Journal of Statistics*, 18(1):63–75, 1991.

[199] Aad W van der Vaart. New Donsker classes. *The Annals of Probability*, 24(4):2128–2124, 1996.

[200] Aad W van der Vaart. *Asymptotic statistics*. Cambridge University Press, 1998.

[201] Aad W. van der Vaart and Jon A. Wellner. *Weak convergence and empirical processes*. Springer Science & Business Media, 1996.

[202] J.H. van Zanten and Aad W van der Vaart. Reproducing kernel Hilbert spaces of Gaussian priors. In *Pushing the limits of contemporary statistics: contributions in honor of Jayanta K. Ghosh*, pages 200–222. Institute of Mathematical Statistics, 2008.

[203] V. S. Varadarajan. On the convergence of sample probability distributions. *Sankhyā: The Indian Journal of Statistics (1933-1960)*, 19(1/2):23–26, Feb 1958.

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

[205] Titouan Vayer, Laetitia Chapel, Rémi Flamary, Romain Tavenard, and Nicolas Courty. Optimal transport for structured data with application on graphs. In *Proceedings of the 38th International Conference on Machine Learning*, 2019.

[206] Titouan Vayer, Rémi Flamary, Nicolas Courty, Romain Tavenard, and Laetitia Chapel. Sliced Gromov-Wasserstein. In *Proceedings of the 32nd Annual Conference on Neural Information Processing Systems*, 2019.

[207] Cédric Villani. *Topics in optimal transportation*. American Mathematical Society, 2003.

[208] Cédric Villani. *Optimal transport: Old and new*. Springer Science & Business Media, 2008.

[209] Cédric Vincent-Cuaz, Rémi Flamary, Marco Corneli, Titouan Vayer, and Nicolas Courty. Semi-relaxed Gromov Wasserstein divergence with applications on graphs. In *Proceedings of the 10th International Conference on Learning Representations*, 2022.

[210] Feng-Yu Wang and Jian Wang. Functional inequalities for convolution probability measures. *Annales de l'Institut Henri Poincaré Probabilités et Statistiques*, 52(2):898–914, 2016.

[211] Tao Wang and Ziv Goldfeld. Neural estimation of entropic optimal transport. In *Proceedings of the 2024 IEEE International Symposium on Information Theory*, pages 2116–2121. IEEE, 2024.

[212] Holger Wendland. *Scattered data approximation*. Cambridge University Press, 2004.

[213] Eric Wong, Frank Schmidt, and Zico Kolter. Wasserstein adversarial examples via projected sinkhorn iterations. In *Proceedings of the 36th International Conference on Machine Learning*, pages 6808–6817, 2019.

[214] Jiaqi Xi and Jonathan Niles-Weed. Distributional convergence of the sliced Wasserstein process. In *Proceedings of the 35th Annual Conference on Neural Information Processing Systems*, pages 13961–13973, 2022.

[215] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

[216] Hongteng Xu, Dixin Luo, and Lawrence Carin. Scalable Gromov-Wasserstein learning for graph partitioning and matching. *Advances in neural information processing systems*, 32, 2019.

[217] Hongteng Xu, Dixin Luo, Hongyuan Zha, and Lawrence Carin. Gromov-Wasserstein learning for graph matching and node embedding. In *Proceedings of the 36th International Conference on Machine Learning*, pages 6932–6941. PMLR, 2019.

[218] 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.

[219] Yuguang Yan, Wen Li, Hanrui Wu, Huaqing Min, Mingkui Tan, and Qingyao Wu. Semi-supervised optimal transport for heterogeneous domain adaptation. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pages 2969–2975, 2018.

[220] Kōsaku Yosida. *Functional analysis*. Springer Science & Business Media, 1995.

[221] Yixing Zhang, Xiuyuan Cheng, and Galen Reeves. Convergence of Gaussian-smoothed optimal transport distance with sub-gamma distributions and dependent samples. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics*, pages 2422–2430, 2021.

[222] Zhengxin Zhang, Ziv Goldfeld, Kristjan Greenewald, Youssef Mroueh, and Bharath K Sriperumbudur. Gradient flows and Riemannian structure in the

Gromov-Wasserstein geometry. *Submitted to the Foundations of Computational Mathematics*, 2024. arXiv preprint arXiv:2407.11800.

[223] 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.

[224] Zhengxin Zhang, Youssef Mroueh, Ziv Goldfeld, and Bharath Sriperumbudur. Cycle consistent probability divergences across different spaces. In *Proceedings of the 39th International Conference on Artificial Intelligence and Statistics*, pages 7257–7285, 2022.