Semi-discrete Gromov-Wasserstein Distances: Existence of Gromov-Monge Maps and Statistical Theory

Gabriel Rioux
Center for Applied Mathematics
Cornell University
Ithaca, NY 14853.
ger84@cornell.edu

Ziv Goldfeld
Department of Electrical and Computer Engineering
Cornell University
Ithaca, NY 14853.
goldfeld@cornell.edu

Kengo Kato
Department of Statistics and Data Science
Cornell University
Ithaca, NY 14853.
kk976@cornell.edu

Abstract

The Gromov-Wasserstein (GW) distance serves as a discrepancy measure between metric measure spaces. Despite recent theoretical developments, its structural properties, such as existence of optimal maps, remain largely unaccounted for. In this work, we analyze the semi-discrete regime for the GW problem wherein one measure is finitely supported. Notably, we derive a primitive condition which guarantees the existence of optimal maps. This condition also enables us to derive the asymptotic distribution of the empirical semi-discrete GW distance under proper centering and scaling. As a complement to this asymptotic result, we also derive expected empirical convergence rates. As is the case with the standard Wasserstein distance, the rate we derive in the semi-discrete GW case, $n^{-1/2}$, is dimension-independent which is in stark contrast to the curse of dimensionality rate obtained in general.

1 Introduction

The Gromov-Wasserstein (GW) distance, introduced by Mémoli in [35], enables a comparison between abstract metric measure (mm) spaces and provides an alignment plan between them. Precisely, for two mm spaces, $(\mathcal{X}, \mathsf{d}_\mathcal{X}, \mu_0)$ and $(\mathcal{Y}, \mathsf{d}_\mathcal{Y}, \mu_1)$, their (p, q)-GW distance is given by

$$\mathsf{D}_{p,q}(\mu_0, \mu_1) = \left(\inf_{\pi \in \Pi(\mu_0, \mu_1)} \iint \left| \mathsf{d}^p_\mathcal{X}(x, x') - \mathsf{d}^p_\mathcal{Y}(y, y') \right|^q d\pi \otimes \pi(x, y, x', y') \right)^{1/q}, \tag{1}$$

where $\Pi(\mu_0, \mu_1)$ denotes the set of all couplings of μ_0, μ_1. The GW distances thus equals the least amount of distance distortion one can achieve between the mm spaces when optimizing over all possible alignments thereof (as modeled by couplings). Remarkably, $\mathsf{D}_{p,q}$ defines a metric on the quotient space of all mm spaces obtained by identifying isomorphic mm spaces (i.e., when the underlying measures μ_0, μ_1 are such that $\mu_1 = f_\sharp \mu_0$ for an isometry $f : \mathcal{X} \to \mathcal{Y}$). The ability of the GW distance to meaningfully compare heterogeneous data has spurred its usage in many applications, including generative modelling [5], graph matching [55], heterogeneous domain adaptation [43, 56], spectral clustering [8], graph classification [48], object matching [34, 35] and shape analysis [19, 46].

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

Despite these virtuous properties, there remains large gaps in our understanding of solutions to (1). For instance, sufficient conditions for the existence of optimal couplings induced by a deterministic map, dubbed Gromov-Monge maps, is still an open question (see Question 2.4 in [36]). Another important drawback of the GW distance is that it suffers from the curse of dimensionality in statistical estimation [57]. This work posits the semi-discrete setting for the GW problem as a natural class of problems for which both of these issues can be addressed.

As a special instance of the general GW problem, the semi-discrete GW problem (SDGW) is obtained when one of the marginals is discretely supported and the other is continuous. This setting falls well within the scope of the GW problem due to the underlying heterogeneity and as it enables a comparison of distributions supported on different spaces. To our knowledge, this paradigm has not been explored in the context of the GW problem despite the interest in standard semi-discrete optimal transport (SDOT). Indeed, the SDOT problem has seen use in diverse applications, ranging from computer graphics [12, 31] and generative modelling [2, 7, 27] to fluid dynamics [13, 22] and cosmology [32]. This interest is due, in large part, to the strong structural and statistical properties inherent to the semi-discrete setting. For instance, the structure of optimal maps for the SDOT problem as well as its stability properties were studied in [3, 28] whereas [15] proves a parametric empirical convergence rate and limit distributions for the empirical SDOT cost (see the subsequent literature review for details). Inspired by these results, the present paper studies the $(2, 2)$-GW problem for marginals supported in Euclidean spaces under the semi-discrete paradigm. Our main contributions are to first establish existence of Gromov-Monge maps for the SDGW problem under primitive conditions and, subsequently, to prove finite sample convergence rates and limit distributions for the empirical distance.

Literature review. In contrast to the standard optimal transport (OT) problem [1, 52, 53], little is known about the structure of solutions to the quadratic GW (QGW) problem, $\mathsf{D}_{2,2}(\mu_0, \mu_1)$, with $(\mu_0, \mu_1) \in \mathcal{P}(\mathbb{R}^{d_0}) \times \mathcal{P}(\mathbb{R}^{d_1})$. Indeed, conditions guaranteeing the existence of optimal plans for the QGW problem which are induced by a map are generally unknown. A first result in this direction is Theorem 9.21 in [47] which proves the existence of optimal maps for absolutely continuous measures which are rotationally invariant about their barycenter. Proposition 4.2.4 in [51] proves such a result under the assumption of compact support, absolute continuity of μ_0, that $d_0 \geq d_1$, and that the cross-correlation matrix of an optimal coupling ($\int xy^\intercal d\pi(x, y)$) is full rank along with an abstract condition on the map. Under the same conditions, Theorem 5 in [18] shows that if the rank of the cross-correlation matrix is at most $d_1 - 2$, an optimal map exists and, otherwise there exists an optimal plan induced by a bi-map (viz. two-way map). To our knowledge, these are the only such results currently available in the literature. Until quite recently, a dual representation for the QGW distance was also unavailable. This issue was addressed in [57], where a connection between the QGW distance and an optimization problem involving a standard optimal transport problem with a cost depending on the decision variables (5). By leveraging this variational formulation, the authors of that work were able to obtain the first sample complexity result for the empirical Gromov-Wasserstein distance, proving dimension-dependent rates which suffer from the curse of dimensionality. This rate is improved as to depend on the lesser of the two ambient dimensions in [26].

We conclude with a brief survey of recent statistical developments for SDOT. From the statistical lens, the empirical SDOT distance converges to its population counterpart in expectation at the rate $\sqrt{N}n^{-1/2}$, where n is the number of samples when the cost is Euclidean [15]. This result contrasts the dimension dependent rate, $n^{-1/d}$, of standard optimal transport with the Euclidean cost for probability measures on $\mathcal{P}(\mathbb{R}^d)$ [21, 54]. [15] equally establishes the asymptotic distribution of empirical SDOT for general non-negative costs (upon centering by the population quantity and scaling by $\sqrt{n}$) and proves sufficient conditions for asymptotic normality. Their work also covers limit distributions for the optimal potential under certain regularity conditions. This result is used in [42] to derive limit distributions for the L^p error ($p \geq 1$) of the empirical SDOT map and for linear functionals thereof.

2 Notation and preliminaries

For a nonempty subset S of a topological space $\mathcal{S}$, $\ell^\infty(S)$ denotes the Banach space of bounded real functions on S equipped with the supremum norm $\|\cdot\|_{\infty,S} = \sup_S |\cdot|$. The closure of $S \subset \mathcal{S}$ is denoted $\overline{S}^{\mathcal{S}}$. We denote by $\mathcal{P}(\mathbb{R}^d)$ the set of all probability measures on $\mathbb{R}^d$. For a measurable

map $T : \mathbb{R}^d \to \mathbb{R}^{d'}$ and $\mu \in \mathcal{P}(\mathbb{R}^d)$, $T_\sharp \mu \in \mathcal{P}(\mathbb{R}^{d'})$ is the pushforward measure defined by $T_\sharp \mu(A) = \mu(T^{-1}(A))$ for every Borel set $A \subset \mathbb{R}^{d'}$. For $\mu \in \mathcal{P}(\mathbb{R}^d)$, $\bar{\mu} = (\mathrm{Id} - \mathbb{E}_\mu[X])_\sharp \mu$ is the centered version of μ. For $p \geq 1$, $M_p(\mu) = \int \|\cdot\|^p d\mu$ is the p-th moment of μ. The support of μ is denoted $\mathrm{spt}(\mu)$. The weak convergence of probability measures is denoted by $\xrightarrow{w}$ and convergence in distribution by $\xrightarrow{d}$ (in the sense of Hoffmann-Jørgensen if necessary).

For matrices $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{d_0 \times d_1}$, $\langle \boldsymbol{A}, \boldsymbol{B} \rangle_F = \mathrm{Tr}\left(\boldsymbol{A}^\intercal \boldsymbol{B}\right)$ is the Frobenius inner product, $\|\cdot\|_F$ is the induced norm, and $B_F(r) = \{\|\cdot\|_F \leq r\}$. For a set $A \subset \mathbb{R}^d$, $\mathrm{conv}(A)$ is the convex hull of A, $\mathrm{lin}(A)$ is the linear hull of A, and, if A is compact, $\|A\|_\infty = \sup_{x \in A} \|x\|$. $\mathbb{N}_0 = \mathbb{N} \cup \{0\}$. For an open set $U \subset \mathbb{R}^d$, $\mathcal{C}^\infty_M(U)$ denotes the set of smooth functions, $f : U \to \mathbb{R}$, for which $\max_{\alpha \in \mathbb{N}_0^d} \|\partial^\alpha f\|_{\infty, U} \leq M$. We adopt the shorthand notation $a \vee b = \max(a, b)$ and $a \wedge b = \min(a, b)$.

2.1 Optimal transport

We now recall some standard results from optimal transport (OT) theory. Fix measurable sets $\mathcal{X} \subset \mathbb{R}^{d_0}$ and $\mathcal{Y} \subset \mathbb{R}^{d_0}$. For $\mu_0 \in \mathcal{P}(\mathcal{X}), \mu_1 \in \mathcal{P}(\mathcal{Y})$ and a continuous cost function $c : \mathcal{X} \times \mathcal{Y} \mapsto \mathbb{R}$ satisfying $\int c d\mu_0 \otimes \mu_1 < \infty$ and $c(x, y) \geq a(x) + b(y)$ for $a \in L^1(\mu_0), b \in L^1(\mu_1)$,

$$\mathsf{OT}_c(\mu_0, \mu_1) := \min_{\pi \in \Pi(\mu_0, \mu_1)} \int c d\pi = \max_{\psi \in L^1(\mu_1)} \left\{ \int \psi^c d\mu_0 + \int \psi d\mu_1 \right\}, \tag{2}$$

where $\psi^c : x \in \mathcal{X} \mapsto \inf_{\mathcal{Y}} \left(c(x, \cdot) - \psi\right)$ is the c-conjugate of ψ (cf. e.g. Theorem 5.10 and Remark 5.14 in [53]). The first problem is dubbed the primal problem whereas the second is called the dual problem and both admit solutions under these conditions. Moreover, solutions to the dual problem can be assumed to be c-concave in the sense that $\psi(y) = \phi^c(y) = \inf_{\mathcal{X}} \left(c(\cdot, y) - \phi\right)$ for some function ϕ. If ψ is c-concave and solves the dual problem, we call (ψ, ψ^c) conjugate potentials for $\mathsf{OT}_c(\mu_0, \mu_1)$. We say that conjugate potentials are unique up to constants if any two pairs of conjugate potentials (ψ_0, ψ_0^c), (ψ_1, ψ_1^c) are such that $\psi_0 = \psi_1 + a$ and $\psi_0^c = \psi_1^c - a$ for some $a \in \mathbb{R}$. A solution, π, to the primal problem is called an optimal plan for $\mathsf{OT}_c(\mu_0, \mu_1)$ and, if $\pi = (\mathrm{Id}, T)_\sharp \mu_0$ for a measurable map $T : \mathcal{X} \to \mathcal{Y}$, we call T an optimal map.

In the semi-discrete setting, $\mu_0 \in \mathcal{P}(\mathbb{R}^{d_0})$ is supported in $\mathcal{X}$, an open ball centered at 0 with finite radius, and $\mu_1 \in \mathcal{P}(\mathbb{R}^{d_1})$ is supported on the points $(y^{(i)})_{i=1}^N = \mathcal{Y}$. In this case, the dual OT problem (2) reads

$$\mathsf{OT}_c(\mu_0, \mu_1) = \max_{z \in \mathbb{R}^N} \left\{ \sum_{i=1}^N z_i \mu_1\left(\{y^{(i)}\}\right) + \int \min_{1 \leq i \leq N} \left\{ c(\cdot, y^{(i)}) - z_i \right\} d\mu_0 \right\}. \tag{3}$$

We call a solution to (3) an optimal vector for $\mathsf{OT}_c(\mu_0, \mu_1)$.

2.2 Gromov-Wasserstein distance

We focus on the $(2,2)$-GW distance between probability measures, $\mu_0 \in \mathcal{P}(\mathbb{R}^{d_0}), \mu_1 \in \mathcal{P}(\mathbb{R}^{d_1})$ with finite fourth moments,

$$\mathsf{D}(\mu_0, \mu_1) = \left(\min_{\pi \in \Pi(\mu_0, \mu_1)} \iint \left| \|x - x'\|^2 - \|y - y'\|^2 \right|^2 d\pi \otimes \pi(x, y, x', y') \right)^{1/2}, \tag{4}$$

which admits the following variational form (see Corollary 1 in [57]),

$$\begin{gathered} \mathsf{D}(\bar{\mu}_0, \bar{\mu}_1)^2 = \mathsf{D}_1(\bar{\mu}_0, \bar{\mu}_1) + \mathsf{D}_2(\bar{\mu}_0, \bar{\mu}_1), \\ \mathsf{D}_1(\mu_0, \mu_1) = \int \|x - x'\|^4 d\mu_0 \otimes \mu_0(x, x') + \int \|y - y'\|^4 d\mu_1 \otimes \mu_1(y, y') - 4 M_2(\mu_0) M_2(\mu_1), \\ \mathsf{D}_2(\mu_0, \mu_1) = \min_{\boldsymbol{A} \in \mathbb{R}^{d_0 \times d_1}} \left\{ 32 \|\boldsymbol{A}\|_F^2 + \mathsf{OT}_{\boldsymbol{A}}(\mu_0, \mu_1) \right\}, \end{gathered} \tag{5}$$

where $\mathsf{OT}_{\boldsymbol{A}}(\mu_0, \mu_1) = \mathsf{OT}_{c_{\boldsymbol{A}}}(\mu_0, \mu_1)$ for $c_{\boldsymbol{A}} : (x, y) \mapsto -4\|x\|^2 \|y\|^2 - 32 x^\intercal \boldsymbol{A} y$ and we recall that $\bar{\mu}_i = (\mathrm{Id} - \mathbb{E}_{\mu_i}[X])_\sharp \mu_i$ for $i = 0, 1$. Of note is that D_1 is an explicit constant whereas D_2 is a minimization problem with objective function,

$$\Phi_{(\mu_0, \mu_1)} : \boldsymbol{A} \in \mathbb{R}^{d_0 \times d_1} \mapsto 32 \|\boldsymbol{A}\|_F^2 + \mathsf{OT}_{\boldsymbol{A}}(\mu_0, \mu_1). \tag{6}$$

Further, if $\pi^\star$ is optimal for (4), then $\boldsymbol{A}^\star = \frac{1}{2}\int xy^\intercal d\pi^\star(x,y)$ is optimal for (6) (see the proof of Theorem 1 in [57]). Our first result is to further characterize solutions of (6),

Theorem 1 (On minimizers of (6))**.** *If $\mu_0 \in \mathcal{P}(\mathbb{R}^{d_0})$, $\mu_1 \in \mathcal{P}(\mathbb{R}^{d_1})$ are compactly supported, then*

1. *$\Phi_{(\mu_0,\mu_1)}$ is locally Lipschitz continuous and coercive. If all optimal plans for $\mathsf{OT}_{\boldsymbol{A}}(\mu_0,\mu_1)$ admit the same cross-correlation matrix, $\int xy^\intercal d\pi_{\boldsymbol{A}}(x,y)$, then $\Phi_{(\mu_0,\mu_1)}$ is Fréchet differentiable at $\boldsymbol{A} \in \mathbb{R}^{d_0\times d_1}$ with $\left(D\Phi_{(\mu_0,\mu_1)}\right)_{[\boldsymbol{A}]}(\boldsymbol{B}) = 64\langle \boldsymbol{A} - \frac{1}{2}\int xy^\intercal d\pi_{\boldsymbol{A}}(x,y), \boldsymbol{B}\rangle_F$.*

2. *If $\boldsymbol{A}^\star$ minimizes (6), then $2\boldsymbol{A}^\star = \int xy^\intercal d\pi^\star(x,y) \in B_F(\sqrt{M_2(\mu_0)M_2(\mu_1)})$ for some optimal plan $\pi^\star$ for $\mathsf{OT}_{\boldsymbol{A}^\star}(\mu_0,\mu_1)$. If μ_0,μ_1 are centered, then $\pi^\star$ solves (4).*

Theorem 1 shows that all minimizers of (6) are contained in $B_F(\sqrt{M_2(\mu_0)M_2(\mu_1)})$ and, given such a minimizer, a solution to (4) 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 $\mathsf{OT}_{\boldsymbol{A}}(\mu_0,\mu_1)$ admits multiple optimal plans, $\Phi_{(\mu_0,\mu_1)}$ may fail to be differentiable at $\boldsymbol{A}$. Thus, the proof of Theorem 1 (Appendix A.1) uses the Clarke subdifferential [10] to formalize this approach.

The assumption that all optimal plans for $\mathsf{OT}_{\boldsymbol{A}}(\mu_0,\mu_1)$ admit the same cross-correlation matrix appears non-trivial to verify in general. For instance, a classical result guaranteeing uniqueness of optimal plans and existence of optimal maps (Theorem 10.28 in [53]) requires that $\nabla_x c(x,\cdot)$ is injective. As $\nabla_x c_{\boldsymbol{A}}(x,y) = -8x\|y\|^2 - 32\boldsymbol{A}y$, this condition can fail even in anodyne situations (e.g. $0 \in \mathrm{int}(\mathrm{spt}(\mu_0))$ and there exists $y, y' \in \mathrm{spt}(\mu_1) \cap \ker(\boldsymbol{A})$). The failure of this condition constitutes a roadblock to proving the existence of optimal maps for (4), called Gromov-Monge maps. In what follows, we demonstrate that, in the semi-discrete case, a simple condition guarantees uniqueness of optimal plans and existence of optimal maps for $\mathsf{OT}_{\boldsymbol{A}}(\mu_0,\mu_1)$.

3 Structural properties

In the sequel, we restrict our attention to the semi-discrete Gromov-Wasserstein (SDGW) problem wherein $\mu_0 \in \mathcal{P}(\mathbb{R}^{d_0})$ is supported in $\mathcal{X}$, an open ball centered at 0 with finite radius, and $\mu_1 \in \mathcal{P}(\mathbb{R}^{d_1})$ is supported on the points $(y^{(i)})_{i=1}^N = \mathcal{Y}$. In light of Theorem 1, in the absence of a precise characterization of the minimizers of $\Phi_{(\mu_0,\mu_1)}$ in (6), a condition ensuring uniqueness of optimal couplings for $\mathsf{OT}_{\boldsymbol{A}}(\mu_0,\mu_1)$ will prove useful in analyzing the SDGW problem. To this end, we consider the following assumption.

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

$$\mu_0\left(\left(c_{\boldsymbol{A}}(\cdot,y^{(i)}) - c_{\boldsymbol{A}}(\cdot,y^{(j)})\right)^{-1}(t)\right) = 0, \text{ for } \boldsymbol{A} \in \mathbb{R}^{d_0\times d_1}. \tag{7}$$

Assumption 1 involves standard conditions on μ_0 as to guarantee uniqueness of optimal couplings and existence of optimal maps for OT_c with $c(x,y) = h(x-y)$ for h strictly convex (see Theorem 1.2 in [23]). Condition (7) is related to $\mathcal{Y}$ and is seen to hold under the following primitive condition.

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

The proof of Proposition 1 follows essentially from the structure of the cost function, see Appendix A.2. Of note is that Proposition 1 provides a necessary and sufficient condition for Assumption 1 to hold at every $\boldsymbol{A} \in \mathbb{R}^{d_0\times d_1}$ simultaneously. This condition is crucial in our study of limit distributions for the empirical SDGW distance to guarantee e.g. uniqueness up to constants of optimal potentials for $\mathsf{OT}_{\boldsymbol{A}}(\mu_0,\mu_1)$ uniformly in $\boldsymbol{A}$.

Under Assumption 1, solutions of $\mathsf{OT}_{\boldsymbol{A}}(\mu_0,\mu_1)$ admit the following characterization.

Proposition 2 (On solutions of $\mathsf{OT}_{\boldsymbol{A}}(\mu_0,\mu_1)$)**.** *Fix $\boldsymbol{A} \in \mathbb{R}^{d_0\times d_1}$ and let $z^{\boldsymbol{A}}$ be an optimal vector for $\mathsf{OT}_{\boldsymbol{A}}(\mu_0,\mu_1)$. Under Assumption 1, the optimal plan is unique and is induced by the map*

$$T_{\boldsymbol{A}} : x \in \mathcal{X} \mapsto y^{(I_{z^{\boldsymbol{A}}}(x))}, \text{ where } I_{z^{\boldsymbol{A}}}(x) \in \mathrm{argmin}_{1\leq i\leq N}\left(c_{\boldsymbol{A}}(x,y^{(i)}) - z_i^{\boldsymbol{A}}\right), \tag{8}$$

which is uniquely defined μ_0-almost everywhere.

The proof of Proposition 2 is standard, see Appendix A.3 for details. The following result is a direct consequence of Theorem 1 and Proposition 2.

Theorem 2 (Existence of Gromov-Monge maps)**.** *Let $\boldsymbol{A}^\star$ minimize $\Phi_{(\mu_0,\mu_1)}$. If Assumption 1 holds at $\boldsymbol{A}^\star$, then there exists an optimal plan for the SDGW problem which is induced by the map $T_{\boldsymbol{A}^\star}$ given in* (8)*. Furthermore, $\boldsymbol{A}^\star = \frac{1}{2}\sum_{i=1}^N \int_{\mathrm{Lag}_{z^{\boldsymbol{A}^\star},i}} x d\mu_0(x) \left(y^{(i)}\right)^\intercal$, where $\mathrm{Lag}_{z^{\boldsymbol{A}^\star},i} = \{x \in \mathcal{X} : i \in \mathrm{argmin}_{1\leq i\leq N} \left(c_{\boldsymbol{A}^\star}(x, y^{(i)}) - z_i^{\boldsymbol{A}^\star}\right)\}$.*

The sets Lag defined in (2) are known as Laguerre cells. Proposition 1 provides a simple condition for verifying Assumption 1 at $\boldsymbol{A}^\star$ along with a condition guaranteeing that it holds at every $\boldsymbol{A} \in \mathbb{R}^{d_0\times d_1}$. To our knowledge, this is the first result providing simple, explicit conditions which guarantee the existence of Gromov-Monge maps for (4). Indeed, as mentioned previously, other results require symmetry of the marginals or *a priori* knowledge of the rank of the cross-correlation matrix of an optimal plan for (4); such conditions are restrictive, but may hold beyond the semi-discrete setting.

Remark 1 (Structure of solutions)**.** *If Assumption 1 holds at every $\boldsymbol{A} \in \mathbb{R}^{d_0\times d_1}$ then a minimizer $\boldsymbol{A}^\star$ of $\Phi_{(\mu_0,\mu_1)}$ must be such that $\boldsymbol{A}^\star_{i(\cdot)} \in \mathrm{lin}\,(\mathcal{Y})$ for $i = 1, \ldots, N$. This observation can be used to reduce the dimensionality of the problem defining D_2 if $\dim(\mathrm{lin}(\mathcal{Y})) < d_1$. In the limiting case that the points $(y^{(i)})_{i=1}^N$ are colinear, the resulting problem is d_0 dimensional for instance. Furthermore, this endows us with an* a priori *upper bound on the rank of the cross-correlation matrix for $\pi_{\boldsymbol{A}^\star}$ solving $\mathsf{OT}_{\boldsymbol{A}^\star}(\mu_0, \mu_1)$; such conditions are broadly useful in the structural study of the GW problem as evidenced by Proposition* 4.2.5 *in [51] and Theorem 5 in [18] described in the literature review.*

4 Statistical properties

Throughout, we let $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$ be independent and identically distributed samples from μ_0 and μ_1 respectively. We let $\hat{\mu}_{0,n} = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ and $\hat{\mu}_{1,n} = \frac{1}{n}\sum_{i=1}^n \delta_{Y_i}$ denote the corresponding empirical measures. In what follows, we consider both the sample complexity of the empirical SDGW distance $\mathsf{D}(\hat{\mu}_{0,n}, \hat{\mu}_{1,n})$ as well as its asymptotic distribution under proper centering and scaling. Our analysis uses the variational formula (5) along with the dual form for semi-discrete OT (3). Overall, our proof technique is similar to that used to treat the standard SDOT cost [15]. However, we stress that the need for uniformity over costs $c_{\boldsymbol{A}}$ substantially complicates the analysis and requires developing new techniques.

4.1 Sample complexity

To derive the expected rate of convergence of $\mathsf{D}(\hat{\mu}_{0,n}, \hat{\mu}_{1,n})$ to $\mathsf{D}(\mu_0, \mu_1)$, we show that the associated potential vectors and their $c_{\boldsymbol{A}}$-conjugates lie in suitable classes of functions. As such, let $M = \frac{1}{2}\|\mathcal{X}\|_\infty\|\mathcal{Y}\|_\infty \geq \frac{1}{2}\sqrt{M_2(\nu_0)M_2(\nu_1)}$ for any $\nu_0 \in \mathcal{P}(\mathcal{X}), \nu_1 \in \mathcal{P}(\mathcal{Y})$ and define

$$\mathcal{G}_{0,K} = \left\{ x \in \mathcal{X} \mapsto \min_{1\leq i\leq N} \left\{ c_{\boldsymbol{A}}(x, y^{(i)}) - z_i \right\} : \boldsymbol{A} \in B_F(M), z \in \mathbb{R}^N, \|z\|_\infty \leq K \right\},$$
$$\mathcal{G}_{1,K} = \{ f : \mathcal{Y} \to \mathbb{R} : \|f\|_{\infty,\mathcal{Y}} \leq K \}.$$

We show in Proposition 3 ahead that a dual vector to $\mathsf{OT}_{\boldsymbol{A}}(\nu_0, \nu_1)$ can always be identified with an element of $\mathcal{G}_{1,K}$ for a choice of K that is independent of $(\nu_0, \nu_1) \in \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{Y})$ and $\boldsymbol{A} \in B_F(M)$ (a straightforward modification of that argument shows that $K = 8\|\mathcal{X}\|_\infty\|^2\mathcal{Y}\|_\infty^2 + 64M\|\mathcal{X}\|_\infty\|\mathcal{Y}\|_\infty$ suffices). Evidently, its $c_{\boldsymbol{A}}$-conjugate lies in $\mathcal{G}_{0,K}$ when $\boldsymbol{A} \in B_F(M)$. In what follows, we identify $\nu_0 \in \mathcal{P}(\mathcal{X})$ and $\nu_1 \in \mathcal{P}(\mathcal{Y})$ with elements of $\ell^\infty(\mathcal{G}_{0,K})$ and $\ell^\infty(\mathcal{G}_{1,K})$ respectively by setting $\int g_i d\nu_i = \nu_i(g_i)$ for every $g_i \in \mathcal{G}_{i,K}$ and $i = 0, 1$. We now establish the expected rate of convergence of $\hat{\mu}_{i,n}$ to μ_i in $\ell^\infty(\mathcal{G}_{i,K})$ $(i = 0, 1)$.

Lemma 1. *The classes $\mathcal{G}_{0,K}$ and $\mathcal{G}_{1,K}$ are, respectively, μ_0-, μ_1-Donsker. Moreover,*

$$\mathbb{E}\left[\|\hat{\mu}_{0,n} - \mu_0\|_{\infty,\mathcal{G}_{0,K}}\right] \lesssim_{K,\mathcal{Y},\mathcal{X}} \sqrt{\frac{N + d_0 d_1}{n}} \text{ and } \mathbb{E}\left[\|\hat{\mu}_{1,n} - \mu_1\|_{\infty,\mathcal{G}_{1,K}}\right] \lesssim_K \sqrt{\frac{N}{n}}$$

The proof of Lemma 1 follows similar lines to the proof of Theorem 2.6 in [15] with the added complexity that the costs depend on $\boldsymbol{A}$ varying in $B_F(M)$. The crux of the argument is that these function classes are indexed by parameters varying in compact subsets of $\mathbb{R}^{N+d_0d_1}$ for $\mathcal{G}_{0,K}$ or $\mathbb{R}^N$ for $\mathcal{G}_{1,K}$ and hence are simple enough as to admit finite bracketing entropy and uniform entropy respectively (cf. e.g. [50]) whilst containing all relevant optimal potentials. The derived rates are similar to that obtained in [15] for SOT with the addition of the factor d_0d_1 to account for the varying costs, see Appendix A.4 for full details. With this, we obtain the first sample complexity results for the SDGW problem.

Theorem 3. *Let $R = \operatorname{diam}(\mathcal{X}) \vee \operatorname{diam}(\mathcal{Y})$, then*

$$\mathbb{E}\left[|\mathsf{D}(\mu_0,\mu_1)^2 - \mathsf{D}(\hat{\mu}_{0,n},\hat{\mu}_{1,n})^2|\right] \lesssim_{K,\mathcal{X},\mathcal{Y}} \left(2R^4 + \sqrt{N+d_0d_1} + \sqrt{N}\right) n^{-1/2},$$
$$\mathbb{E}\left[|\mathsf{D}(\mu_0,\mu_1) - \mathsf{D}(\hat{\mu}_{0,n},\hat{\mu}_{1,n})|\right] \lesssim_{K,\mathcal{X},\mathcal{Y}} \mathsf{D}(\mu_0,\mu_1)^{-1} \left(2R^4 + \sqrt{N+d_0d_1} + \sqrt{N}\right) n^{-1/2}.$$

The proof of Theorem 3 (Appendix A.5) leverages the variational form of the SDGW problem (5) along with the existence of a minimizer of $\Phi_{(\nu_0,\nu_1)}$ in $B_F(M)$ for any choice of $(\nu_0,\nu_1) \in \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{Y})$ (see Theorem 1). With this, the question of sample complexity for empirical SDGW reduces to that of empirical OT uniformly over the collection of costs $(c_{\boldsymbol{A}})_{\boldsymbol{A}\in B_F(M)}$ which can be addressed using Lemma 1. Of note is that $\mathsf{D}(\mu_0,\mu_1) > 0$ in the semi-discrete case.

4.2 Limit distribution theory

We now derive the asymptotic distribution of the empirical SDGW distance. Our approach is based on the extended functional delta method. This approach requires deriving a first order expansion of the SDGW distance as a function of its marginals and proving that the relevant empirical processes converge in a suitable space.

Precisely, we treat D as a functional on the set

$$\mathfrak{P} = \{\nu_0 \otimes \nu_1 : \nu_0 \in \mathcal{P}(\mathcal{X}), \operatorname{spt}(\nu_0) \subset \operatorname{spt}(\mu_0), \nu_1 \in \mathcal{P}(\mathcal{Y})\},$$

which we treat as a subset of the space $\ell^\infty(\mathcal{F}_K^\oplus)$, where $\mathcal{F}_K^\oplus = \{f_0 \oplus f_1 : f_0 \in \mathcal{F}_{0,K}, f_1 \in \mathcal{G}_{1,K}\}$, for $\mathcal{F}_{0,K} = \mathcal{C}_K^\infty(\mathcal{X}) + \mathcal{H}_{0,K} \cup \{0\}$,

$$\mathcal{H}_{0,K} = \left\{x \in \mathcal{X} \mapsto \min_{1\le i\le N}\left\{c_{\boldsymbol{A}}(x-\xi, y^{(i)}) - z_i\right\} : \boldsymbol{A} \in B_F(M), z \in \mathbb{R}^N, \|z\|_\infty \le K, \xi \in \mathcal{X}\right\}.$$

Here, for any $\nu_0 \otimes \nu_1 \in \mathfrak{P}$ and $f_0 \oplus f_1 \in \mathcal{F}_K^\oplus$, $\nu_0 \otimes \nu_1(f_0 \oplus f_1) = \nu_0(f_0) + \nu_1(f_1) = \int f_0 d\nu_0 + \int f_1 d\nu_1$. Note that $\tau : \{(\nu_0,\nu_1) \in \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{Y}) : \operatorname{spt}(\nu_0) \subset \operatorname{spt}(\mu_0)\} \mapsto \nu_0 \otimes \nu_1 \in \mathfrak{P}$ is one-to-one (see Proposition 6 in Appendix B), hence any functional on the latter set can be identified with a functional on $\mathfrak{P}$; for convenience, we denote both functionals with the same symbol.

In what follows it will be convenient to work with the following extensions of the optimal vector and its $c_{\boldsymbol{A}}$-conjugate which we call extended potentials.

Definition 1 (Extended potentials)**.** *Let $\mathcal{X}^\circ = 2\mathcal{X}$ and $\mathcal{Y}^\circ$ be an open ball centered at 0 with radius $r > 2\|\mathcal{Y}\|_\infty$. Let $(\nu_0,\nu_1) \in \mathcal{P}(\mathcal{X}^\circ) \times \mathcal{P}(\mathcal{Y}^\circ)$ be such that ν_1 is distributed on N points $(y^{(i)})_{i=1}^N$. For $\boldsymbol{A} \in \mathbb{R}^{d_0\times d_1}$, let $z^{\boldsymbol{A}}$ be an optimal vector for $\mathsf{OT}_{\boldsymbol{A}}(\nu_0,\nu_1)$. Then, the extended potentials for $\mathsf{OT}_{\boldsymbol{A}}(\nu_0,\nu_1)$ are given by*

$$\varphi^{\boldsymbol{A}} : x \in \mathcal{X}^\circ \mapsto \min_{1\le i\le N}\left\{c_{\boldsymbol{A}}(x, y^{(i)}) - z_i^{\boldsymbol{A}}\right\},$$
$$\psi^{\boldsymbol{A}} : y \in \mathcal{Y}^\circ \mapsto \inf_{x\in\mathcal{X}^\circ}\left\{c_{\boldsymbol{A}}(x,y) - \varphi^{\boldsymbol{A}}(x)\right\}.$$

These potentials are defined on $\mathcal{X}^\circ$ and $\mathcal{Y}^\circ$, as if $\nu \in \mathcal{P}(\mathbb{R}^d)$ is such that $\operatorname{spt}(\nu) \subset B(r)$ for $r > 0$, then $\|\mathbb{E}_\nu[X]\| \le \mathbb{E}_\nu\|X\| \le \|\operatorname{spt}(\nu)\|_\infty < r$, so $B(r) - \mathbb{E}_\nu[X] \subset 2B(r)$. Consequently, the extended potentials $(\varphi^{\boldsymbol{A}}, \psi^{\boldsymbol{A}})$ are well-defined on $\operatorname{spt}(\bar{\nu}_0) \times \operatorname{spt}(\bar{\nu}_1)$ for any $\nu_0 \otimes \nu_1 \in \mathfrak{P}$.

The extended potentials satisfy many of the useful properties exhibited by standard optimal potentials. Among other results, we estabish that $\psi^{\boldsymbol{A}}$ is a proper extension of $z^{\boldsymbol{A}}$, proofs can be found in Appendix A.6.

Proposition 3 (Properties of extended potentials)**.** *Fix $\nu_0 \in \mathcal{P}(\mathcal{X}^\circ), \nu_1 \in \mathcal{P}(\mathcal{Y}^\circ)$ with $\mathrm{spt}(\nu_1) = (\bar{y}^{(i)})_{i=1}^N$, and $\boldsymbol{A} \in B_F(M)$. Let $z^{\boldsymbol{A}}$ be an optimal vector for $\mathsf{OT}_{\boldsymbol{A}}(\nu_0, \nu_1)$ and $(\varphi^{\boldsymbol{A}}, \psi^{\boldsymbol{A}})$ be extended potentials. Then,*

1. *$\varphi^{\boldsymbol{A}}$ and $\psi^{\boldsymbol{A}}$ are concave and Lipschitz continuous with a shared Lipschitz constant which is independent of $\boldsymbol{A}, \nu_0, \nu_1$.*
2. *let $\Lambda_j = \left\{x \in \mathcal{X}^\circ : \mathrm{argmin}_{1\le i\le N}\left(c_{\boldsymbol{A}}\left(x, y^{(i)}\right) - z_i^{\boldsymbol{A}}\right) = \{j\}\right\}$ for $j = 1, \dots, N$. If Assumption 1 holds at $\boldsymbol{A}$, then $(\Lambda_j)_{j=1}^N$ partitions $\mathcal{X}^\circ$ up to a measure zero set and $\varphi^{\boldsymbol{A}}$ is differentiable at $x \in \Lambda_j$ with*
$$\nabla\varphi^{\boldsymbol{A}}(x) = -8x\|y^{(j)}\|^2 - 32\boldsymbol{A}y^{(j)}, \quad D^2\varphi^{\boldsymbol{A}}(x) = -8\,\mathrm{Id}\,\|y^{(j)}\|^2.$$
3. *For $i = 1, \dots, N$, $\psi^{\boldsymbol{A}}\left(\bar{y}^{(i)}\right) = z_i^{\boldsymbol{A}}$. Further, $z^{\boldsymbol{A}}, \varphi^{\boldsymbol{A}}, \psi^{\boldsymbol{A}}$ can be chosen such that $\|z^{\boldsymbol{A}}\|_\infty \vee \|\varphi^{\boldsymbol{A}}\|_{\infty,\mathcal{X}^\circ} \vee \|\psi^{\boldsymbol{A}}\|_{\infty,\mathcal{Y}^\circ} \le C_{M,\mathcal{X}^\circ,\mathcal{Y}^\circ}$, where $C_{M,\mathcal{X}^\circ,\mathcal{Y}^\circ}$ depends only on $M, \mathcal{X}^\circ, \mathcal{Y}^\circ$.*
4. *If Assumption 1 holds at $\boldsymbol{A}$, $\mathrm{spt}(\nu_0)$ has negligible boundary, and $\mathrm{int}(\mathrm{spt}(\nu_0))$ is connected, then the pair $(\varphi^{\boldsymbol{A}}, \psi^{\boldsymbol{A}})$ is unique up to additive constants on $\mathrm{int}(\mathrm{spt}(\nu_0)) \times \mathrm{spt}(\nu_1)$.*

With Proposition 3 in hand, we show stability of the SDGW distance in the sense of Hadamard.

Definition 2 (Hadamard directional derivative [41, 44])**.** *Let $\mathfrak{D}, \mathfrak{E}$ be normed spaces and fix a non-empty set $\Theta \subset \mathfrak{D}$. For $\theta \in \Theta$, the tangent cone to Θ at θ is given by*
$$\mathcal{T}_\Theta(\theta) = \left\{h \in \mathfrak{D} : h = \lim_{n\to\infty} \frac{\theta_n - \theta}{t_n}, \text{ for some } \theta_n \in \Theta, \theta_n \to \theta, t_n \downarrow 0\right\}.$$
A map $f : \Theta \to \mathfrak{E}$ is Hadamard directionally differentiable at $\theta \in \Theta$ if there exists a map $f'_\theta : \mathcal{T}_\Theta(\theta) \to \mathbb{E}$ satisfying
$$\lim_{n\to\infty} \frac{f(\theta + t_n h_n) - f(\theta)}{t_n} = f'_\theta(h),$$
for any $h \in \mathcal{T}_\Theta(\theta)$, $t_n \downarrow 0$, and $h_n \to h$ in $\mathfrak{D}$ with $\theta + t_n h_n \in \Theta$.

In our application, D is a functional on $\mathfrak{P} \subset \ell^\infty(\mathcal{F}_K^\oplus)$. It is readily seen that $\mathfrak{P}$ is convex as a subset of $\ell^\infty(\mathcal{F}_K^\oplus)$, so $\mu_0 \otimes \mu_1 + t(\nu_0 \otimes \nu_1 - \mu_0 \otimes \mu_1) \in \mathfrak{P}$ for $t \in [0, 1]$ and $\mathcal{T}_{\mathfrak{P}}(\mu_0 \otimes \mu_1) = \overline{\{t^{-1}(\nu_0 \otimes \nu_1 - \mu_0 \otimes \mu_1) : \nu_0 \otimes \nu_1 \in \mathfrak{P}, t > 0\}}^{\ell^\infty(\mathcal{F}_K^\oplus)}$ [41]. Given this expression, if $f_0 \oplus f_1 \in \mathcal{F}_K^\oplus$, then for any $\eta \in \mathcal{T}_{\mathfrak{P}}(\mu_0 \otimes \mu_1)$ and $\alpha \in \mathbb{R}$, $\alpha\eta(f_0 \oplus f_1) = \lim_{n\to\infty} t_n^{-1}(\nu_{0,n} \otimes \nu_{1,n} - \mu_0 \otimes \mu_1)(\alpha(f_0 \oplus f_1))$ for some $t_n > 0$, $\nu_{0,n} \otimes \nu_{1,n} \in \mathfrak{P}$ such that η extends uniquely to $\alpha\mathcal{F}_K^\oplus$. Such extensions are used in the following results.

Theorem 4 (Stability of SDGW)**.** *Assume that $\bar{y}^{(i)} = y^{(i)} - \mathbb{E}_{\mu_1}[X]$ $(i = 1, \dots, N)$ is such that $\|\bar{y}^{(i)}\| \ne \|\bar{y}^{(j)}\|$ for $i \ne j$, $\bar{\mu}_0$ satisfies Assumption 1, $\mathrm{spt}(\bar{\mu}_0)$ has a negligible boundary, and $\mathrm{int}\left(\mathrm{spt}(\bar{\mu}_0)\right)$ is connected. Then, the map $\nu_0 \otimes \nu_1 \in \mathfrak{P} \mapsto \mathsf{D}(\bar{\nu}_0, \bar{\nu}_1)^2$ is Hadamard directionally differentiable at $\mu_0 \otimes \mu_1$ with derivative*
$$\eta \in \mathcal{T}_{\mathfrak{P}}(\mu_0 \otimes \mu_1) \mapsto \eta(f_0 \oplus f_1) + \inf_{\mathrm{argmin}_{B_F(M)}\left(\Phi_{(\bar{\mu}_0,\bar{\mu}_1)}\right)} \left\{\eta\left(L_0(\varphi^{(\cdot)}) \oplus L_1(\psi^{(\cdot)})\right)\right\},$$
where $f_i = 2\int \|\cdot - x\|^4 d\mu_i(x) - 4M_2(\mu_i)\int \|\cdot - \mathbb{E}_{\mu_i}[X]\|^2$ for $i = 0, 1$, $(\varphi^{\boldsymbol{A}}, \psi^{\boldsymbol{A}})$ is any pair of extended potentials for $\mathsf{OT}_{\boldsymbol{A}}(\bar{\mu}_0, \bar{\mu}_1)$ with $\boldsymbol{A} \in B_F(M)$, and
$$L_0(\varphi^{(\cdot)}) : x \in \mathcal{X} \mapsto \varphi^{(\cdot)}(x - \mathbb{E}_{\mu_0}[X]) - \mathbb{E}_{\bar{\mu}_0}[\nabla\varphi^{(\cdot)}(X)]^\intercal x,$$
$$L_1(\psi^{(\cdot)}) : y \in \mathcal{Y} \mapsto \psi^{(\cdot)}(y - \mathbb{E}_{\mu_1}[X]) - 8\mathbb{E}_{(X,Y)\sim\pi_{(\cdot)}}[\|X\|^2 Y]^\intercal y,$$
where $\pi_{\boldsymbol{A}}$ is the unique optimal plan for $\mathsf{OT}_{\boldsymbol{A}}(\bar{\mu}_0, \bar{\mu}_1)$.

The proof of Theorem 4 relies on a connection between Hadamard directional differentiability and Gâteaux directional differentiability (on $\mathfrak{P} - \mu_0 \otimes \mu_1$) for Lipschitz continuous maps Proposition 7 in Appendix B. Using the decomposition $\mathsf{D}(\bar{\nu}_0, \bar{\nu}_1)^2 = \mathsf{D}_1(\bar{\nu}_0, \bar{\nu}_1) + \mathsf{D}_2(\bar{\nu}_0, \bar{\nu}_1)$, it suffices to separately prove Hadamard directional differentiability of each summand at $\mu_0 \otimes \mu_1$. The Hadamard directional

derivative of $\mathsf{D}_1(\bar{\nu}_0, \bar{\nu}_1)$ is relatively straightforward to derive, whereas that of $\mathsf{D}_2(\bar{\nu}_0, \bar{\nu}_1)$ requires a more careful analysis.

Pending differentiability of $\nu_0 \otimes \nu_1 \in \mathfrak{P} \mapsto \mathsf{OT}_{(\cdot)}(\bar{\nu}_0, \bar{\nu}_1) \in \ell^\infty(B_F(M))$, the chain rule for Hadamard directionally differentiable maps along with a known result establishing Hadamard differentiability of infimum-type functionals [6] prove differentiability of $\mathsf{D}_2(\bar{\nu}_0, \bar{\nu}_1)$ given its variational form (5). A major obstacle to proving differentiability of $\mathsf{OT}_{(\cdot)}(\bar{\nu}_0, \bar{\nu}_1)$ is that the centered perturbations of μ_1 may not be supported on $\mathcal{Y} - \mathbb{E}_{\mu_1}[X]$. To account for this, we write $t^{-1}\left(\mathsf{OT}_{(\cdot)}(\bar{\mu}_{0,t}, \bar{\mu}_{1,t}) - \mathsf{OT}_{(\cdot)}(\bar{\mu}_0, \bar{\mu}_1)\right)$, for $\mu_{i,t} = \mu_i + t(\nu_i - \mu_i)$ $(i = 0, 1)$, $\nu_0 \otimes \nu_1 \in \mathfrak{P}$ as

$$t^{-1}\left(\mathsf{OT}_{(\cdot)}(\bar{\mu}_{0,t}, \bar{\mu}_{1,t}) - \mathsf{OT}_{(\cdot)}(\bar{\mu}_{0,t}, (\mathrm{Id} - \mathbb{E}_{\mu_1}[X])_\sharp \mu_{1,t})\right) + t^{-1}\left(\mathsf{OT}_{(\cdot)}(\bar{\mu}_{0,t}, (\mathrm{Id} - \mathbb{E}_{\mu_1}[X])_\sharp \mu_{1,t}) - \mathsf{OT}_{(\cdot)}(\bar{\mu}_0, \bar{\mu}_1)\right),$$

and analyze each term separately. Complete details are included in Appendix A.7.

Of note is that Theorem 4 requires the optimal potentials to be unique up to additive constants for every $\boldsymbol{A} \in B_F(M)$, which can be guaranteed by appealing to Proposition 1 and Proposition 3.

Given Theorem 4, the subsequent limit distribution result follows by applying the extended functional delta method [17, 20, 41, 45] once Donskerness of the class $\mathcal{F}_{0,K}$ has been established (see Appendix A.9 for a complete proof).

Theorem 5 (Semi-discrete GW limit distribution)**.** *In the setting of Theorem 4, for any $K > 0$, there exists a tight μ_0-Brownian bridge process G_{μ_0} in $\ell^\infty(\mathcal{F}_{0,K})$, and a tight μ_1-Brownian bridge process G_{μ_1} in $\ell^\infty(\mathcal{G}_{1,K})$ for which*

$$\sqrt{n}\left(\mathsf{D}(\hat{\mu}_{0,n}, \hat{\mu}_{1,n})^2 - \mathsf{D}(\mu_0, \mu_1)^2\right) \xrightarrow{d} G_{\mu_0}(f_0) + G_{\mu_1}(f_1) + \inf_{\mathrm{argmin}_{B_F(M)}\left(\Phi_{(\bar{\mu}_0, \bar{\mu}_1)}\right)} \left\{G_{\mu_0}\left(L_0(\varphi^{(\cdot)})\right) + G_{\mu_1}\left(L_1(\psi^{(\cdot)})\right)\right\}.$$

In the absence of conditions guaranteeing the uniqueness of minimizers to $\Phi_{(\bar{\mu}_0, \bar{\mu}_1)}$, the derived limit distribution involves an $\inf$ over Gaussian processes and hence may fail to be normal. Remark that the power of 2 in the empirical semi-discrete GW distance can be shed by applying the standard delta method with the function $\sqrt{(\cdot)}$.

5 Conclusion

In this paper, we have provided a primitive condition which guarantees the existence of Gromov-Monge maps for the SDGW problem. To our knowledge, this is the first result that does not require high level conditions such as symmetry of the marginals or knowledge of the rank of the cross-correlation matrix of an optimal coupling. This condition also enabled us to establish the limit distribution of the empirical SDGW problem, where it is used to guarantee uniqueness of the extended potentials up to constants. To complement this asymptotic result, we also derivethe finite-sample performance of the empirical SDGW estimator, showing that it converges to its population-level counterpart in expectation at a parametric rate. This result is in stark contrast to the dimension-dependent rate obtained in the continuous regime [26, 57].

Acknowledgments and Disclosure of Funding

Z. Goldfeld is partially supported by NSF grants CCF-2046018, DMS-2210368, and CCF-2308446, and the IBM Academic Award. K. Kato is partially supported by the NSF grants DMS-1952306, DMS-2014636, and DMS-2210368. G. Rioux is partially supported by the NSERC Postgraduate Fellowship PGSD-567921-2022.

References

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

[2] Dongsheng An, Yang Guo, Na Lei, Zhongxuan Luo, Shing-Tung Yau, and Xianfeng Gu, *Ae-ot: A new generative model based on extended semi-discrete optimal transport*, ICLR 2020 (2019).

[3] Mohit Bansil and Jun Kitagawa, *Quantitative stability in the geometry of semi-discrete optimal transport*, International Mathematics Research Notices **2022** (2022), no. 10, 7354–7389.

[4] Dimitri P Bertsekas, *Nonlinear programming*, 2 ed., Athena Scientific, 1999.

[5] Charlotte Bunne, David Alvarez-Melis, Andreas Krause, and Stefanie Jegelka, *Learning generative models across incomparable spaces*, International conference on machine learning, PMLR, 2019, pp. 851–861.

[6] Javier Cárcamo, Antonio Cuevas, and Luis-Alberto Rodríguez, *Directional differentiability for supremum-type functionals: statistical applications*, Bernoulli **26** (2020), no. 3, 2143–2175.

[7] Yucheng Chen, Matus Telgarsky, Chao Zhang, Bolton Bailey, Daniel Hsu, and Jian Peng, *A gradual, semi-discrete approach to generative network training via explicit wasserstein minimization*, International Conference on Machine Learning, PMLR, 2019, pp. 1071–1080.

[8] Samir Chowdhury and Tom Needham, *Generalized spectral clustering via gromov-wasserstein learning*, International Conference on Artificial Intelligence and Statistics, PMLR, 2021, pp. 712–720.

[9] Francis H Clarke, Yuri S Ledyaev, Ronald J Stern, and Peter R Wolenski, *Nonsmooth analysis and control theory*, vol. 178, Springer Science & Business Media, 2008.

[10] Frank H Clarke, *Generalized gradients and applications*, Transactions of the American Mathematical Society **205** (1975), 247–262.

[11] ———, *Optimization and nonsmooth analysis*, SIAM, 1990.

[12] Fernando De Goes, Katherine Breeden, Victor Ostromoukhov, and Mathieu Desbrun, *Blue noise through optimal transport*, ACM Transactions on Graphics (TOG) **31** (2012), no. 6, 1–11.

[13] Fernando De Goes, Corentin Wallez, Jin Huang, Dmitry Pavlov, and Mathieu Desbrun, *Power particles: an incompressible fluid solver based on power diagrams.*, ACM Trans. Graph. **34** (2015), no. 4, 50–1.

[14] Eustasio del Barrio, Alberto González-Sanz, and Jean-Michel Loubes, *Central limit theorems for general transportation costs*, arXiv preprint arXiv:2102.06379 (2021).

[15] ———, *Central limit theorems for semidiscrete wasserstein distances*, arXiv preprint arXiv:2202.06380 (2022).

[16] Richard M Dudley, *Uniform central limit theorems*, vol. 142, Cambridge university press, 2014.

[17] Lutz Dümbgen, *On nondifferentiable functions and the bootstrap*, Probability Theory and Related Fields **95** (1993), no. 1, 125–140.

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

[19] Danielle Ezuz, Justin Solomon, Vladimir G Kim, and Mirela Ben-Chen, *Gwcnn: A metric alignment layer for deep shape analysis*, Computer Graphics Forum, vol. 36, Wiley Online Library, 2017, pp. 49–57.

[20] Zheng Fang and Andres Santos, *Inference on directionally differentiable functions*, The Review of Economic Studies **86** (2019), 377–412.

[21] Nicolas Fournier and Arnaud Guillin, *On the rate of convergence in wasserstein distance of the empirical measure*, Probability Theory and Related Fields **162** (2015), no. 3, 707–738.

[22] Thomas O Gallouët and Quentin Mérigot, *A lagrangian scheme à la brenier for the incompressible euler equations*, Foundations of Computational Mathematics **18** (2018), no. 4, 835–865.

[23] Wilfrid Gangbo and Robert J. McCann, *The geometry of optimal transportation*, Acta Mathematica **177** (1996), no. 2, 113–161.

[24] David Gilbarg and Neil S Trudinger, *Elliptic partial differential equations of second order*, vol. 224, Springer, 2015.

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

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

[27] Antoine Houdard, Arthur Leclaire, Nicolas Papadakis, and Julien Rabin, *On the existence of optimal transport gradient for learning generative models*, arXiv preprint arXiv:2102.05542 (2021).

[28] Jun Kitagawa, Quentin Mérigot, and Boris Thibert, *Convergence of a newton algorithm for semi-discrete optimal transport*, Journal of the European Mathematical Society **21** (2019), no. 9, 2603–2651.

[29] Achim Klenke, *Probability theory: a comprehensive course*, Springer Science & Business Media, 2013.

[30] Andrei Nikolaevich Kolmogorov and Vladimir Mikhailovich Tikhomirov, *ε-entropy and ε-capacity of sets in function spaces*, Uspekhi Matematicheskikh Nauk **14** (1959), no. 2, 3–86.

[31] Bruno Lévy, *A numerical algorithm for l_{2} semi-discrete optimal transport in 3d*, ESAIM: Mathematical Modelling and Numerical Analysis-Modélisation Mathématique et Analyse Numérique **49** (2015), no. 6, 1693–1715.

[32] Bruno Lévy, Roya Mohayaee, and Sebastian von Hausegger, *A fast semidiscrete optimal transport algorithm for a unique reconstruction of the early universe*, Monthly Notices of the Royal Astronomical Society **506** (2021), no. 1, 1165–1185.

[33] Pascal Massart, *Concentration inequalities and model selection: Ecole d'eté de probabilités de saint-flour xxxiii-2003*, Springer, 2007.

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

[35] ———, *Gromov-Wasserstein distances and the metric approach to object matching*, Found. Comput. Math. **11** (2011), no. 4, 417–487.

[36] Facundo Mémoli and Tom Needham, *Distance distributions and inverse problems for metric measure spaces*, 2021.

[37] Mina Ossiander, *A central limit theorem under metric entropy with l2 bracketing*, The Annals of Probability (1987), 897–919.

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

[39] R. T. Rockafellar, *Convex analysis*, Princeton Univ. Press, New Jersey, 1970.

[40] T. R. Rockafellar and R. J-B Wets, *Variational Analysis*, vol. 317, Springer Science & Business Media, 2009.

[41] Werner Römisch, *Delta method, infinite dimensional*, Encyclopedia of Statistical Sciences, Wiley, 2004.

[42] Ritwik Sadhu, Ziv Goldfeld, and Kengo Kato, *Limit theorems for semidiscrete optimal transport maps*, arXiv preprint arXiv:2303.10155 (2023).

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

[44] Alexander Shapiro, *On concepts of directional differentiability*, Journal of Optimization Theory and Applications **66** (1990), 477–487.

[45] Alexander Shapiro, *Asymptotic analysis of stochastic programs*, Annals of Operations Research **30** (1991), no. 1, 169–186.

[46] Justin Solomon, Gabriel Peyré, Vladimir G Kim, and Suvrit Sra, *Entropic metric alignment for correspondence problems*, ACM Transactions on Graphics (ToG) **35** (2016), no. 4, 1–13.

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

[48] Vayer Titouan, Nicolas Courty, Romain Tavenard, and Rémi Flamary, *Optimal transport for structured data with application on graphs*, International Conference on Machine Learning, PMLR, 2019, pp. 6275–6284.

[49] Aad W van der Vaart, *New Donsker classes*, The Annals of Probability **24** (1996), no. 4, 2128–2124.

[50] Aad W van der Vaart and Jon A Wellner, *Weak convergence and empirical processes*, Springer, 1996.

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

[52] Cédric Villani, *Topics in optimal transportation*, no. 58, American Mathematical Soc., 2003.

[53] Cédric Villani, *Optimal transport: Old and new*, Springer, 2008.

[54] J. Weed and F. Bach, *Sharp asymptotic and finite-sample rates of convergence of empirical measures in Wasserstein distance*, Bernoulli **25** (2019), no. 4A, 2620–2648.

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

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

[57] Zhengxin Zhang, Ziv Goldfeld, Youssef Mroueh, and Bharath K. Sriperumbudur, *Gromov-Wasserstein distances: entropic regularization, duality, and sample complexity*, arXiv preprint arXiv:2212.12848 (2022).

A Proof of main results

A.1 Proof of Theorem 1

To simplify the proof of Theorem 1, we separately prove each statement as its own lemma.

Lemma 2. *$\Phi_{(\mu_0,\mu_1)}$ is Fréchet differentiable at $\boldsymbol{A} \in \mathbb{R}^{d_0 \times d_1}$ with derivative $\left(D\Phi_{(\mu_0,\mu_1)}\right)_{[\boldsymbol{A}]}(\boldsymbol{B}) = 64\langle \boldsymbol{A} - \frac{1}{2}\int xy^\intercal d\pi_{\boldsymbol{A}}(x,y), \boldsymbol{B}\rangle_F$ provided all optimal couplings for $\mathsf{OT}_{\boldsymbol{A}}(\mu_0,\mu_1)$ admit the same cross-correlation matrix $\frac{1}{2}\int xy^\intercal d\pi_{\boldsymbol{A}}(x,y)$.*

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

$$\mathsf{OT}_{\boldsymbol{A}+\boldsymbol{H}}(\mu_0,\mu_1) - \mathsf{OT}_{\boldsymbol{A}}(\mu_0,\mu_1) \leq \int c_{\boldsymbol{A}+\boldsymbol{H}} d\pi_{\boldsymbol{A}} - \int c_{\boldsymbol{A}} d\pi_{\boldsymbol{A}} = -32 \int x^\intercal \boldsymbol{H} y d\pi_{\boldsymbol{A}}(x,y), \quad (9)$$

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

$$\mathsf{OT}_{\boldsymbol{A}+\boldsymbol{H}}(\mu_0,\mu_1) - \mathsf{OT}_{\boldsymbol{A}}(\mu_0,\mu_1) \geq -32 \int x^\intercal \boldsymbol{H} y d\pi_{\boldsymbol{A}+\boldsymbol{H}}(x,y), \quad (10)$$

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

$$\sup_{\substack{x\in\mathrm{spt}(\mu_0)\\ y\in\mathrm{spt}(\mu_1)}} |c_{\boldsymbol{A}+\boldsymbol{H}_n}(x,y) - c_{\boldsymbol{A}}(x,y)| = \sup_{\substack{x\in\mathrm{spt}(\mu_0)\\ y\in\mathrm{spt}(\mu_1)}} |32 x^\intercal \boldsymbol{H}_n y| \leq 32 \sup_{\mathrm{spt}(\mu_0)} \|\cdot\| \sup_{\mathrm{spt}(\mu_1)} \|\cdot\| \|\boldsymbol{H}_n\|_F \to 0,$$

hence $c_{\boldsymbol{A}+\boldsymbol{H}_n} \to c_{\boldsymbol{A}}$ uniformly on $\mathrm{spt}(\mu_0) \times \mathrm{spt}(\mu_1)$. It follows from Theorem 5.20 in [53] that, for any subsequence n' of n there exists a further subsequence n'' along which $\pi_{\boldsymbol{A}+\boldsymbol{H}_{n''}} \xrightarrow{w} \pi$ for some optimal coupling π for $\mathsf{OT}_{\boldsymbol{A}}(\mu_0,\mu_1)$. Thus $\int xy^\intercal d\pi_{\boldsymbol{A}+\boldsymbol{H}_{n''}}(x,y) \to \int xy^\intercal d\pi(x,y) = \int xy^\intercal d\pi_{\boldsymbol{A}}(x,y)$ by assumption. As the limit is the same regardless of the choice of subsequence, conclude that $\int xy^\intercal d\pi_{\boldsymbol{A}+\boldsymbol{H}}(x,y) \to \int xy^\intercal d\pi_{\boldsymbol{A}}(x,y)$ as $\boldsymbol{H} \to 0$, thus

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

which proves the claim. □

Lemma 3. *$\Phi_{(\mu_0,\mu_1)}$ is locally Lipschitz continuous and coercive.*

Proof. Fix a compact set $K \subset \mathbb{R}^{d_0 \times d_1}$. For any $\boldsymbol{A}, \boldsymbol{A}' \in K$, it follows from (9) and (10) that,

$$\begin{aligned}
\mathsf{OT}_{\boldsymbol{A}'}(\mu_0,\mu_1) - \mathsf{OT}_{\boldsymbol{A}}(\mu_0,\mu_1) &\leq -32 \int x^\intercal (\boldsymbol{A}' - \boldsymbol{A}) y d\pi_{\boldsymbol{A}}(x,y) \\
&\leq 32 \|\boldsymbol{A}' - \boldsymbol{A}\|_F \left\| \int xy^\intercal d\pi_{\boldsymbol{A}}(x,y) \right\|_F, \\
\mathsf{OT}_{\boldsymbol{A}'}(\mu_0,\mu_1) - \mathsf{OT}_{\boldsymbol{A}}(\mu_0,\mu_1) &\geq -32 \|\boldsymbol{A}' - \boldsymbol{A}\|_F \left\| \int xy^\intercal d\pi_{\boldsymbol{A}'}(x,y) \right\|_F,
\end{aligned}$$

that is,

$$|\mathsf{OT}_{\boldsymbol{A}'}(\mu_0,\mu_1) - \mathsf{OT}_{\boldsymbol{A}}(\mu_0,\mu_1)| \leq 32 \|\boldsymbol{A}' - \boldsymbol{A}\|_F \left(\left\| \int xy^\intercal d\pi_{\boldsymbol{A}'}(x,y) \right\|_F \bigvee \left\| \int xy^\intercal d\pi_{\boldsymbol{A}}(x,y) \right\|_F \right).$$

Observe that, for any coupling $\pi \in \Pi(\mu_0,\mu_1)$, $\left\|\int xy^\intercal d\pi(x,y)\right\|_F \leq \int \|x\|\|y\| d\pi(x,y) \leq \sqrt{M_2(\mu_0)M_2(\mu_1)}$ by Jensen's inequality and the Cauchy-Schwarz inequality. Conclude that

$$|\mathsf{OT}_{\boldsymbol{A}'}(\mu_0,\mu_1) - \mathsf{OT}_{\boldsymbol{A}}(\mu_0,\mu_1)| \leq 32\sqrt{M_2(\mu_0)M_2(\mu_1)} \|\boldsymbol{A}' - \boldsymbol{A}\|_F.$$

Further, $\left|\|\boldsymbol{A}\|_F^2 - \|\boldsymbol{A}'\|_F^2\right| = (\|\boldsymbol{A}\|_F + \|\boldsymbol{A}'\|_F)\,|\|\boldsymbol{A}\|_F - \|\boldsymbol{A}'\|_F| \leq 2\sup_K \|\cdot\|_F \|\boldsymbol{A} - \boldsymbol{A}'\|_F$.

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

$$\int -4\|x\|^2\|y\|^2 - 32x^\intercal \boldsymbol{A} y d\pi_{\boldsymbol{A}}(x,y) \geq -4\sqrt{M_4(\mu_0)M_4(\mu_1)} - 32\sqrt{M_2(\mu_0)M_2(\mu_1)}\|\boldsymbol{A}\|_F$$

Hence $32\|\boldsymbol{A}\|_F + \frac{\mathsf{OT}_{\boldsymbol{A}}(\mu_0,\mu_1)}{\|\boldsymbol{A}\|_F} \to \infty$ as $\|\boldsymbol{A}\|_F \to \infty$ proving coercivity. □

Lemma 4. *Let $\pi \in \Pi(\mu_0, \mu_1)$ be arbitrary, then $\int xy^\intercal d\pi(x,y) \in B_F(\sqrt{M_2(\mu_0)M_2(\mu_1)})$.*

Proof. By Jensen's inequality, $\left\|\int xy^\intercal d\pi(x,y)\right\|_F \leq \int \|xy^\intercal\|_F \, d\pi(x,y) = \int \|x\|\|y\| d\pi(x,y)$. 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. □

We now prove point 2, concluding the proof of Theorem 1.

Proof of Theorem 1 2. By Rademacher's theorem, $\Phi_{(\mu_0,\mu_1)}$ is differentiable on a set Λ of full measure and, by Theorem 8.1 in [9], the Clarke subdifferential of $\Phi_{(\mu_0,\mu_1)}$ at $\boldsymbol{A} \in \mathbb{R}^{d_0 \times d_1}$ can be defined as $\partial\Phi_{(\mu_0,\mu_1)}(\boldsymbol{A}) = \operatorname{conv}\left(\left\{\lim_{\Omega \ni \boldsymbol{A}_n \to \boldsymbol{A}} (D\Phi_{(\mu_0,\mu_1)})_{[\boldsymbol{A}_n]}\right\}\right)$ where Ω is any subset of Λ for which $\Lambda \backslash \Omega$ is negligible and it is presupposed that the limit converges. From Lemma 2, $(D\Phi_{(\mu_0,\mu_1)})_{[\boldsymbol{A}_n]}$ can be identified with $64\left(\boldsymbol{A}_n - \frac{1}{2}\int xy^\intercal d\pi_{\boldsymbol{A}_n}(x,y)\right)$ where $\pi_{\boldsymbol{A}_n}$ is any optimal plan for $\mathsf{OT}_{\boldsymbol{A}_n}(\mu_0,\mu_1)$ (by assumption all such cross-correlation matrices are identical). As $\boldsymbol{A}_n \to \boldsymbol{A}$, it follows from the proof of Lemma 2 that $\pi_{\boldsymbol{A}_n} \overset{w}{\to} \pi_{\boldsymbol{A}}$ up to a subsequence, where $\pi_{\boldsymbol{A}}$ is some optimal plan for $\mathsf{OT}_{\boldsymbol{A}}(\mu_0,\mu_1)$. It follows that $\partial\Phi_{(\mu_0,\mu_1)}(\boldsymbol{A}) = 64\boldsymbol{A} - 32\operatorname{conv}\left(\left\{\int xy^\intercal d\pi_{\boldsymbol{A}}(x,y) : \exists\, \pi_{\boldsymbol{A}_n} \overset{w}{\to} \pi_{\boldsymbol{A}}\right\}\right)$ such that $\partial\Phi_{(\mu_0,\mu_1)}(\boldsymbol{A}) = D\left(\Phi_{(\mu_0,\mu_1)}\right)_{[\boldsymbol{A}]}$ for $\boldsymbol{A} \in \Lambda$. If $\boldsymbol{A} \notin \Lambda$, the previous convex hull is simply a subset of all cross-correlation matrices for some optimal plan for $\mathsf{OT}_{\boldsymbol{A}}(\mu_0,\mu_1)$. By Proposition 2.3.2 in [11], if $\bar{\boldsymbol{A}}$ is a local minimizer for $\Phi_{(\mu_0,\mu_1)}$, then $0 \in \partial\Phi_{(\mu_0,\mu_1)}(\bar{\boldsymbol{A}})$; in any case there must exist an optimal plan $\pi_{\bar{\boldsymbol{A}}}$ for $\mathsf{OT}_{\bar{\boldsymbol{A}}}(\mu_0,\mu_1)$ satisfying $2\bar{\boldsymbol{A}} = \int xy^\intercal d\pi_{\bar{\boldsymbol{A}}}(x,y) \in B_F(\sqrt{M_2(\mu_0)M_2(\mu_1)})$ by Lemma 4. As the global minimum is known to be attained, at least one local minimizer is globally optimal.

Now, assume that μ_0, μ_1 are centered. It is straightforward to see that if $\boldsymbol{A}^\star$ solves (6), then the associated optimal plan $\pi_{\boldsymbol{A}^\star}$ satisfies

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

such that $\pi_{\boldsymbol{A}^\star}$ is optimal for (4) (see Section 5.3 in [57] for details). □

A.2 Proof of Proposition 1

It suffices to show that $c_{\boldsymbol{A}}(\cdot, y^{(i)}) - c_{\boldsymbol{A}}(\cdot, y^{(j)})$ is nonconstant on any set of positive (Lebesgue) measure. To this end, for any $x \in \mathcal{X}$ and $i \neq j$, observe that

$$\nabla_x \left(c_{\boldsymbol{A}}(x, y^{(i)}) - c_{\boldsymbol{A}}(x, y^{(j)})\right) = -8x\left(\|y^{(i)}\|^2 - \|y^{(j)}\|^2\right) - 32\boldsymbol{A}(y^{(i)} - y^{(j)}).$$

As such, if $\|y^{(i)}\| \neq \|y^{(j)}\|$, $c_{\boldsymbol{A}}(x, y^{(i)}) - c_{\boldsymbol{A}}(x, y^{(j)})$ is constant on at most a set of measure zero, as its gradient vanishes on at most one point. If $\|y^{(i)}\| = \|y^{(j)}\|$, $c_{\boldsymbol{A}}(\cdot, y^{(i)}) \not\equiv c_{\boldsymbol{A}}(\cdot, y^{(j)})$ on $\mathcal{X}$ if and only if $y^{(i)} - y^{(j)} \notin \ker(\boldsymbol{A})$. □

A.3 Proof of Proposition 2

We first address uniqueness of the optimal plan. For any $z \in \mathbb{R}^N$, let $\phi_z : x \in \mathcal{X} \mapsto \min_{1 \leq i \leq N}\left\{c_{\boldsymbol{A}}(x, y^{(i)}) - z_i\right\}$. If $x \in \mathcal{X}$ is such that $\phi_z(x) + \phi_z^c(y^{(i)}) = c_{\boldsymbol{A}}(x, y^{(i)})$ and $\phi_z(x) + \phi_z^c(y^{(j)}) = c_{\boldsymbol{A}}(x, y^{(j)})$ for some $i \neq j$, then $c_{\boldsymbol{A}}(x, y^{(j)}) - c_{\boldsymbol{A}}(x, y^{(i)}) = \phi_z^c(y^{(j)}) - \phi_z^c(y^{(i)})$.

By Assumption 1, the previous equality occurs on a μ_0-negligible set. It follows from Theorem 5.30 in [53] that the optimal plan for $\mathsf{OT}_{\boldsymbol{A}}(\mu_0, \mu_1)$ is unique and induced by a map.

We now derive the expression for the optimal map. Let $\zeta(y^{(i)}) = z_i^{\boldsymbol{A}}$ for $i = 1, \dots, N$. By (2) and strong duality,

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

As $\phi_{z^{\boldsymbol{A}}}(x) + \zeta(y) \leq c_{\boldsymbol{A}}(x, y)$ for $(x, y) \in \mathcal{X} \times \mathcal{Y}$ by definition, it follows that $\phi_{z^{\boldsymbol{A}}}(x) + \zeta(y) = c_{\boldsymbol{A}}(x, y)$ $\pi_{\boldsymbol{A}}$ almost surely. As $\pi_{\boldsymbol{A}}$ is induced by a map $T_{\boldsymbol{A}} : \mathcal{X} \to \mathcal{Y}$, $\phi_{z^{\boldsymbol{A}}}(x) = c_{\boldsymbol{A}}(x, T_{\boldsymbol{A}}(x)) - \zeta(T_{\boldsymbol{A}}(x))$ μ_0 almost everywhere. However, $\phi_{z^{\boldsymbol{A}}}(x) = \min_{1 \leq i \leq N} \left\{ c_{\boldsymbol{A}}(x, y^{(i)}) - \zeta(y^{(i)}) \right\}$, so the prior equality can only occur if $T_{\boldsymbol{A}}(x) = y^{(I_{z^{\boldsymbol{A}}}(x))}$; by Assumption 1, $I_{z^{\boldsymbol{A}}}(x)$ is a singleton for μ_0 almost every $x \in \mathcal{X}$. □

A.4 Proof of Lemma 1

For any $\boldsymbol{A}, \boldsymbol{A}' \in B_F(M)$ and $z, z' \in \mathbb{R}^N$ with $\|z\|_\infty \vee \|z'\|_\infty \leq K$,

$$\begin{aligned}\left| \min_{1 \leq i \leq N} \left\{ c_{\boldsymbol{A}}(x, y^{(i)}) - z_i \right\} - \min_{1 \leq i \leq N} \left\{ c_{\boldsymbol{A}'}(x, y^{(i)}) - z_i' \right\} \right| &\leq \max_{1 \leq i \leq N} | - 32 x^\intercal (\boldsymbol{A} - \boldsymbol{A}') y^{(i)} - (z_i - z_i')| \\ &\leq C(\|z - z'\|_\infty + \|\boldsymbol{A} - \boldsymbol{A}'\|_F),\end{aligned}$$

for $C = 1 \vee 32 \|\mathcal{X}\|_\infty \|\mathcal{Y}\|_\infty$. Let $T = \left\{ (z, \boldsymbol{A}) \in \mathbb{R}^N \times B_F(M) : \|z\|_\infty \leq K \right\}$ and identify it with a subset of $\mathbb{R}^{N + d_0 d_1}$ via $\tau : (z, \boldsymbol{A}) \in T \mapsto (z, A_{(\cdot)1}, \dots, A_{(\cdot)d1})$. Let $\|\tau(z, \boldsymbol{A})\|_{\tau T} = \|z\|_\infty + \|(A_{(\cdot)1}, \cdots, A_{(\cdot)d_1})\|$. By Theorem 2.7.11 in [50], $N_{[\,]}(2C\epsilon, \mathcal{G}_{0,K}, L^2(\mu_0)) \leq N(\epsilon, \tau T, \|\cdot\|_{\tau T})$.

We now upper bound the covering number of τT. Observe that $\tau_T \subset B_{\tau T}(K + M)$, the closed ball in $\mathbb{R}^{N + d_0 d_1}$ of radius $K + M$ with respect to $\|\cdot\|_{\tau T}$. The following argument is standard (cf. e.g. Lemma 4.14 in [33]). Let S_ϵ be an ϵ-net for $B_{\tau T}(K + M)$ with respect to $\|\cdot\|_{\tau T}$. By definition, $\cup_{x \in S_\epsilon} \left(x + \frac{\epsilon}{2} B_{\tau T}(1) \right) \subset \left(K + M + \frac{\epsilon}{2} \right) B_{\tau T}(1)$ and, as the sets on the left hand side are disjoint, $|S_\epsilon| \left(\frac{\epsilon}{2} \right)^{N + d_0 d_1} \operatorname{vol}\left(B_{\tau T}(1) \right) \leq \left(K + M + \frac{\epsilon}{2} \right)^{N + d_0 d_1} \operatorname{vol}\left(B_{\tau T}(1) \right)$ i.e. $|S_\epsilon| \leq \left(\frac{2(K+M)}{\epsilon} + 1 \right)^{N + d_0 d_1}$. It follows that $N(\epsilon, \tau T, \|\cdot\|_{\tau T}) \leq \left(\frac{2(K+M)}{\epsilon} \mathbb{1}_{\{\epsilon \leq K + M\}} + 1 \right)^{N + d_0 d_1}$.

We now provide an envelope for the class $\mathcal{G}_{0,K}$. For any $j \in \{1, \dots, N\}$,

$$\begin{aligned}&\min_{1 \leq i \leq N} \left\{ -8\|x\|^2 \|y^{(i)}\|^2 + 32 \|x\| \|\boldsymbol{A}\|_F \|y^{(i)}\| - z_i \right\} \\ &\qquad \leq \min_{1 \leq i \leq N} \left\{ c_{\boldsymbol{A}}(x, y^{(i)}) - z_i \right\} \leq -8\|x\|^2 \|y^{(j)}\|^2 + 32 \|x\| \|\boldsymbol{A}\|_F \|y^{(j)}\| - z_j.\end{aligned}$$

Letting, $\left| \min_{1 \leq i \leq N} \left\{ c_{\boldsymbol{A}}(x, y^{(i)}) - z_i \right\} \right| \leq 8\|x\|^2 \|\mathcal{Y}\|_\infty^2 + 32 M \|\mathcal{Y}\|_\infty \|x\| + K = F(x)$ such that $F(x)$ serves as an envelope for $\mathcal{G}_{0,K}$.

By Theorem A.2. in [49],

$$\begin{aligned}&\sqrt{n} \mathbb{E}\left[\|\hat{\mu}_{0,n} - \mu_0\|_{\infty, \mathcal{G}_{0,K}} \right] \\ &\lesssim \|F\|_{L^2(\mu_0)} \int_0^1 \sqrt{1 + \log N_{[\,]}(\epsilon \|F\|_{L^2(\mu_0)}, \mathcal{G}_{0,K}, L^2(\mu_0))} d\epsilon, \\ &\leq 2C(K + M) \int_0^{\frac{\|F\|_{L^2(\mu_0)}}{2C(K+M)}} \sqrt{1 + (N + d_0 d_1) \log \left(\frac{2}{\epsilon} \mathbb{1}_{\{\epsilon \leq 1\}} + 1 \right)} d\epsilon, \\ &\leq \|F\|_{L^2(\mu_0)} + 2C(K + M) \sqrt{N + d_0 d_1} C',\end{aligned}$$

for $C' = \int_0^\infty \sqrt{\log\left(\frac{2}{\epsilon} \mathbb{1}_{\{\epsilon \leq 1\}} + 1 \right)} d\epsilon < \infty$, where the first inequality holds up to a universal constant. Consequently, $\sqrt{n} \left[\mathbb{E} \|\hat{\mu}_{0,n} - \mu_0\|_{\infty, \mathcal{G}_{0,K}} \right] \lesssim_{K, \mathcal{Y}, \mathcal{X}} (N + d_0 d_1)^{1/2}$ proving the claim. Finiteness of the first integral in the display implies μ_0-Donskerness of $\mathcal{G}_{0,K}$ by Theorem 3.1 in [37].

Next, let $f, g \in \mathcal{G}_{1,K}$. Then, for any $\nu \in \mathcal{P}(\mathcal{Y})$,

$$\|f-g\|_{L^2(\nu)}^2 = \sum_{i=1}^N \left(f(y^{(i)}) - g(y^{(i)})\right)^2 \nu(\{y^{(i)}\}) \leq \max_{1\leq i\leq N} \left(f(y^{(i)}) - g(y^{(i)})\right)^2 \sum_{i=1}^N \nu(\{y^{(i)}\}),$$

that is, $\|f-g\|_{L^2(\nu)} \leq \|z_f - z_g\|_\infty$, where $z_f = \left(f\left(y^{(1)}\right), \ldots, f\left(y^{(N)}\right)\right)$ satisfies $\|z_f\|_\infty \leq K$ and z_g is defined analogously. Thus, $N(\epsilon, \mathcal{G}_{1,K}, L^2(\nu)) \leq N(\epsilon, \{\|\cdot\|_\infty \leq K\}, \|\cdot\|_\infty) \leq \left(\frac{2K}{\epsilon}\mathbb{1}_{\{\epsilon\leq K\}} + 1\right)^N$ as before and, from Equation 2 in [49],

$$\begin{aligned}
\sqrt{n}\mathbb{E}\left[\|\hat{\mu}_{1,n} - \mu_1\|_{\infty,\mathcal{G}_{1,K}}\right] &\lesssim K \int_0^1 \sup_{\nu\in\mathcal{P}(\mathcal{Y})} \sqrt{1 + \log N(\epsilon K, \mathcal{G}_{1,K}, L^2(\nu))} d\epsilon, \\
&\leq \int_0^K \sqrt{1 + N\log\left(\frac{2}{\epsilon}\mathbb{1}_{\{\epsilon\leq 1\}} + 1\right)} d\epsilon, \\
&\leq K + \sqrt{N}C',
\end{aligned}$$

such that $\sqrt{n}\left[\mathbb{E}\|\hat{\mu}_{1,n} - \mu_1\|_{\infty,\mathcal{G}_{1,K}}\right] \lesssim_K N^{1/2}$. Again, finiteness of the above integral implies that $\mathcal{G}_{1,K}$ is μ_1-Donsker (see Theorem 2.5.2 in [50]). □

A.5 Proof of Theorem 3

Assume without loss of generality that μ_0, μ_1 are centered. By equations (26)-(28) in the proof of Theorem 3 in [57],

$$\mathbb{E}\left[\left|\mathsf{D}(\hat{\mu}_{0,n}, \hat{\mu}_{1,n})^2 - \mathsf{D}(\mu_0,\mu_1)^2\right|\right] \lesssim 2R^4 n^{-1/2} + \mathbb{E}\left[\left|\mathsf{D}_2(\hat{\mu}_{0,n}, \hat{\mu}_{1,n}) - \mathsf{D}_2(\mu_0,\mu_1)\right|\right],$$

up to universal constants. By the variational formulation of the SDGW problem (5) and Theorem 1,

$$\left|\mathsf{D}_2(\hat{\mu}_{0,n}, \hat{\mu}_{1,n}) - \mathsf{D}_2(\mu_0,\mu_1)\right| \leq \sup_{\boldsymbol{A}\in B_F(M)} \left|\mathsf{OT}_{\boldsymbol{A}}(\hat{\mu}_{0,n}, \hat{\mu}_{1,n}) - \mathsf{OT}_{\boldsymbol{A}}(\mu_0,\mu_1)\right|.$$

For any $\boldsymbol{A} \in B_F(M)$, let $z^{\boldsymbol{A}}, z^{\boldsymbol{A},n}$ be optimal vectors for $\mathsf{OT}_{\boldsymbol{A}}(\mu_0,\mu_1)$ and $\mathsf{OT}_{\boldsymbol{A}}(\hat{\mu}_{0,n}, \hat{\mu}_{1,n})$ respectively with $\|z^{\boldsymbol{A}}\| \vee \|z^{\boldsymbol{A},n}\| \leq K$. Let $\psi^{\boldsymbol{A}}(y^{(i)}) = z_i^{\boldsymbol{A}}$, $\psi_n^{\boldsymbol{A}}(y^{(i)}) = z_i^{\boldsymbol{A},n}$ for $i = 1, \ldots, N$. Then,

$$\begin{aligned}
\mathsf{OT}_{\boldsymbol{A}}(\hat{\mu}_{0,n}, \hat{\mu}_{1,n}) - \mathsf{OT}_{\boldsymbol{A}}(\mu_0,\mu_1) &\leq \int (\psi_n^{\boldsymbol{A}})^{c_{\boldsymbol{A}}} d(\hat{\mu}_{0,n} - \mu_0) + \int \psi_n^{\boldsymbol{A}} d(\hat{\mu}_{1,n} - \mu_1), \\
\mathsf{OT}_{\boldsymbol{A}}(\hat{\mu}_{0,n}, \hat{\mu}_{1,n}) - \mathsf{OT}_{\boldsymbol{A}}(\mu_0,\mu_1) &\geq \int (\psi^{\boldsymbol{A}})^{c_{\boldsymbol{A}}} d(\hat{\mu}_{0,n} - \mu_0) + \int \psi^{\boldsymbol{A}} d(\hat{\mu}_{1,n} - \mu_1).
\end{aligned}$$

As $\psi^{\boldsymbol{A}}, \psi_n^{\boldsymbol{A}} \in \mathcal{G}_{1,K}$ and $(\psi^{\boldsymbol{A}})^{c_{\boldsymbol{A}}}, (\psi_n^{\boldsymbol{A}})^{c_{\boldsymbol{A}}} \in \mathcal{G}_{0,K}$, $\left|\mathsf{OT}_{\boldsymbol{A}}(\hat{\mu}_{0,n}, \hat{\mu}_{1,n}) - \mathsf{OT}_{\boldsymbol{A}}(\mu_0,\mu_1)\right| \leq \|\hat{\mu}_{0,n} - \mu_0\|_{\infty,\mathcal{G}_{0,K}} + \|\hat{\mu}_{1,n} - \mu_1\|_{\infty,\mathcal{G}_{1,K}}$. Conclude from Lemma 1 that

$$\mathbb{E}\left[\left|\mathsf{D}(\hat{\mu}_{0,n}, \hat{\mu}_{1,n})^2 - \mathsf{D}(\mu_0,\mu_1)^2\right|\right] \lesssim_{K,\mathcal{X},\mathcal{Y}} \left(2R^4 + \sqrt{N + d_0 d_1} + \sqrt{N}\right) n^{-1/2}.$$

The claimed result follows from the inequality $|x-y| \leq y^{-1}|x^2 - y^2|$ for $x \geq 0, y > 0$; noting that $\mathsf{D}(\mu_0,\mu_1)$ nullifies if and only if μ_0 and μ_1 can be related by an isometric map. □

A.6 Proof of Proposition 3

For 1, concavity follows from the fact that $\varphi^{\boldsymbol{A}}$ and $\psi^{\boldsymbol{A}}$ are pointwise infima of concave functions (cf. e.g. Theorem 5.5 in [39]). As for Lipschitz continuity, if $x, x' \in \mathcal{X}^\circ$ and

$$|\varphi^{\boldsymbol{A}}(x) - \varphi^{\boldsymbol{A}}(x')| \leq \max_{1\leq i\leq N} \left|c_{\boldsymbol{A}}\left(x, y^{(i)}\right) - c_{\boldsymbol{A}}\left(x', y^{(i)}\right)\right|.$$

Further, $\left|c_{\boldsymbol{A}}\left(x, \bar{y}^{(i)}\right) - c_{\boldsymbol{A}}\left(x', \bar{y}^{(i)}\right)\right| \leq 4\|\mathcal{Y}^\circ\|_\infty^2 \left|\|x\|^2 - \|x'\|^2\right| + 32\|\mathcal{Y}^\circ\|_\infty M\|x - x'\|$, and $\left|\|x\|^2 - \|x'\|^2\right| = \left|\|x\| - \|x'\|\right| (\|x\| + \|x'\|) \leq 2\|\mathcal{X}^\circ\|_\infty \|x - x'\|$, proving Lipschitz continuity with a constant which is independent of $\boldsymbol{A}, \nu_0, \nu_1$. A similar argument applies to $\psi^{\boldsymbol{A}}$. The maximum of the two Lipschitz constants serves as a shared Lipschitz constant.

As for 2, it follows from Rademacher's theorem and (1) (cf. e.g. Theorem 9.60 in [40]) that $\varphi^{\boldsymbol{A}}$ is differentiable a.e. on $\mathcal{X}^\circ$. On the other hand, $\varphi^{\boldsymbol{A}}$ is differentiable at $x \in \mathcal{X}^\circ$ if and only $\operatorname{conv}\left(\left\{\nabla_x c_{\boldsymbol{A}}(x, \bar{y}^{(i)}) : i \in \operatorname{argmin}_{1\le i\le N}\left(c_{\boldsymbol{A}}(x, \bar{y}^{(i)}) - z_i^{\boldsymbol{A}}\right)\right\}\right)$ is a singleton by Danskin's theorem (see Proposition B.25 in [4] and Theorem 25.1 in [39], observing that $c_{\boldsymbol{A}}(\cdot, \bar{y}^{(i)})$ is concave). As $\nabla_x c_{\boldsymbol{A}}(x, \bar{y}^{(i)}) = -8x\|\bar{y}^{(i)}\|^2 - 32\boldsymbol{A}\bar{y}^{(i)}$, $\nabla_x c_{\boldsymbol{A}}(x, \bar{y}^{(i)}) = \nabla_x c_{\boldsymbol{A}}(x, \bar{y}^{(j)})$ for at most one x if $\|\bar{y}^{(i)}\| \neq \|\bar{y}^{(j)}\|$ for $i \neq j$ (namely $x = 4\boldsymbol{A}\frac{\left(\bar{y}^{(i)} - \bar{y}^{(j)}\right)}{\|\bar{y}^{(j)}\|^2 - \|\bar{y}^{(i)}\|^2}$) and at no points x if $\|\bar{y}^{(i)}\| = \|\bar{y}^{(j)}\|$ for $i \neq j$ as $\bar{y}^{(i)} - \bar{y}^{(j)} \notin \ker(\boldsymbol{A})$ by Proposition 1. As such, $(\Lambda_j)_{j=1}^N$ must partition $\mathcal{X}^\circ$ up to a negligible set and $\nabla\varphi^{\boldsymbol{A}} = \nabla_x c_{\boldsymbol{A}}(x, \bar{y}^{(j)})$ on Λ_j. It is easy to see that Λ_j is an open set such that $D^2\varphi^{\boldsymbol{A}}$ can be computed classically on Λ_j.

To prove 3, by optimality of $z^{\boldsymbol{A}}$, $\varphi^{\boldsymbol{A}}(x) + z_i^{\boldsymbol{A}} = c_{\boldsymbol{A}}(x, \bar{y}^{(i)})$ $\pi_{\boldsymbol{A}}$ almost surely for any optimal coupling $\pi_{\boldsymbol{A}}$ for $\mathsf{OT}_{\boldsymbol{A}}(\nu_0, \nu_1)$ as in the proof of Proposition 2. Thus, there exists $x^{(i)} \in \mathcal{X}^\circ$ for which $\varphi^{\boldsymbol{A}}(x^{(i)}) + z_i^{\boldsymbol{A}} = c_{\boldsymbol{A}}(x^{(i)}, \bar{y}^{(i)})$ for $i = 1, \dots, N$. Hence,

$$\psi^{\boldsymbol{A}}(\bar{y}^{(i)}) = \inf_{\mathcal{X}^\circ}\left\{c_{\boldsymbol{A}}(x, y^{(i)}) - \varphi^{\boldsymbol{A}}(x)\right\} \ge z_i^{\boldsymbol{A}},$$

and this lower bound is attained at $x^{(i)}$, so $\psi^{\boldsymbol{A}}(\bar{y}^{(i)}) = z_i^{\boldsymbol{A}}$ for $i = 1, \dots, N$. For the second part,

$$-\|c_{\boldsymbol{A}}\|_{\infty,\mathcal{X}^\circ\times\mathcal{Y}^\circ} - \max_{1\le i\le N} z_i^{\boldsymbol{A}} \le \min_{1\le i\le N}\left\{c_{\boldsymbol{A}}(x, \bar{y}^{(i)}) - z_i^{\boldsymbol{A}}\right\} \le \|c_{\boldsymbol{A}}\|_{\infty,\mathcal{X}^\circ\times\mathcal{Y}^\circ} - \max_{1\le i\le N} z_i^{\boldsymbol{A}},$$
$$-\|c_{\boldsymbol{A}}\|_{\infty,\mathcal{X}^\circ\times\mathcal{Y}^\circ} - \sup_{\mathcal{X}^\circ} \varphi^{\boldsymbol{A}} \le \inf_{\mathcal{X}^\circ}\left(c_{\boldsymbol{A}}(\cdot, y) - \varphi^{\boldsymbol{A}}\right) \le \|c_{\boldsymbol{A}}\|_{\infty,\mathcal{X}^\circ\times\mathcal{Y}^\circ} - \sup_{\mathcal{X}^\circ} \varphi^{\boldsymbol{A}},$$

as $z^{\boldsymbol{A}} + a$ and its c-transform are also optimal for the dual problem, we may assume that $\max_{1\le i\le N} z_i^{\boldsymbol{A}} = 0$ such that $|\varphi^{\boldsymbol{A}}| \le \|c_{\boldsymbol{A}}\|_{\infty,\mathcal{X}^\circ\times\mathcal{Y}^\circ}$ on $\mathcal{X}^\circ$ by the top equation, and, from the bottom equation, $-2\|c_{\boldsymbol{A}}\|_{\infty,\mathcal{X}^\circ\times\mathcal{Y}^\circ} \le \psi^{\boldsymbol{A}} \le \|c_{\boldsymbol{A}}\|_{\infty,\mathcal{X}^\circ\times\mathcal{Y}^\circ}$ on $\mathcal{Y}^\circ$. As in 1, $\|c_{\boldsymbol{A}}\|_{\infty,\mathcal{X}^\circ\times\mathcal{Y}^\circ}$ can be bounded by a constant which does not depend on $\boldsymbol{A}$.

For 4, let $\pi^{\boldsymbol{A}}$ be the unique optimal plan for $\mathsf{OT}_{\boldsymbol{A}}(\nu_0, \nu_1)$ (see Proposition 2). As before, $\varphi^{\boldsymbol{A}}(x) = c_{\boldsymbol{A}}(x, y) - \psi^{\boldsymbol{A}}(y)$ $\pi^{\boldsymbol{A}}$-almost everywhere. As $\varphi^{\boldsymbol{A}}$ is differentiable a.e. on $\mathcal{X}^\circ$, $\nabla\varphi^{\boldsymbol{A}}(x) = \nabla_x c_{\boldsymbol{A}}(x, y)$ $\pi^{\boldsymbol{A}}$-almost everywhere. As $\operatorname{spt}(\bar{\mu}_0)$ has negligible boundary, $\nabla\varphi^{\boldsymbol{A}}$ is uniquely defined (up to a negligible set) on $\operatorname{int}(\operatorname{spt}(\bar{\mu}_0))$ which is connected, hence by a simple adaption of Theorem 2.6 in [14], any other extended potentials $(\tilde{\varphi}^{\boldsymbol{A}}, \tilde{\psi}^{\boldsymbol{A}})$ must satisfy $\nabla\varphi^{\boldsymbol{A}} = \nabla\tilde{\varphi}^{\boldsymbol{A}}$ almost everywhere on $\operatorname{int}(\operatorname{spt}(\bar{\mu}_0))$, so $\tilde{\varphi}^{\boldsymbol{A}} = \varphi^{\boldsymbol{A}} + a$ on $\operatorname{int}(\operatorname{spt}(\bar{\mu}_0))$ for some $a \in \mathbb{R}$. By the previous deliberations, for $i = 1, \dots, N$, $\tilde{\psi}^{\boldsymbol{A}}(\bar{y}^{(i)}) = c_{\boldsymbol{A}}(x, \bar{y}^{(i)}) - \tilde{\varphi}^{\boldsymbol{A}}(x) = c_{\boldsymbol{A}}(x, \bar{y}^{(i)}) - (\varphi^{\boldsymbol{A}}(x) + a) = \psi^{\boldsymbol{A}}(\bar{y}^{(i)}) - a$ on a set of positive μ_0 measure. □

A.7 Proof of Theorem 4

The proof of Theorem 4 is broken down as follows. Appendix A.7.1 proves Hadamard directional differentiability of $\mathsf{D}_1(\bar{\nu}_0, \bar{\nu}_1)$ at $\mu_0 \otimes \mu_1$ and Appendix A.8 proves differentiability of $\mathsf{D}_2(\bar{\nu}_0, \bar{\nu}_1)$ at $\mu_0 \otimes \mu_1$. Together, this shows that $\mathsf{D}(\bar{\nu}_0, \bar{\nu}_1)^2 = \mathsf{D}_1(\bar{\nu}_0, \bar{\nu}_1) + \mathsf{S}_2(\bar{\nu}_0, \bar{\nu}_1)$ is differentiable at $\mu_0 \otimes \mu_1$.

A.7.1 Hadamard derivative of D_1

The main result of this section is the following.

Proposition 4. *The map $\nu_0 \otimes \nu_1 \in \mathfrak{P} \mapsto \mathsf{D}_1(\bar{\nu}_0, \bar{\nu}_1)$ is Hadamard directionally differentiable at $\mu_0 \otimes \mu_1$ with derivative*

$$\eta \in \mathcal{T}_{\mathfrak{P}}(\mu_0 \otimes \mu_1) \mapsto \eta\left(2\int \|x - \cdot\|^4 d\mu_0(x) - 4M_2(\bar{\mu}_1)\| \cdot -\mathbb{E}_{\mu_0}[X]\|^2 \right.$$
$$\left. \oplus\, 2\int \|y - \cdot\|^4 d\mu_1(y) - 4M_2(\bar{\mu}_0)\| \cdot -\mathbb{E}_{\mu_1}[X]\|^2\right).$$

To prove Proposition 4, we first compute the Gâteaux directional derivative and subsequently show Lipschitz continuity.

Lemma 5. *Fix $\nu_0 \otimes \nu_1 \in \mathfrak{P}$. Let $\mu_{i,t} := \mu_i + t(\nu_i - \mu_i)$ for $i = 0, 1$ and $t \in [0, 1]$. Then,*

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

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 $i = 0, 1$,

$$\begin{aligned}&\iint \|x - x'\|^4 d\bar{\mu}_{i,t}(x) d\bar{\mu}_{i,t}(x') - \iint \|x - x'\|^4 d\bar{\mu}_i(x) d\bar{\mu}_i(x') \\ &= 2t \iint \|x - x'\|^4 d\mu_i(x) d(\nu_i - \mu_i)(x') + t^2 \iint \|x - x'\|^4 d(\nu_i - \mu_i)(x) d(\nu_i - \mu_i)(x').\end{aligned} \tag{11}$$

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 $t \in [0, 1]$,

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

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, combined with (11) prove the claim.

□

Lemma 6. *For any $\nu_0 \otimes \nu_1, \rho_0 \otimes \rho_1 \in \mathfrak{P}$, there exists a universal finite constant C for which $|\mathsf{D}_1(\bar{\nu}_0, \bar{\nu}_1) - \mathsf{D}_1(\bar{\rho}_0, \bar{\rho}_1)| \le C \|\nu_0 \otimes \nu_1 - \rho_0 \otimes \rho_1\|_{\infty, \mathcal{F}_K^{\oplus}}$.*

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

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

As the functions $x \in \mathcal{X} \mapsto \int \|x-x'\|^4 d\nu_0(x'), x \in \mathcal{X} \mapsto \int \|x-x'\|^4 d\rho_0(x')$ are smooth with uniformly bounded derivatives of all orders (for some constant depending only on $\mathcal{X}$) and $\|\int \|\cdot - y'\|^4 d\nu_1(y')\|_{\infty,\mathcal{Y}} \vee \|\int \|\cdot - y'\|^4 d\rho_1(y')\|_{\infty,\mathcal{Y}}$ is bounded by a universal constant, there exists a universal constant C_1 for which

$$\left|\sum_{i=0}^{1}\left(\iint \|x-x'\|^4 d\bar{\nu}_i(x)d\bar{\nu}_i(x') - \iint \|x-x'\|^4 d\bar{\rho}_i(x)d\bar{\rho}_i(x')\right)\right| \leq C_1\|\nu_0\otimes\nu_1 - \rho_0\otimes\rho_1\|_{\infty,\mathcal{F}_K^{\oplus}}. \tag{12}$$

Next, observe that

$$\iint \|x\|^2\|y\|^2 d\bar{\nu}_0(x)d\bar{\nu}_1(y) = \iint \|x-\mathbb{E}_{\nu_0}[X]\|^2\|y-\mathbb{E}_{\nu_1}[X]\|^2 d\nu_0(x)d\nu_1(y),$$

where, for $i=0,1$, $\|x-\mathbb{E}_{\nu_i}[X]\|^2 = \|x-\mathbb{E}_{\rho_i}[X]\|^2 + 2\langle x-\mathbb{E}_{\rho_i}[X], \mathbb{E}_{\rho_i}[X]-\mathbb{E}_{\nu_i}[X]\rangle + \|\mathbb{E}_{\rho_i}[X]-\mathbb{E}_{\nu_i}[X]\|^2$, such that $\int \|x-\mathbb{E}_{\nu_i}[X]\|^2 d\nu_i(x) = \int \|x-\mathbb{E}_{\rho_i}[X]\|^2 d\nu_i(x) - \|\mathbb{E}_{\rho_i}[X]-\mathbb{E}_{\nu_i}[X]\|^2$ and

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

Consequently,

$$\begin{aligned}
&\iint \|x\|^2\|y\|^2 d\bar{\nu}_0(x)d\bar{\nu}_1(y) - \iint \|x\|^2\|y\|^2 d\bar{\rho}_0(x)d\bar{\rho}_1(y)\\
&\qquad = \iint \|x-\mathbb{E}_{\rho_0}[X]\|^2\|y-\mathbb{E}_{\rho_1}[X]\|^2 d(\nu_0-\rho_0)(x)d\nu_1(y)\\
&\qquad + \iint \|x-\mathbb{E}_{\rho_0}[X]\|^2\|y-\mathbb{E}_{\rho_1}[X]\|^2 d\rho_0(x)d(\nu_1-\rho_1)(y)\\
&\qquad - \int \|x-\mathbb{E}_{\rho_0}[X]\|^2 d\nu_0(x)\|\mathbb{E}_{\nu_1}[X]-\mathbb{E}_{\rho_1}[X]\|^2\\
&\qquad - \int \|y-\mathbb{E}_{\rho_1}[X]\|^2 d\nu_1(y)\|\mathbb{E}_{\nu_0}[X]-\mathbb{E}_{\rho_0}[X]\|^2\\
&\qquad + \|\mathbb{E}_{\nu_0}[X]-\mathbb{E}_{\rho_0}[X]\|^2\|\mathbb{E}_{\nu_1}[X]-\mathbb{E}_{\rho_1}[X]\|^2.
\end{aligned} \tag{13}$$

As in the first part of the proof, $x \in \mathcal{X} \mapsto \|x-\mathbb{E}_{\rho_0}[X]\|^2$ is smooth with uniformly bounded derivatives of all orders independently of the choice of ρ_0 and $\|\|\cdot - \mathbb{E}_{\rho_1}[X]\|^2\|_{\infty,\mathcal{Y}}$ is bounded uniformly in the choice of ρ_1. Furthermore, $\|\mathbb{E}_{\nu_0}[X]-\mathbb{E}_{\rho_0}[X]\| \leq \sqrt{d_0}\|\mathbb{E}_{\nu_0}[X]-\mathbb{E}_{\rho_0}[X]\|_\infty \lesssim_{\mathcal{X}} \sqrt{d_0}\|\nu_0-\rho_0\|_{\infty,\mathcal{C}_K^{\lfloor \frac{d_0}{2}\rfloor+1}(\mathcal{X})}$ as the coordinate projections are evidently smooth with uniformly bounded derivatives on $\mathcal{X}$. Moreover, $\|\mathbb{E}_{\nu_1}[X]-\mathbb{E}_{\rho_1}[X]\| \leq \sqrt{d_1}\|\mathbb{E}_{\nu_1}[X]-\mathbb{E}_{\rho_1}[X]\|_\infty \lesssim_{\mathcal{Y}} \sqrt{d_1}\|\nu_1-\rho_1\|_{\infty,\mathcal{G}_{1,K}}$ as the coordinate projections are uniformly bounded on $\mathcal{Y}$. Since $\mathcal{C}_K^{\lfloor \frac{d_0}{2}\rfloor+1}(\mathcal{X})$ and $\mathcal{G}_{1,K}$ are symmetric (in the sense that $\mathcal{G}_{1,K} = -\mathcal{G}_{1,K}$) and contain 0, $\|\nu_0-\rho_0\|_{\infty,\mathcal{C}_K^{\lfloor \frac{d_0}{2}\rfloor+1}(\mathcal{X})} + \|\nu_1-\rho_1\|_{\infty,\mathcal{G}_{1,K}} \leq \|\nu_0\otimes\nu_1-\rho_0\otimes\rho_1\|_{\infty,\mathcal{F}_K^{\oplus}}$. Applying this bound to (13) and combining with (12) proves the claim. □

Proof of Proposition 4. The proof of Proposition 4 follows from Lemmas 5 and 6 by applying Proposition 7, noting that $x \in \mathcal{X} \mapsto 2\int \|x-x'\|^4 d\mu_0(x') - 4M_2(\bar{\mu}_1)\|x-\mathbb{E}_{\mu_0}[X]\|^2$ is smooth with uniformly bounded derivatives of all orders, and $y \in \mathcal{Y} \mapsto 2\int \|y-y'\|^4 d\mu_1(y') - 4M_2(\bar{\mu}_0)\|y-\mathbb{E}_{\mu_1}[X]\|^2$ is uniformly bounded. Therefore, these functions lie, respectively, in $\mathcal{F}_{0,K}$ and $\mathcal{G}_{1,K}$ for K sufficiently large. □

A.8 Hadamard derivative of D_2

Throughout, we always choose versions of the extended potentials which are uniformly bounded by Proposition 3. The main result of this section is as follows.

Proposition 5. *The functional $\nu_0 \otimes \nu_1 \in \mathfrak{P} \mapsto \mathsf{D}_2(\bar{\nu}_0, \bar{\nu}_1)$ is Hadamard directionally differentiable at $\mu_0 \otimes \mu_1$ with derivative*

$$\eta \in \mathcal{T}_{\mathfrak{P}}(\mu_0 \otimes \mu_1) \mapsto \inf_{\operatorname{argmin}_{\boldsymbol{A} \in B_F(M)} \left(\Phi_{(\bar{\mu}_0, \bar{\mu}_1)}\right)} \eta\Big(\varphi^{\boldsymbol{A}}(\cdot - \mathbb{E}_{\mu_0}[X]) - \mathbb{E}_{\bar{\mu}_0}[\nabla \varphi^{\boldsymbol{A}}(X)]^{\intercal}(\cdot) \\ \oplus \psi^{\boldsymbol{A}}(\cdot - \mathbb{E}_{\mu_1}[X]) - 8\mathbb{E}_{(X,Y)\sim\pi_{\boldsymbol{A}}}\left[\|X\|^2 Y\right]^{\intercal}(\cdot)\Big).$$

As aforementioned, it suffices to establish Hadamard directional differentiability of the functional $\nu_0 \otimes \nu_1 \in \mathfrak{P} \mapsto \mathsf{OT}_{(\cdot)}(\bar{\nu}_0 \otimes \bar{\nu}_1) \in \ell^\infty(B_F(M))$ such that differentiability of $\mathsf{D}_2(\bar{\mu}_0, \bar{\mu}_1)$ follows from the chain rule and a known result for differentiability of infimum-type functionals [6].

Lemma 7. *Fix $\nu_0 \otimes \nu_1 \in \mathfrak{P}$. Let $\mu_{i,t} := \mu_i + t(\nu_i - \mu_i)$ for $i = 0, 1$ and $t \in [0, 1]$. Then, the following limit holds in $\ell^\infty(B_F(M))$*

$$\begin{aligned} \frac{\mathsf{OT}_{(\cdot)}(\bar{\mu}_{0,t}, \bar{\mu}_{1,t}) - \mathsf{OT}_{(\cdot)}(\bar{\mu}_0, \bar{\mu}_1)}{t} \to & \int \varphi^{(\cdot)}(x - \mathbb{E}_{\mu_0}[X]) d(\nu_0 - \mu_0)(x) \\ & + \int \psi^{(\cdot)}(x - \mathbb{E}_{\mu_1}[X]) d(\nu_1 - \mu_1)(x) \\ & - \int \nabla \varphi^{(\cdot)}(x)^{\intercal}(\mathbb{E}_{\nu_0}[X] - \mathbb{E}_{\mu_0}[X]) d\bar{\mu}_0(x) \\ & - 8 \int \|x\|^2 \langle y, (\mathbb{E}_{\nu_1}[X] - \mathbb{E}_{\mu_1}[X])\rangle d\pi_{(\cdot)}(x, y), \end{aligned} \tag{14}$$

where, for $\boldsymbol{A} \in B_F(M)$, $(\varphi^{\boldsymbol{A}}, \psi^{\boldsymbol{A}})$ are extended potentials for $\mathsf{OT}_{\boldsymbol{A}}(\bar{\mu}_0, \bar{\mu}_1)$ and $\pi_{\boldsymbol{A}}$ is the unique optimal coupling.

To prove Lemma 7, we decompose $\mathsf{OT}_{\boldsymbol{A}}(\bar{\mu}_{0,t}, \bar{\mu}_{1,t}) - \mathsf{OT}_{\boldsymbol{A}}(\bar{\mu}_0, \bar{\mu}_1)$ as

$$\mathsf{OT}_{\boldsymbol{A}}(\bar{\mu}_{0,t}, \bar{\mu}_{1,t}) - \mathsf{OT}_{\boldsymbol{A}}(\bar{\mu}_0, \bar{\mu}_1) = \mathsf{OT}_{\boldsymbol{A}}(\bar{\mu}_{0,t}, \bar{\mu}_{1,t}) - \mathsf{OT}_{\boldsymbol{A}}(\bar{\mu}_{0,t}, \tilde{\mu}_{1,t}) + \mathsf{OT}_{\boldsymbol{A}}(\bar{\mu}_{0,t}, \tilde{\mu}_{1,t}) - \mathsf{OT}_{\boldsymbol{A}}(\bar{\mu}_0, \bar{\mu}_1),$$

where $\tilde{\mu}_{1,t} = (\mathrm{Id} - \mathbb{E}_{\mu_1}[X])_\sharp \, \mu_{1,t}$ and separately analyze the two differences.

Lemma 8. *Let $\pi_{\boldsymbol{A}}$ be an optimal coupling for $\mathsf{OT}_{\boldsymbol{A}}(\bar{\mu}_0, \bar{\mu}_1)$ for $\boldsymbol{A} \in B_F(M)$. Then,*

$$\frac{\mathsf{OT}_{(\cdot)}(\bar{\mu}_{0,t}, \bar{\mu}_{1,t}) - \mathsf{OT}_{(\cdot)}(\bar{\mu}_{0,t}, \tilde{\mu}_{1,t})}{t} \to -8 \int \|x\|^2 \langle y, (\mathbb{E}_{\nu_1}[X] - \mathbb{E}_{\mu_1}[X])\rangle d\pi_{(\cdot)}(x, y) \text{ in } \ell^\infty(B_F(M)).$$

Proof. For $t \in [-1, 1]$, let $\tau_t = \mathrm{Id} - t\,(\mathbb{E}_{\nu_1}[X] - \mathbb{E}_{\mu_1}[X])$ such that $\bar{\mu}_{1,t} = (\tau_t)_\sharp \tilde{\mu}_{1,t}$ and $\tilde{\mu}_{1,t} = (\tau_{-t})_\sharp \, \bar{\mu}_{1,t}$. Observe that if $\bar{\pi} \in \Pi\,(\bar{\mu}_{0,t}, \bar{\mu}_{1,t})$, then $(\mathrm{Id}, \tau_t)_\sharp \, \bar{\pi} \in \Pi(\bar{\mu}_{0,t}, \tilde{\mu}_{1,t})$ and, conversely, if $\tilde{\pi} \in \Pi\,(\bar{\mu}_{0,t}, \tilde{\mu}_{1,t})$, then $(\mathrm{Id}, \tau_{-t})_\sharp \tilde{\pi} \in \Pi\,(\bar{\mu}_{0,t}, \bar{\mu}_{1,t})$. Now, fix $\boldsymbol{A} \in B_F(M)$ and let $\bar{\pi}_{\boldsymbol{A},t}$ and $\tilde{\pi}_{\boldsymbol{A},t}$ be optimal plans for $\mathsf{OT}_{\boldsymbol{A}}(\bar{\mu}_{0,t}, \bar{\mu}_{1,t})$ and $\mathsf{OT}_{\boldsymbol{A}}(\bar{\mu}_{0,t}, \tilde{\mu}_{1,t})$ respectively. Then,

$$\mathsf{OT}_{\boldsymbol{A}}(\bar{\mu}_{0,t}, \bar{\mu}_{1,t}) - \mathsf{OT}_{\boldsymbol{A}}(\bar{\mu}_{0,t}, \tilde{\mu}_{1,t}) \geq \int c_{\boldsymbol{A}}(x, y) d\bar{\pi}_{\boldsymbol{A},t}(x, y) - \int c_{\boldsymbol{A}}(x, \tau_t(y)) d\bar{\pi}_{\boldsymbol{A},t}(x, y),$$

$$\mathsf{OT}_{\boldsymbol{A}}(\bar{\mu}_{0,t}, \bar{\mu}_{1,t}) - \mathsf{OT}_{\boldsymbol{A}}(\bar{\mu}_{0,t}, \tilde{\mu}_{1,t}) \leq \int c_{\boldsymbol{A}}(x, \tau_{-t}(y)) d\tilde{\pi}_{\boldsymbol{A},t}(x, y) - \int c_{\boldsymbol{A}}(x, y) d\tilde{\pi}_{\boldsymbol{A},t}(x, y).$$

Note that

$$\begin{aligned} c_{\boldsymbol{A}}(x, \tau_t(y)) &= -4\|x\|^2 \|y - t\,(\mathbb{E}_{\nu_1}[X] - \mathbb{E}_{\mu_1}[X])\,\|^2 - 32 x^{\intercal} \boldsymbol{A}(y - t\,(\mathbb{E}_{\nu_1}[X] - \mathbb{E}_{\mu_1}[X])) \\ &= 8t\|x\|^2 \langle y, (\mathbb{E}_{\nu_1}[X] - \mathbb{E}_{\mu_1}[X])\rangle - 4t^2 \|x\|^2 \|\mathbb{E}_{\nu_1}[X] - \mathbb{E}_{\mu_1}[X]\|^2 \\ &\quad + 32 t x^{\intercal} \boldsymbol{A}(\mathbb{E}_{\nu_1}[X] - \mathbb{E}_{\mu_1}[X]) + c_{\boldsymbol{A}}(x, y). \end{aligned}$$

As $\bar{\mu}_{0,t}$ is mean-zero,

$$\int c_{\boldsymbol{A}}(x, \tau_{-t}(y)) - c_{\boldsymbol{A}}(x,y) d\tilde{\pi}_{\boldsymbol{A},t}(x,y) = -8t \int \|x\|^2 \langle y, (\mathbb{E}_{\nu_1}[X] - \mathbb{E}_{\mu_1}[X])\rangle d\tilde{\pi}_{\boldsymbol{A},t}(x,y) - 4t^2 M_2(\bar{\mu}_{0,t}) \|\mathbb{E}_{\nu_1}[X] - \mathbb{E}_{\mu_1}[X]\|^2,$$

$$\int c_{\boldsymbol{A}}(x,y) - c_{\boldsymbol{A}}(x, \tau_t(y)) d\bar{\pi}_{\boldsymbol{A},t}(x,y) = -8t \int \|x\|^2 \langle y, (\mathbb{E}_{\nu_1}[X] - \mathbb{E}_{\mu_1}[X])\rangle d\bar{\pi}_{\boldsymbol{A},t}(x,y) + 4t^2 M_2(\bar{\mu}_{0,t}) \|\mathbb{E}_{\nu_1}[X] - \mathbb{E}_{\mu_1}[X]\|^2.$$

Observe that $\bar{\mu}_{0,t} \xrightarrow{w} \bar{\mu}_0, \bar{\mu}_{1,t} \xrightarrow{w} \bar{\mu}_1, \tilde{\mu}_{1,t} \xrightarrow{w} \bar{\mu}_1$ as $t \downarrow 0$, hence also $\tilde{\pi}_{\boldsymbol{A},t} \xrightarrow{w} \pi_{\boldsymbol{A}}$ and $\bar{\pi}_{\boldsymbol{A},t} \xrightarrow{w} \pi_{\boldsymbol{A}}$, where $\pi_{\boldsymbol{A}}$ is the unique optimal plan for $\mathsf{OT}_{\boldsymbol{A}}(\bar{\mu}_0, \bar{\mu}_1)$ (see Theorem 5.20 in [53]). Consider

$$\begin{aligned}
&\sup_{\boldsymbol{A} \in B_F(M)} \left| \frac{\mathsf{OT}_{\boldsymbol{A}}(\bar{\mu}_{0,t}, \bar{\mu}_{1,t}) - \mathsf{OT}_{\boldsymbol{A}}(\bar{\mu}_{0,t}, \tilde{\mu}_{1,t})}{t} + 8 \int \|x\|^2 \langle y, (\mathbb{E}_{\nu_1}[X] - \mathbb{E}_{\mu_1}[X])\rangle d\pi_{\boldsymbol{A}}(x,y) \right| \\
&\leq 8 \sup_{\boldsymbol{A} \in B_F(M)} \left| \int \|x\|^2 \langle y, (\mathbb{E}_{\nu_1}[X] - \mathbb{E}_{\mu_1}[X])\rangle d(\pi_{\boldsymbol{A}} - \tilde{\pi}_{\boldsymbol{A},t})(x,y) \right| \\
&+ 8 \sup_{\boldsymbol{A} \in B_F(M)} \left| \int \|x\|^2 \langle y, (\mathbb{E}_{\nu_1}[X] - \mathbb{E}_{\mu_1}[X])\rangle d(\pi_{\boldsymbol{A}} - \bar{\pi}_{\boldsymbol{A},t})(x,y) \right| \\
&+ 4t M_2(\bar{\mu}_{0,t}) \|\mathbb{E}_{\nu_1}[X] - \mathbb{E}_{\mu_1}[X]\|^2.
\end{aligned} \tag{15}$$

The final term on the right hand side evidently converges to 0 as $t \downarrow 0$. We now show that the first term converges to 0 as well; convergence of the second term to 0 follows by analogy.

Let $t_n \downarrow 0$ with $t_n \leq 1$ be arbitrary and fix $\epsilon > 0$. Let $\boldsymbol{A}_n \in B_F(M)$ be such that

$$\begin{aligned}
&\sup_{\boldsymbol{A} \in B_F(M)} \left| \int \|x\|^2 \langle y, (\mathbb{E}_{\nu_1}[X] - \mathbb{E}_{\mu_1}[X])\rangle d(\pi_{\boldsymbol{A}} - \tilde{\pi}_{\boldsymbol{A},t})(x,y) \right| \\
&\leq \left| \int \|x\|^2 \langle y, (\mathbb{E}_{\nu_1}[X] - \mathbb{E}_{\mu_1}[X])\rangle d(\pi_{\boldsymbol{A}_n} - \tilde{\pi}_{\boldsymbol{A}_n,t_n})(x,y) \right| + \epsilon.
\end{aligned}$$

For any subsequence n', there exists a further subsequence n'' along which $\boldsymbol{A}_{n''} \to \boldsymbol{A} \in B_F(M)$ by the Bolzano-Weierstrass theorem. As $c_{\boldsymbol{A}_{n''}} \to c_{\boldsymbol{A}}$ uniformly on compact sets, $\pi_{\boldsymbol{A}_{n''}} \xrightarrow{w} \pi_{\boldsymbol{A}}$ and $\tilde{\pi}_{\boldsymbol{A}_{n''}} \xrightarrow{w} \pi_{\boldsymbol{A}}$ by Theorem 5.20 in [53]. As such,

$$\int \|x\|^2 \langle y, (\mathbb{E}_{\nu_1}[X] - \mathbb{E}_{\mu_1}[X])\rangle d(\pi_{\boldsymbol{A}_{n''}} - \tilde{\pi}_{\boldsymbol{A}_{n''},t_{n''}})(x,y) \to 0,$$

and, as this limit is independent of the choice of subsequence, the convergence holds along the original sequence, yielding

$$\limsup_{t \downarrow 0} \sup_{\boldsymbol{A} \in B_F(M)} \left| \int \|x\|^2 \langle y, (\mathbb{E}_{\nu_1}[X] - \mathbb{E}_{\mu_1}[X])\rangle d(\pi_{\boldsymbol{A}} - \tilde{\pi}_{\boldsymbol{A},t})(x,y) \right| < \epsilon.$$

As ϵ is arbitrary, conclude that

$$\lim_{t \downarrow 0} \sup_{\boldsymbol{A} \in B_F(M)} \left| \int \|x\|^2 \langle y, (\mathbb{E}_{\nu_1}[X] - \mathbb{E}_{\mu_1}[X])\rangle d(\pi_{\boldsymbol{A}} - \tilde{\pi}_{\boldsymbol{A},t})(x,y) \right| = 0.$$

Similarly, the remaining term in (15) converges to 0 which concludes the proof. □

As for the second limit, we first establish an auxiliary lemma.

Lemma 9. *Let $(\varphi_t^{\boldsymbol{A}}, \psi_t^{\boldsymbol{A}})$ be extended potentials for $\mathsf{OT}_{\boldsymbol{A}}(\bar{\mu}_{0,t}, \tilde{\mu}_{1,t})$ for $t \in [0,1]$. For $s \in (-1,1)$,*

$$\int \varphi_t^{\boldsymbol{A}}(\cdot - s(\mathbb{E}_{\nu_0}[X] - \mathbb{E}_{\mu_0}[X])) d\bar{\mu}_0 - \int \varphi_t^{\boldsymbol{A}} d\bar{\mu}_0 + s \int \nabla \varphi_t^{\boldsymbol{A}}(\cdot)^\intercal (\mathbb{E}_{\nu_0}[X] - \mathbb{E}_{\mu_0}[X]) d\bar{\mu}_0 = r_{\boldsymbol{A},t}(s),$$

where $r_{\boldsymbol{A},t}(s) = O(s^2)$ as $s \to 0$ uniformly in the choice of t and $\boldsymbol{A}$.

Proof. Observe that $\mathrm{spt}(\tilde{\mu}_{1,t}) = \mathrm{spt}(\bar{\mu}_1) = \bar{\mathcal{Y}}$ such that Proposition 3 2 can be applied. Let Λ_j^t be the corresponding Λ_j sets for $t \in [0,1]$. For any $(x, s') \in \mathcal{X}^\circ \times (-1,1)$ satisfying $x - s'(\mathbb{E}_{\nu_0}[X] - \mathbb{E}_{\mu_0}[X]) \in \Lambda_j^t$, the map $g_{x,t} : s \in (-1,1) \mapsto \varphi_t^{\boldsymbol{A}}(x - s(\mathbb{E}_{\nu_0}[X] - \mathbb{E}_{\mu_0}[X]))$ is differentiable at s' with

$$\begin{aligned}
g'_{x,t}(s') &= -\nabla\varphi_t^{\boldsymbol{A}}(x - s'(\mathbb{E}_{\nu_0}[X] - \mathbb{E}_{\mu_0}[X]))^\intercal(\mathbb{E}_{\nu_0}[X] - \mathbb{E}_{\mu_0}[X]),\\
g''_{x,t}(s') &= (\mathbb{E}_{\nu_0}[X] - \mathbb{E}_{\mu_0}[X])^\intercal D^2\varphi_t^{\boldsymbol{A}}(x - s'(\mathbb{E}_{\nu_0}[X] - \mathbb{E}_{\mu_0}[X]))(\mathbb{E}_{\nu_0}[X] - \mathbb{E}_{\mu_0}[X]),\\
g'''_{x,t}(s') &= 0,
\end{aligned}$$

where the final equality follows from the fact that all derivatives of $\varphi_t^{\boldsymbol{A}}$ must be zero on Λ_j^t, as its Hessian is constant. For $s \in (-1,1)$, let $\Lambda_{j,s}^t = \left\{x \in \mathcal{X}^\circ : x - s(\mathbb{E}_{\nu_0}[X] - \mathbb{E}_{\mu_0}[X]) \in \Lambda_j^t\right\}$, then $\bar{\mu}_0\left(\cup_{j=1}^N \Lambda_{j,s}^t\right) = 1$. It readily follows that $h_t : s \in (-1,1) \mapsto \int g_{x,t}(s) d\bar{\mu}_0(x)$ satisfies (cf. e.g. Theorem 6.28 [29])

$$h'_t(s) = \int g'_{x,t}(s) d\bar{\mu}_0(x), \quad h''_t(s) = \int g''_{x,t}(s) d\bar{\mu}_0(x), \quad h'''_t(s) = 0.$$

Let $F_x(s) = x - s\left(\mathbb{E}_{\nu_0}[X] - \mathbb{E}_{\mu_0}[X]\right)$ and τ be sufficiently small that $s + \tau \in (-1,1)$, then

$$\begin{aligned}
h'_t(s+\tau) &= \sum_{j=1}^N \int_{\Lambda_{j,s+\tau}^t} g'_{x,t}(s+\tau) d\bar{\mu}_0(x)\\
&= \sum_{j=1}^N \int_{\Lambda_{j,s+\tau}^t} -\nabla\varphi_t^{\boldsymbol{A}}(F_x(s+\tau))^\intercal(\mathbb{E}_{\nu_0}[X] - \mathbb{E}_{\mu_0}[X]) d\bar{\mu}_0(x)\\
&= \sum_{j=1}^N \int \mathbb{1}_{\Lambda_{j,s+\tau}^t}(x) \left(8\|\bar{y}^{(j)}\|^2 F_x(s+\tau) + 32\boldsymbol{A}\bar{y}^{(j)}\right)^\intercal (\mathbb{E}_{\nu_0}[X] - \mathbb{E}_{\mu_0}[X]) d\bar{\mu}_0(x).
\end{aligned}$$

As $\mathbb{1}_{\Lambda_{j,s+\tau}^t}(x) \to \mathbb{1}_{\Lambda_{j,s}^t}(x)$ and $F_x(s+\tau) \to F_x(s)$ pointwise as $\tau \to 0$ (recall from the proof of Proposition 3 point 2 that Λ_j^t is an open set), it follows from the dominated convergence theorem that $h'_t(s)$ is continuous; the same argument applies to the higher order derivatives. Thus, we can apply a Taylor expansion to obtain that

$$h_t(s) - h_t(0) - sh'_t(0) = \frac{s^2}{2} h''_t(0),$$

as the third derivative vanishes. Observe that

$$|h''_t(0)| \leq \sum_{j=1}^N \int_{\Lambda_j^t} 8\|\bar{y}^{(j)}\|^2 \|\mathbb{E}_{\nu_0}[X] - \mathbb{E}_{\mu_0}[X]\|^2 d\bar{\mu}_0(x) \leq 8\,\mathrm{diam}(\mathcal{X})^2 \max_{j=1}^N \|y^{(j)}\|^2,$$

proving the claim. □

Lemma 10. *The following limit holds in $\ell^\infty(B_F(M))$,*

$$\begin{aligned}
\frac{\mathsf{OT}_{(\cdot)}(\bar{\mu}_{0,t}, \tilde{\mu}_{1,t}) - \mathsf{OT}_{(\cdot)}(\bar{\mu}_0, \bar{\mu}_1)}{t} \to &\int \varphi^{(\cdot)}(x - \mathbb{E}_{\mu_0}[X]) d(\nu_0 - \mu_0)(x)\\
&+ \int \psi^{(\cdot)}(x - \mathbb{E}_{\mu_1}[X]) d(\nu_1 - \mu_1)(x)\\
&- \int \nabla\varphi^{(\cdot)}(x)^\intercal(\mathbb{E}_{\nu_0}[X] - \mathbb{E}_{\mu_0}[X]) d\bar{\mu}_0(x).
\end{aligned}$$

Proof. For $\boldsymbol{A} \in B_F(M)$ and $t \in [0,1]$, let $(\varphi_t^{\boldsymbol{A}}, \psi_t^{\boldsymbol{A}})$ denote the uniformly bounded extended potentials for $\mathsf{OT}_{\boldsymbol{A}}(\bar{\mu}_{0,t}, \tilde{\mu}_{1,t})$ afforded by Proposition 3. Observe that

$$\begin{aligned}
\mathsf{OT}_{\boldsymbol{A}}(\bar{\mu}_{0,t}, \tilde{\mu}_{1,t}) - \mathsf{OT}_{\boldsymbol{A}}(\bar{\mu}_0, \bar{\mu}_1) &\leq \int \varphi_t^{\boldsymbol{A}} d\bar{\mu}_{0,t} + \int \psi_t^{\boldsymbol{A}} d\tilde{\mu}_{1,t} - \int \varphi_t^{\boldsymbol{A}} d\bar{\mu}_0 - \int \psi_t^{\boldsymbol{A}} d\bar{\mu}_1\\
&= \int \varphi_t^{\boldsymbol{A}} d(\bar{\mu}_{0,t} - \bar{\mu}_0) + \int \psi_t^{\boldsymbol{A}} d(\tilde{\mu}_{1,t} - \bar{\mu}_1),
\end{aligned}$$

and

$$\mathsf{OT}_{\boldsymbol{A}}(\bar{\mu}_{0,t},\tilde{\mu}_{1,t})-\mathsf{OT}_{\boldsymbol{A}}(\bar{\mu}_0,\bar{\mu}_1)\geq\int\varphi_0^{\boldsymbol{A}}d\bar{\mu}_{0,t}+\int\psi_0^{\boldsymbol{A}}d\tilde{\mu}_{1,t}-\int\varphi_0^{\boldsymbol{A}}d\bar{\mu}_0-\int\psi_0^{\boldsymbol{A}}d\bar{\mu}_1$$
$$=\int\varphi_0^{\boldsymbol{A}}d(\bar{\mu}_{0,t}-\bar{\mu}_0)+\int\psi_0^{\boldsymbol{A}}d(\tilde{\mu}_{1,t}-\bar{\mu}_1).$$

Letting $\tau_t=\mathrm{Id}-\mathbb{E}_{\mu_0}[X]-t\left(\mathbb{E}_{\nu_0}[X]-\mathbb{E}_{\mu_0}[X]\right)$, we have by definition that

$$\int\psi_t^{\boldsymbol{A}}d(\tilde{\mu}_{1,t}-\bar{\mu}_1)=t\int\psi_t^{\boldsymbol{A}}(\cdot-\mathbb{E}_{\mu_1}[X])d(\nu_1-\mu_1)$$
$$\int\varphi_t^{\boldsymbol{A}}d(\bar{\mu}_{0,t}-\bar{\mu}_0)=\int\varphi_t^{\boldsymbol{A}}\circ\tau_t-\varphi_t^{\boldsymbol{A}}(\cdot-\mathbb{E}_{\mu_0}[X])d\mu_0+t\int\varphi_t^{\boldsymbol{A}}\circ\tau_t d(\nu_0-\mu_0)$$
$$=r_{\boldsymbol{A},t}(t)-t\int\nabla\varphi_t^{\boldsymbol{A}}(\cdot)^{\intercal}(\mathbb{E}_{\nu_0}[X]-\mathbb{E}_{\mu_0}[X])d\bar{\mu}_0+t\int\varphi_t^{\boldsymbol{A}}\circ\tau_t d(\nu_0-\mu_0),$$

where the final equality follows from Lemma 9. As $\varphi_0^{\boldsymbol{A}}=\varphi^{\boldsymbol{A}}$ and $\psi_0^{\boldsymbol{A}}=\psi^{\boldsymbol{A}}$,

$$\begin{aligned}
&\sup_{\boldsymbol{A}\in B_F(M)}\left|\frac{\mathsf{OT}_{\boldsymbol{A}}(\mu_{0,t},\mu_{1,t})-\mathsf{OT}_{\boldsymbol{A}}(\mu_0,\mu_1)}{t}-\int\varphi^{\boldsymbol{A}}(\cdot-\mathbb{E}_{\mu_0}[X])d(\nu_0-\mu_0)\right.\\
&\qquad\left.-\int\psi^{\boldsymbol{A}}(\cdot-\mathbb{E}_{\mu_1}[X])d(\nu_1-\mu_1)+\int\nabla\varphi^{\boldsymbol{A}}(\cdot)^{\intercal}(\mathbb{E}_{\nu_0}[X]-\mathbb{E}_{\mu_0}[X])d\bar{\mu}_0\right|\\
&\leq\sup_{\boldsymbol{A}\in B_F(M)}\left|\int\varphi_t^{\boldsymbol{A}}\circ\tau_t-\varphi^{\boldsymbol{A}}(\cdot-\mathbb{E}_{\mu_0}[X])d(\nu_0-\mu_0)\right| \qquad (16)\\
&+\sup_{\boldsymbol{A}\in B_F(M)}\left|\int\psi_t^{\boldsymbol{A}}(\cdot-\mathbb{E}_{\mu_1}[X])-\psi^{\boldsymbol{A}}(\cdot-\mathbb{E}_{\mu_1}[X])d(\nu_1-\mu_1)\right|\\
&+\sup_{\boldsymbol{A}\in B_F(M)}\left|\int\left(\nabla\varphi_t^{\boldsymbol{A}}(\cdot)-\nabla\varphi^{\boldsymbol{A}}(\cdot)\right)^{\intercal}(\mathbb{E}_{\nu_0}[X]-\mathbb{E}_{\mu_0}[X])d\bar{\mu}_0\right|+o(1),
\end{aligned}$$

recalling that $\sup_{t\in[0,1]}\sup_{\boldsymbol{A}\in B_F(M)}|r_{\boldsymbol{A},t}(s)|=O(s^2)$ as $s\to 0$ by Lemma 9. We will show that each of the remaining terms on the right hand side converge to 0 as $t\downarrow 0$, starting with the first.

Fix $\epsilon>0$. For any sequence $t_n\downarrow 0$ with $t_n\leq 1$, let $\boldsymbol{A}_n\in B_F(M)$ be such that

$$\sup_{\boldsymbol{A}\in B_F(M)}\left|\int\varphi_{t_n}^{\boldsymbol{A}}\circ\tau_{t_n}-\varphi^{\boldsymbol{A}}(\cdot-\mathbb{E}_{\mu_0}[X])d(\nu_0-\mu_0)\right|$$
$$\leq\left|\int\varphi_{t_n}^{\boldsymbol{A}_n}\circ\tau_{t_n}-\varphi^{\boldsymbol{A}_n}(\cdot-\mathbb{E}_{\mu_0}[X])d(\nu_0-\mu_0)\right|+\epsilon.$$

Fix an arbitrary subsequence $t_{n'}$. Then, by the theorems of Bolzano-Weierstrass and Arzelà-Ascoli, there exists a further subsequence n'' along which $\boldsymbol{A}_{n''}\to\boldsymbol{A}\in B_F(M)$ and $(\varphi_{t_{n''}}^{\boldsymbol{A}_{n''}},\psi_{t_{n''}}^{\boldsymbol{A}_{n''}})\to(\varphi,\psi)$ uniformly on $\overline{\mathcal{X}^\circ}\times\overline{\mathcal{Y}^\circ}$ for some pair of continuous functions (up to extending the Lipschitz potentials to the closures of the respective sets). We now show that (φ,ψ) is a pair of optimal potentials for $\mathsf{OT}_{\boldsymbol{A}}(\bar{\mu}_0,\bar{\mu}_1)$. As $\varphi_{t_{n''}}^{\boldsymbol{A}_{n''}}(x)+\psi_{t_{n''}}^{\boldsymbol{A}_{n''}}(y)\leq c_{\boldsymbol{A}_{n''}}(x,y)$, $\varphi(x)+\psi(y)\leq c_{\boldsymbol{A}}(x,y)$. As the potentials are uniformly bounded,

$$\int\varphi_{t_{n''}}^{\boldsymbol{A}_{n''}}d\bar{\mu}_{0,t}+\int\psi_{t_{n''}}^{\boldsymbol{A}_{n''}}d\bar{\mu}_{1,t}\to\int\varphi d\bar{\mu}_0+\int\psi d\bar{\mu}_1,$$

and, by Theorem 5.20 in [53], $\int\varphi d\bar{\mu}_0+\int\psi d\bar{\mu}_1=\int c_{\boldsymbol{A}}(x,y)d\pi_{\boldsymbol{A}}(x,y)$ such that $\varphi(x)+\psi(y)=c_{\boldsymbol{A}}(x,y)$ $\pi_{\boldsymbol{A}}$-a.e. In particular, for $\bar{\mu}_0$-a.e. $x\in\mathcal{X}^\circ$, $\varphi(x)=\inf_{1\leq i\leq N}\left(c_{\boldsymbol{A}}(x,y^{(i)})-\psi(y^{(i)})\right)$ such that $(\psi(y^{(1)}),\ldots,\psi(y^{(N)}))$ is an optimal dual vector for $\mathsf{OT}_{\boldsymbol{A}}(\bar{\mu}_0,\bar{\mu}_1)$. By Proposition 3, $\varphi=\varphi^{\boldsymbol{A}}+a$ on $\mathrm{int}(\mathrm{spt}(\bar{\mu}_0))$ and $\psi=\psi^{\boldsymbol{A}}-a$ on $\mathcal{Y}-\mathbb{E}_{\mu_1}[X]$ for some $a\in\mathbb{R}$. Thus, $\int\varphi_{t_{n''}}^{\boldsymbol{A}_{n''}}\circ\tau_{t_{n''}}d(\nu_0-\mu_0)\to\int\varphi^{\boldsymbol{A}}(\cdot-\mathbb{E}_{\mu_0}[X])d(\nu_0-\mu_0)$ (recalling that $\mathrm{spt}(\nu_0)\subset\mathrm{spt}(\mu_0)$ by definition of $\mathfrak{P}$). A similar argument shows that $\int\varphi^{\boldsymbol{A}_{n'''}}(\cdot-\mathbb{E}_{\mu_0}[X])d(\nu_0-\mu_0)\to\int\varphi^{\boldsymbol{A}}(\cdot-\mathbb{E}_{\mu_0}[X])d(\nu_0-\mu_0)$ along a further subsequence n''' such that $\int\varphi_{t_{n'''}}^{\boldsymbol{A}_{n'''}}\circ\tau_{t_{n''}}d(\nu_0-\mu_0)-\int\varphi^{\boldsymbol{A}_{n'''}}d(\nu_0-\mu_0)\to 0$. As

the limit thus obtained is independent of the choice of subsequence, convergence holds along the entire subsequence, so

$$\limsup_{n\to\infty}\left\|\int \varphi_{t_n}^{(\cdot)}\circ\tau_{t_n}-\varphi^{(\cdot)}(x-\mathbb{E}_{\mu_0}[X])d(\nu_0-\mu_0)(x)\right\|_{\infty,B_F(M)}\leq\epsilon,$$

and as ϵ is arbitrary, $\lim_{t\downarrow 0}\left\|\int \varphi_{t_n}^{(\cdot)}\circ\tau_{t_n}-\varphi^{(\cdot)}(x-\mathbb{E}_{\mu_0}[X])d(\nu_0-\mu_0)(x)\right\|_{\infty,B_F(M)}=0$. Similarly, $\lim_{t\downarrow 0}\left\|\int \psi_t^{(\cdot)}(y-\mathbb{E}_{\mu_1}[X])-\psi^{(\cdot)}(y-\mathbb{E}_{\mu_1}[X])d(\nu_1-\mu_1(y)\right\|_{\infty,B_F(M)}=0$.

As for the penultimate term, $\sup_{\boldsymbol{A}\in B_F(M)}\left|\int\left(\nabla\varphi_t^{\boldsymbol{A}}(\cdot)-\nabla\varphi^{\boldsymbol{A}}(\cdot)\right)^{\intercal}(\mathbb{E}_{\nu_0}[X]-\mathbb{E}_{\mu_0}[X])d\bar{\mu}_0\right|$, fix a sequence $t_n\downarrow 0$ with $t_n\leq 1$, $\epsilon>0$, and let $\boldsymbol{A}_n$ be such that

$$\begin{aligned}&\sup_{\boldsymbol{A}\in B_F(M)}\left|\int\left(\nabla\varphi_t^{\boldsymbol{A}}(\cdot)-\nabla\varphi^{\boldsymbol{A}}(\cdot)\right)^{\intercal}(\mathbb{E}_{\nu_0}[X]-\mathbb{E}_{\mu_0}[X])d\bar{\mu}_0\right|\\&\qquad\leq\left|\int\left(\nabla\varphi_{t_n}^{\boldsymbol{A}_n}(\cdot)-\nabla\varphi^{\boldsymbol{A}_n}(\cdot)\right)^{\intercal}(\mathbb{E}_{\nu_0}[X]-\mathbb{E}_{\mu_0}[X])d\bar{\mu}_0\right|+\epsilon.\end{aligned}$$

From the previous part of the proof, for any subsequence n' there is a further subsequence n'' along which $\boldsymbol{A}_{n''}\to\boldsymbol{A}\in B_F(M)$ and $\varphi_{t_{n''}}^{\boldsymbol{A}_{n''}}\to\varphi$ uniformly on $\overline{\mathcal{X}^\circ}$, where φ is continuous and coincides with $\varphi^{\boldsymbol{A}}+a$ on $\operatorname{int}(\operatorname{spt}(\bar{\mu}_0))$ for some $a\in\mathbb{R}$. As $\varphi_{t_{n''}}^{\boldsymbol{A}_{n''}}$ is a collection of concave functions on $\mathcal{X}^\circ$ (see Proposition 3), φ is concave on $\mathcal{X}^\circ$ by Theorem 10.8 in [39]. In particular, φ is differentiable a.e. on $\mathcal{X}^\circ$ by Rademacher's theorem, so the functions

$$\begin{aligned}h_{n''}:t\in(-1,1)&\mapsto\int\varphi_{t_{n''}}^{\boldsymbol{A}_{n''}}(\cdot-t(\mathbb{E}_{\nu_0}[X]-\mathbb{E}_{\mu_0}[X]))d\bar{\mu}_0,\\h:t\in(-1,1)&\mapsto\int\varphi(\cdot-t(\mathbb{E}_{\nu_0}[X]-\mathbb{E}_{\mu_0}[X]))d\bar{\mu}_0,\end{aligned}$$

are concave, differentiable, and $h_{n''}\to h$ pointwise. It follows from Theorem 25.7 in [39] that $h'_{n''}(0)\to h'(0)$, that is,

$$\int\nabla\varphi_{t_{n''}}^{\boldsymbol{A}_{n''}}(\cdot)^{\intercal}(\mathbb{E}_{\nu_0}[X]-\mathbb{E}_{\mu_0}[X])d\bar{\mu}_0\to\int\nabla\varphi(\cdot)^{\intercal}(\mathbb{E}_{\nu_0}[X]-\mathbb{E}_{\mu_0}[X])d\bar{\mu}_0.$$

As $\varphi=\varphi^{\boldsymbol{A}}+a$ on $\operatorname{int}(\operatorname{spt}(\bar{\mu}_0))$, $\nabla\varphi=\nabla\varphi^{\boldsymbol{A}}$ on $\operatorname{int}(\operatorname{spt}(\bar{\mu}_0))$ as well. A similar argument shows that up to extraction of a further subsequence n''', $\int\nabla\varphi^{\boldsymbol{A}_{n'''}}(\cdot)^{\intercal}(\mathbb{E}_{\nu_0}[X]-\mathbb{E}_{\mu_0}[X])d\bar{\mu}_0\to\int\nabla\varphi^{\boldsymbol{A}}(\cdot)^{\intercal}(\mathbb{E}_{\nu_0}[X]-\mathbb{E}_{\mu_0}[X])d\bar{\mu}_0$. Conclude that

$$\int\nabla\varphi_{t_{n'''}}^{\boldsymbol{A}_{n'''}}(\cdot)^{\intercal}(\mathbb{E}_{\nu_0}[X]-\mathbb{E}_{\mu_0}[X])d\bar{\mu}_0-\int\nabla\varphi^{\boldsymbol{A}_{n'''}}(\cdot)^{\intercal}(\mathbb{E}_{\nu_0}[X]-\mathbb{E}_{\mu_0}[X])d\bar{\mu}_0\to 0,$$

which is independent of the choice of original subsequence, so

$$\limsup_{t\downarrow 0}\sup_{\boldsymbol{A}\in B_F(M)}\left|\int\left(\nabla\varphi_t^{\boldsymbol{A}}(\cdot)-\nabla\varphi^{\boldsymbol{A}}(\cdot)\right)^{\intercal}(\mathbb{E}_{\nu_0}[X]-\mathbb{E}_{\mu_0}[X])d\bar{\mu}_0\right|\leq\epsilon,$$

which shows that $\sup_{\boldsymbol{A}\in B_F(M)}\left|\int\left(\nabla\varphi_t^{\boldsymbol{A}}(\cdot)-\nabla\varphi^{\boldsymbol{A}}(\cdot)\right)^{\intercal}(\mathbb{E}_{\nu_0}[X]-\mathbb{E}_{\mu_0}[X])d\bar{\mu}_0\right|\to 0$ as $t\downarrow 0$. Conclude that the right hand side of (16) converges to 0, proving the claim. □

Lemma 11. *For any $\nu_0\otimes\nu_1,\rho_0\otimes\rho_1\in\mathfrak{P}$, there exists a finite universal constant C for which $\|\mathsf{OT}_{(\cdot)}(\bar{\nu}_0,\bar{\nu}_1)-\mathsf{OT}_{(\cdot)}(\bar{\rho}_0,\bar{\rho}_1)\|_{\infty,B_F(M)}\leq C\|\nu_0\otimes\nu_1-\rho_0\otimes\rho_1\|_{\infty,\mathcal{F}_K^\oplus}$.*

Proof. Fix $\boldsymbol{A}\in B_F(M)$ and let $(\varphi_\nu^{\boldsymbol{A}},\psi_\nu^{\boldsymbol{A}})$ and $(\varphi_\rho^{\boldsymbol{A}},\psi_\rho^{\boldsymbol{A}})$ be extended potentials for $\mathsf{OT}_{\boldsymbol{A}}(\bar{\nu}_0,\bar{\nu}_1)$ and $\mathsf{OT}_{\boldsymbol{A}}(\bar{\rho}_0,\bar{\rho}_1)$ respectively satisfying the bounds from Proposition 3. Then,

$$\begin{aligned}\mathsf{OT}_{\boldsymbol{A}}(\bar{\nu}_0,\bar{\nu}_1)-\mathsf{OT}_{\boldsymbol{A}}(\bar{\rho}_0,\bar{\rho}_1)&\leq\int\varphi_\nu^{\boldsymbol{A}}d\bar{\nu}_0+\int\psi_\nu^{\boldsymbol{A}}d\bar{\nu}_1-\int\varphi_\nu^{\boldsymbol{A}}d\bar{\rho}_0-\int\psi_\nu^{\boldsymbol{A}}d\bar{\rho}_1,\\\mathsf{OT}_{\boldsymbol{A}}(\bar{\nu}_0,\bar{\nu}_1)-\mathsf{OT}_{\boldsymbol{A}}(\bar{\rho}_0,\bar{\rho}_1)&\geq\int\varphi_\rho^{\boldsymbol{A}}d\bar{\nu}_0+\int\psi_\rho^{\boldsymbol{A}}d\bar{\nu}_1-\int\varphi_\rho^{\boldsymbol{A}}d\bar{\rho}_0-\int\psi_\rho^{\boldsymbol{A}}d\bar{\rho}_1.\end{aligned}$$

Let L denote the shared Lipschitz constant of the extended potentials from Proposition 3, then

$$\begin{aligned}\int \varphi_\nu^{\boldsymbol{A}} d\bar{\nu}_0 - \int \varphi_\nu^{\boldsymbol{A}} d\bar{\rho}_0 &= \int \varphi_\nu^{\boldsymbol{A}}(\cdot - \mathbb{E}_{\nu_0}[X]) d\nu_0 - \int \varphi_\nu^{\boldsymbol{A}}(\cdot - \mathbb{E}_{\rho_0}[X]) d\rho_0 \\ &\leq \int \varphi_\nu^{\boldsymbol{A}}(\cdot - \mathbb{E}_{\rho_0}[X]) d(\nu_0 - \rho_0) + L\|\mathbb{E}_{\rho_0}[X] - \mathbb{E}_{\nu_0}[X]\|.\end{aligned}$$

The same argument can be used to bound $\int \psi_\nu^{\boldsymbol{A}} d\bar{\nu}_1 - \int \psi_\nu^{\boldsymbol{A}} d\bar{\rho}_1$. Whence,

$$\begin{aligned}\mathsf{OT}_{\boldsymbol{A}}(\bar{\nu}_0, \bar{\nu}_1) - \mathsf{OT}_{\boldsymbol{A}}(\bar{\rho}_0, \bar{\rho}_1) \leq &\int \varphi_\nu^{\boldsymbol{A}}(\cdot - \mathbb{E}_{\rho_0}[X]) d(\nu_0 - \rho_0) + L\|\mathbb{E}_{\rho_0}[X] - \mathbb{E}_{\nu_0}[X]\| \\ &+ \int \psi_\nu^{\boldsymbol{A}}(\cdot - \mathbb{E}_{\rho_1}[X]) d(\nu_1 - \rho_1) + L\|\mathbb{E}_{\rho_1}[X] - \mathbb{E}_{\nu_1}[X]\|.\end{aligned}$$

Following similar steps for the lower bound yields

$$\begin{aligned}\mathsf{OT}_{\boldsymbol{A}}(\bar{\nu}_0, \bar{\nu}_1) - \mathsf{OT}_{\boldsymbol{A}}(\bar{\rho}_0, \bar{\rho}_1) \geq &\int \varphi_\rho^{\boldsymbol{A}}(\cdot - \mathbb{E}_{\rho_0}[X]) d(\nu_0 - \rho_0) - L\|\mathbb{E}_{\rho_0}[X] - \mathbb{E}_{\nu_0}[X]\| \\ &+ \int \psi_\rho^{\boldsymbol{A}}(\cdot - \mathbb{E}_{\rho_1}[X]) d(\nu_1 - \rho_1) - L\|\mathbb{E}_{\rho_1}[X] - \mathbb{E}_{\nu_1}[X]\|.\end{aligned}$$

As in the proof of Lemma 6, $\|\mathbb{E}_{\rho_0}[X] - \mathbb{E}_{\nu_0}[X]\| + \|\mathbb{E}_{\rho_1}[X] - \mathbb{E}_{\nu_1}[X]\| \lesssim_{\mathcal{X},\mathcal{Y}} \|\nu_0 \otimes \nu_1 - \rho_0 \otimes \rho_1\|_{\infty, \mathcal{F}_K^\oplus}$. Observe that, by choosing K sufficiently large, $\varphi_\rho^{\boldsymbol{A}}(\cdot - \mathbb{E}_{\rho_0}[X])|_{\mathcal{X}} \in \mathcal{H}_{0,K}$ and $\psi_\rho^{\boldsymbol{A}}(\cdot - \mathbb{E}_{\rho_1}[X])|_{\mathcal{Y}} \in \mathcal{G}_{1,K}$ uniformly in the choice of marginals and $\boldsymbol{A} \in B_F(M)$. Hence

$$\left| \int \varphi_\rho^{\boldsymbol{A}}(\cdot - \mathbb{E}_{\rho_0}[X]) d(\nu_0 - \rho_0) + \int \psi_\rho^{\boldsymbol{A}}(\cdot - \mathbb{E}_{\rho_1}[X]) d(\nu_1 - \rho_1) \right| \leq \|\nu_0 \otimes \nu_1 - \rho_0 \otimes \rho_1\|_{\infty, \mathcal{F}_K^\oplus},$$

and the same bound holds for $\left| \int \varphi_\nu^{\boldsymbol{A}}(\cdot - \mathbb{E}_{\rho_0}[X]) d(\nu_0 - \rho_0) + \int \psi_\nu^{\boldsymbol{A}}(\cdot - \mathbb{E}_{\rho_1}[X]) d(\nu_1 - \rho_1) \right|$. Conclude that

$$\|\mathsf{OT}_{(\cdot)}(\bar{\nu}_0, \bar{\nu}_1) - \mathsf{OT}_{(\cdot)}(\bar{\rho}_0, \bar{\rho}_1)\|_{\infty, B_F(M)} \lesssim_{\mathcal{X},\mathcal{Y}} \|\nu_0 \otimes \nu_1 - \rho_0 \otimes \rho_1\|_{\infty, \mathcal{F}_K^\oplus}.$$

□

With these preparations, we now prove Proposition 5.

Proof of Proposition 5. Together, Lemmas 7 and 11 along with Proposition 7 imply Hadamard directional differentiability of the map $\nu_0 \otimes \nu_1 \in \mathfrak{P} \mapsto 32\|\cdot\|_F^2 + \mathsf{OT}_{(\cdot)}(\bar{\nu}_0, \bar{\nu}_1) \in \ell^\infty(B_F(M))$. As the Gâteaux directional derivative of this map at $\mu_0 \otimes \mu_1$ is of the form $(\nu_0 \otimes \nu_1 - \mu_0 \otimes \mu_1)(f_0^{(\cdot)} \oplus f_1^{(\cdot)})$ where $f_0^{\boldsymbol{A}} \oplus f_1^{\boldsymbol{A}} \in \mathcal{F}_K^\oplus$ for $\boldsymbol{A} \in B_F(M)$ and a (uniform) choice of K sufficiently large. By the same steps as the proof of Proposition 7, the Hadamard directional derivative is given by $\eta \in \mathcal{T}_{\mathfrak{P}}(\mu_0 \otimes \mu_1) \mapsto \eta(f_0^{(\cdot)} \oplus f_1^{(\cdot)})$.

Proposition 5 then follows by applying Corollary 2.3 in [6] along with the chain rule for Hadamard directionally differentiable maps pending the condition that $\boldsymbol{A} \in B_F(M) \mapsto 32\|\boldsymbol{A}\|_F^2 + \mathsf{OT}_{\boldsymbol{A}}(\bar{\mu}_0, \bar{\mu}_1)$ is continuous (although the cited result is stated under the assumption that the map is not identically zero, this condition is not necessary) and, for any $\eta \in \mathcal{T}_{\mathfrak{P}}(\mu_0 \otimes \mu_1)$,

$$\begin{aligned}\Upsilon_\eta : \boldsymbol{A} \in B_F(M) \mapsto \eta\Big(&\varphi^{\boldsymbol{A}}(\cdot - \mathbb{E}_{\mu_0}[X]) - \mathbb{E}_{\bar{\mu}_0}[\nabla \varphi^{\boldsymbol{A}}(X)]^\intercal(\cdot) \\ &\oplus \psi^{\boldsymbol{A}}(\cdot - \mathbb{E}_{\mu_1}[X]) - 8\mathbb{E}_{(X,Y)\sim\pi_{\boldsymbol{A}}}\left[\|X\|^2 Y\right]^\intercal(\cdot)\Big),\end{aligned}$$

is continuous.

The former condition was verified in Lemma 3. As for the latter condition, fix $\eta \in \mathcal{T}_{\mathfrak{P}}(\mu_0 \otimes \mu_1)$, $\epsilon > 0$ and let $t > 0$, $\nu_0 \otimes \nu_1 \in \mathfrak{P}$ be such that $\|t(\nu_0 \otimes \nu_1 - \mu_0 \otimes \mu_1) - \eta\|_{\infty, \mathcal{F}_K^\oplus} < \frac{\epsilon}{2}$. Let $(\boldsymbol{A}_n)_{n\in\mathbb{N}} \subset B_F(M)$ converge to $\boldsymbol{A} \in B_F(M)$. As in the proof of Lemma 10 we have that $\varphi^{\boldsymbol{A}_n}(\cdot - \mathbb{E}_{\mu_0}[X]) \to \varphi^{\boldsymbol{A}}(\cdot - \mathbb{E}_{\mu_0}[X])$ on $\mathrm{int}(\mathrm{spt}(\mu_0))$, $\int \nabla \varphi^{\boldsymbol{A}_n}(x) d\bar{\mu}_0(x) \to$

$\int \nabla\varphi^{\boldsymbol{A}}(x)d\bar{\mu}_0(x)$, and $\psi^{\boldsymbol{A}_n}(\cdot - \mathbb{E}_{\mu_1}[X]) \to \psi^{\boldsymbol{A}}(\cdot - \mathbb{E}_{\mu_1}[X])$ on $\mathrm{spt}(\mu_1)$. Further, $\pi_{\boldsymbol{A}_n} \xrightarrow{w} \pi_{\boldsymbol{A}}$ hence $-8\int \|x\|^2 y d\pi_{\boldsymbol{A}_n}(x,y) \to -8\int \|x\|^2 y d\pi_{\boldsymbol{A}}(x,y)$. Hence

$$
\begin{aligned}
|\Upsilon_\eta(\boldsymbol{A}_n) - \Upsilon_\eta(\boldsymbol{A})| \leq \epsilon + t \Bigg| & \int \varphi^{\boldsymbol{A}_n} - \varphi^{\boldsymbol{A}} - \mathbb{E}_{\bar{\mu}_0}[\nabla\varphi^{\boldsymbol{A}_n}(X)]^\intercal(\cdot) + \mathbb{E}_{\bar{\mu}_0}[\nabla\varphi^{\boldsymbol{A}}(X)]^\intercal(\cdot) d(\nu_0 - \mu_0) \\
& + \int \psi^{\boldsymbol{A}_n}(\cdot - \mathbb{E}_{\mu_1}[X]) - \psi^{\boldsymbol{A}}(\cdot - \mathbb{E}_{\mu_1}[X]) d(\nu_1 - \mu_1) \\
& - \int 8\mathbb{E}_{(X,Y)\sim\pi_{\boldsymbol{A}_n}}\left[\|X\|^2 Y\right]^\intercal(\cdot) + 8\mathbb{E}_{(X,Y)\sim\pi_{\boldsymbol{A}}}\left[\|X\|^2 Y\right]^\intercal(\cdot) d(\nu_1 - \mu_1) \Bigg|.
\end{aligned}
$$

By the previous deliberations, the right hand side converges to ϵ as $n \to \infty$ and, as ϵ is arbitrary, Υ_η is continuous, concluding the proof. □

A.9 Proof of Theorem 5

Lemma 12. *The class $\mathcal{F}_{0,K}$ is μ_0-Donsker.*

Proof. For any $x, \xi, \xi' \in \mathcal{X}$, $\boldsymbol{A}, \boldsymbol{A}' \in B_F(M)$, and $z, z' \in \mathbb{R}^N$ with $\|z\|_\infty, \|z'\|_\infty \leq K$,

$$
\begin{aligned}
&\left| \min_{1\leq i\leq N} \left\{ c_{\boldsymbol{A}}(x - \xi, y^{(i)}) - z_i \right\} - \min_{1\leq i\leq N} \left\{ c_{\boldsymbol{A}'}(x - \xi', y^{(i)}) - z'_i \right\} \right| \\
&\leq \max_{1\leq i\leq N} \left| -4(\|x-\xi\|^2 - \|x-\xi'\|^2)\|y^{(i)}\|^2 - 32\left((x-\xi)^\intercal \boldsymbol{A} - (x-\xi')^\intercal \boldsymbol{A}'\right) y^{(i)} - (z_i - z'_i) \right|.
\end{aligned}
$$

Observe that

$$
\left| \|x-\xi\|^2 - \|x-\xi'\|^2 \right| = \left| \|x-\xi\| - \|x-\xi'\| \right| (\|x-\xi\| + \|x-\xi'\|) \leq 2\,\mathrm{diam}(\mathcal{X})\|\xi - \xi'\|,
$$

and $\left|(x-\xi')^\intercal(\boldsymbol{A} - \boldsymbol{A}') - (\xi - \xi')^\intercal \boldsymbol{A}\right| \leq \mathrm{diam}(\mathcal{X})\|\boldsymbol{A} - \boldsymbol{A}'\|_F + \|\boldsymbol{A}\|_F\|\xi - \xi'\|$. That is,

$$
\begin{aligned}
&\left| \min_{1\leq i\leq N} \left\{ c_{\boldsymbol{A}}(x - \xi, y^{(i)}) - z_i \right\} - \min_{1\leq i\leq N} \left\{ c_{\boldsymbol{A}'}(x - \xi', y^{(i)}) - z'_i \right\} \right| \\
&\qquad \lesssim_{M,d_0,d_1,\mathcal{X},\mathcal{Y}} \|\xi - \xi'\| + \|\boldsymbol{A} - \boldsymbol{A}'\|_F + \|z - z'\|_\infty.
\end{aligned}
$$

At this point, it is easy to adapt the proof of Lemma 1 to prove Donskerness of the class $\mathcal{H}_{0,K}$ (hence also $\mathcal{H}_{0,K} \cup \{0\}$). Further, $\mathcal{C}_K^\infty(\mathcal{X})$ is known to be Donsker (cf. e.g. [30] or [49]). As $\mathcal{F}_{0,K} = \mathcal{H}_{0,K} \cup \{0\} + \mathcal{C}_K^\infty(\mathcal{X})$ and Donskerness is preserved under pairwise sums, conclude that $\mathcal{F}_{0,K}$ is a Donsker class. □

Proof of Theorem 5. As the samples X_i and Y_i are assumed to be independent, it follows from Example 1.5.6 in [50], Lemma 3.6 in [16], and Lemmas 1 and 12 that

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

Remark that the map $(l_0, l_1) \in \ell^\infty(\mathcal{F}_{0,K}) \times \ell^\infty(\mathcal{G}_{1,K}) \mapsto \left(f_0 \oplus f_1 \in \mathcal{F}_K^\oplus \mapsto l_0(f_0) + l_1(f_1)\right) \in \ell^\infty(\mathcal{F}_K^\oplus)$ is continuous such that

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

where $G_{\mu_0\otimes\mu_1}(f_0 \oplus f_1) = G_{\mu_0}(f_0) + G_{\mu_1}(f_1)$ for any $f_0 \oplus f_1 \in \mathcal{F}_K^\oplus$. It follows from the portmanteau theorem that $G_{\mu_0\otimes\mu_1} \in \mathcal{T}_{\mathfrak{P}}(\mu_0 \otimes \mu_1)$ with probability 1. The claimed result then follows from the extended functional delta method (see e.g. Theorem 1 in [41]). □

B Auxiliary results

Proposition 6 (Well-definedness of identification)**.** *$\nu_0 \otimes \nu_1 = \nu'_0 \otimes \nu'_1$ as elements of $\mathfrak{P}$ if and only if $\nu_0 = \nu'_0$ and $\nu_1 = \nu'_1$ as measures on $\mathcal{P}(\mathcal{X})$ and $\mathcal{P}(\mathcal{Y})$ respectively.*

Proof. Suppose that $\nu_0 \otimes \nu_1 = \nu_0' \otimes \nu_1'$ as elements of $\mathfrak{P}$. As $\mathcal{F}_{0,K}$ and $\mathcal{G}_{1,K}$ both contain 0, we must have that $\int f_0 d\nu_0 = \int f_0 d\nu_0'$ and $\int f_1 d\nu_1 = \int f_1 d\nu_1'$ for every $f_0 \in \mathcal{F}_{0,K}$ and $f_1 \in \mathcal{G}_{1,K}$. By linearity of integration, these equalities extend, respectively, to $f_0 \in \mathrm{lin}(\mathcal{F}_{0,K})$, $f_1 \in \mathrm{lin}(\mathcal{G}_{1,K})$, where $\mathrm{lin}(\mathcal{G}_{1,K}) = \{f : \mathcal{Y} \to \mathbb{R}\}$, so $\nu_1 = \nu_1'$ as measures. Furthermore, $\mathcal{C}^\infty(\mathcal{X}) \subset \mathrm{lin}(\mathcal{F}_{0,K})$, and this former set is dense in $\mathcal{C}(\mathcal{X})|_{\mathrm{spt}(\mu_0)}$; the set of all restrictions of continuous functions on $\mathcal{X}$ to $\mathrm{spt}(\mu_0)$ (cf. e.g. Lemma 7.1 in [24]). Consequently, $\nu_0 = \nu_0'$ as measures as well. □

Proposition 7 (Hadamard and Gâteaux derivatives)**.** *Let $\mathfrak{E}$ be a Banach space. If $\zeta : \mathfrak{P} \subset \ell^\infty(\mathcal{F}_K^\oplus) \to \mathfrak{E}$ is such that*

$$\|\zeta(\nu_0 \otimes \nu_1) - \zeta(\rho_0 \otimes \rho_1)\|_{\mathfrak{E}} \leq C\|\nu_0 \otimes \nu_1 - \rho_0 \otimes \rho_1\|_{\infty, \mathcal{F}_K^\oplus},$$

for every $\nu_0 \otimes \nu_1, \rho_0 \otimes \rho_1 \in \mathfrak{P}$ and, for any $\nu_0 \otimes \nu_1 \in \mathfrak{P}$, the limit

$$\lim_{t \downarrow 0} \frac{\zeta\left(\mu_0 \otimes \mu_1 + t(\nu_0 \otimes \nu_1 - \mu_0 \otimes \mu_1)\right) - \zeta(\mu_0 \otimes \mu_1)}{t} = \zeta'_{\mu_0 \otimes \mu_1}(\nu_0 \otimes \nu_1 - \mu_0 \otimes \mu_1) \tag{17}$$

exists. Then ζ is Hadamard directionally differentiable at $\mu_0 \otimes \mu_1$ with derivative coinciding with (17) *on $\mathfrak{P} - \mu_0 \otimes \mu_1$. If $\mathfrak{E} = \mathbb{R}$ and $\zeta'_{\mu_0 \otimes \mu_1}(\nu_0 \otimes \nu_1 - \mu_0 \otimes \mu_1) = (\nu_0 \otimes \nu_1 - \mu_0 \otimes \mu_1)(f_0 \oplus f_1)$ for some $f_0 \oplus f_1 \in \mathcal{F}_K^\oplus$, the Hadamard directional derivative is given by $\eta \in \mathcal{T}_{\mathfrak{P}}(\mu_0 \otimes \mu_1) \mapsto \eta\left(f_0 \oplus f_1\right)$.*

Proof. The proof of the first part follows from that of Proposition 1 in [25] or Lemma 8 in [38] and is hence omitted. The expression for the Hadamard directional derivative in the second part follows from the fact that the derivative is positively homogeneous and continuous [41] and that $\mathcal{T}_{\mathfrak{P}}(\mu_0 \otimes \mu_1) = \overline{\{t^{-1}(\nu_0 \otimes \nu_1 - \mu_0 \otimes \mu_1) : \nu_0 \otimes \nu_1 \in \mathfrak{P}, t > 0\}}^{\ell^\infty(\mathcal{F}_K^\oplus)}$. □