Entropic Gromov-Wasserstein Distances: Stability and Algorithms

Gabriel Rioux GER84@CORNELL.EDU
Center for Applied Mathematics
Cornell University
Ithaca, NY 14853, USA

Ziv Goldfeld GOLDFELD@CORNELL.EDU
School of Electrical and Computer Engineering
Cornell University
Ithaca, NY 14853, USA

Kengo Kato KK976@CORNELL.EDU
Department of Statistics and Data Science
Cornell University
Ithaca, NY 14853, USA

Abstract

The Gromov-Wasserstein (GW) distance quantifies discrepancy between metric measure spaces and provides a natural framework for aligning heterogeneous datasets. Alas, as exact computation of GW alignment is NP hard, entropic regularization provides an avenue towards a computationally tractable proxy. Leveraging a recently derived variational representation for the quadratic entropic GW (EGW) distance, this work derives the first efficient algorithms for solving the EGW problem subject to formal, non-asymptotic convergence guarantees. To that end, we derive smoothness and convexity properties of the objective in this variational problem, which enables its resolution by the accelerated gradient method. Our algorithms employs Sinkhorn's fixed point iterations to compute an approximate gradient, which we model as an inexact oracle. We furnish convergence rates towards local and even global solutions (the latter holds under a precise quantitative condition on the regularization parameter), characterize the effects of gradient inexactness, and prove that stationary points of the EGW problem converge towards a stationary point of the unregularized GW problem, in the limit of vanishing regularization. We provide numerical experiments that validate our theory and empirically demonstrate the state-of-the-art empirical performance of our algorithm.

Keywords: Algorithms, convergence rate, global guarantees, Gromov-Wasserstein distances, entropic regularization, inexact gradient methods.

1 Introduction

The Gromov-Wasserstein (GW) distance compares probability distributions that are supported on possibly distinct metric spaces by aligning them with one another. Given two metric measure (mm) spaces $(\mathcal{X}_0, \mathsf{d}_0, \mu_0)$ and $(\mathcal{X}_1, \mathsf{d}_1, \mu_1)$, the (p, q)-GW distance between them is

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

where $\Pi(\mu_0, \mu_1)$ is the set of couplings between μ_0 and μ_1. This approach, proposed in Mémoli (2011), is an optimal transport (OT) based L^p-relaxation of the classical Gromov-Hausdorff distance between metric spaces. The GW distance defines a metric on the space of all mm spaces modulo measure preserving isometries.[1] From an applied standpoint, a solution to the GW problem between two heterogeneous datasets yields not only a quantification of discrepancy, but also an optimal alignment $\pi^\star$ between them. As such, alignment methods inspired by the GW problem have been proposed for many applications, encompassing single-cell genomics (Blumberg et al., 2020; Demetci et al., 2022), alignment of language models (Alvarez-Melis and Jaakkola, 2018), shape matching (Mémoli, 2009; Koehl et al., 2023), graph matching (Xu et al., 2019b,a), heterogeneous domain adaptation (Yan et al., 2018), and generative modeling (Bunne et al., 2019).

Exact computation of the GW distance is a quadratic assignment problem, which is known to be NP-complete (Commander, 2005). To remedy this, various computationally tractable reformulations of the distance have been proposed. We postpone full discussion of such methods to Section 1.2 and focus here on the entropic GW (EGW) distance (Peyré et al., 2016; Solomon et al., 2016)

$$\mathsf{S}^{\varepsilon}_{p,q}(\mu_0, \mu_1) \coloneqq \inf_{\pi \in \Pi(\mu_0, \mu_1)} \iint \left|\mathsf{d}_0^q(x, x') - \mathsf{d}_1^q(y, y')\right|^p d\pi \otimes \pi(x, y, x', y') + \varepsilon \mathsf{D}_{\mathsf{KL}}(\pi \| \mu_0 \otimes \mu_1),$$

which is at the center of this work. Entropic regularization by means of the Kullback-Leibler (KL) divergence above transforms the linear Kantorovich OT problem (Kantorovich, 1942) into a strictly convex one and enables directly solving it using Sinkhorn's algorithm (Cuturi, 2013). Although the EGW problem is, in general, not convex, Solomon et al. (2016) propose to solve it via an iterative approach with Sinkhorn iterations. This method is known to converge to a stationary point of a tight (albeit non-convex) relaxation of the EGW problem, but this is an asymptotic statement and the overall computational complexity is unknown. Similar limitations apply for the popular mirror descent-based approach from Peyré et al. (2016). To the best of our knowledge, there is currently no known algorithm for computing the EGW distance subject to non-asymptotic convergence rate bounds, let alone global optimality guarantees. Further, these methods do not account for the error incurred by using Sinkhorn iterations, nor do they address the behaviour of the solutions they obtain in the limit of vanishing regularization.

The goal of this work is to close the aforementioned computational gaps, targeting algorithms with non-asymptotic guarantees, accounting for inexactness in Sinkhorn's algorithm, characterizing convexity regimes of the EGW problem, and establishing convergence of stationary points to the EGW problem to stationary points of the GW problem as $\varepsilon \downarrow 0$. All of these will be achieved as a consequence of a new stability analysis of the EGW variational representation from Zhang et al. (2022a).

1. Two mm spaces $(\mathcal{X}_0, \mathsf{d}_0, \mu_0)$ and $(\mathcal{X}_1, \mathsf{d}_1, \mu_1)$ are isomorphic if there exists an isometry $T : \mathcal{X}_0 \to \mathcal{X}_1$ for which $\mu_0 \circ T^{-1} = \mu_1$ as measures. The quotient space is then the one induced by this equivalence relation.

1.1 Contributions

Theorem 1 in Zhang et al. (2022a) shows that the EGW distance with quadratic cost between the Euclidean mm spaces $(\mathbb{R}^{d_0}, \|\cdot\|, \mu_0)$ and $(\mathbb{R}^{d_1}, \|\cdot\|, \mu_1)$ can be written as:

$$\mathsf{S}^{\varepsilon}_{2,2}(\mu_0, \mu_1) = C_{\mu_0,\mu_1} + \inf_{\boldsymbol{A} \in \mathcal{D}_M} \left\{ 32\|\boldsymbol{A}\|_F^2 + \mathsf{OT}_{\boldsymbol{A},\varepsilon}(\mu_0, \mu_1) \right\}, \tag{2}$$

where C_{μ_0,μ_1} is a constant that depends only on the moments of the marginals, $\mathcal{D}_M \subset \mathbb{R}^{d_0 \times d_1}$ is a compact rectangle, and $\mathsf{OT}_{\boldsymbol{A},\varepsilon}(\mu_0, \mu_1)$ is an EOT problem with a particular cost function that depends on the auxiliary variable $\boldsymbol{A}$. This representation connects EGW to the well-understood EOT problem, thereby unlocking powerful tools for analysis. To exploit this connection for new computational advancements, we begin by analyzing the stability of the objective in (2) in $\boldsymbol{A}$ and derive its first and second-order Fréchet derivatives. This, in turn, enables us to derive its convexity and smoothness properties and devise new algorithms for solving the EGW problem, subject to formal convergence guarantees.

The Fréchet derivatives of the objective function from (2) in $\boldsymbol{A}$ reveal that this problem falls under the paradigm of smooth constrained optimization. Indeed, the derivatives imply, respectively, upper and lower bounds on the top and bottom eigenvalues of the Hessian matrix of the objective; L-smoothness then follows from the mean value inequality. By further requiring positive semidefiniteness of the Hessian we obtain a sharp and primitive sufficient condition on ε under which (2) becomes convex. These regularity properties are used to lift the accelerated first-order methods for smooth (non-convex) optimization (Ghadimi and Lan, 2016) and convex programming with an inexact oracle (d'Aspremont, 2008) to the EGW problem. This yields the first algorithms with non-asymptotic convergence guarantees toward global or local EGW solutions, depending on whether the problem is convex or not (e.g., under the aforementioned condition). Our algorithms compute not only the EGW cost, but also provide an approximate optimizer—namely a coupling, which serves as the alignment scheme that achieves the said cost.

Specifically, our method iteratively solves the optimization problem in the space of auxiliary matrices $\boldsymbol{A} \in \mathbb{R}^{d_0 \times d_1}$, with each iterate calling the Sinkhorn algorithm to obtain an approximate solution (viz. the inexact oracle) to the corresponding EOT problem. The time complexity of the Sinkhorn algorithm governs the overall runtime, which is therefore $O(N^2)$, for μ_0 and μ_1 as distributions on N points. This presents a significant speedup to the $O(N^3)$ runtime of popular iterative algorithms from Peyré et al. (2016); Solomon et al. (2016). Under certain low-rank assumptions on the cost matrix, Scetbon et al. (2022) recently showed the mirror descent approach from Peyré et al. (2016) can be sped up to run in $O(N^2)$ time, which is comparable to our method. Nevertheless, our algorithms are coupled with formal convergence guarantees, non-asymptotic error bounds, and global optimality claims under the said convexity condition, while no such assurances are available for other methods.

As the derivative of the objective from (2) depends on the optimal coupling for a particular EOT problem, we also account for the error incurred by solving this problem numerically (e.g. via Sinkhorn iterations). In particular, we characterize how well the output of a standard implementation of Sinkhorn's algorithm approximates the true EOT coupling. This effect was not analyzed before, as existing literature focused on approximating the unregularized OT cost, while treating the KL divergence term as a bias.

The EGW distance serves as a proxy to unregularized GW, which renders the vanishing regularization parameter regime, namely $\varepsilon \downarrow 0$, of central importance. We show that stationary points of the variational EGW problem converge (possibly along a subsequence) to a stationary point of the variational GW problem as $\varepsilon \downarrow 0$. Only convergence to a stationary point can be guaranteed in the limit since the underlying variational problem may fail to be convex when ε is small. Nonetheless, this is the first result that provides stationarity guarantees for limiting solutions, which clarifies how the local solutions obtained using our algorithm approximate local solutions to the unregularized problem.

1.2 Literature review

The computational intractability of the GW problem in (1) has inspired several reformulations that aim to alleviate this issue. The sliced GW distance (Vayer et al., 2020) attempts to reduce the computational burden by considering the average of GW distances between one-dimensional projections of the marginals. However, unlike OT in one dimension, the GW problem does not have a known simple solution even in one dimension (Beinert et al., 2022). Another approach is to relax the strict marginal constraints by optimizing over the weights of one of the marginals as in semi-relaxed GW (Vincent-Cuaz et al., 2022) or by using f-divergence penalties; this leads to the unbalanced GW distance (Séjourné et al., 2021), which lends itself well for convex/conic relaxations. A variant that directly optimizes over bi-directional Monge maps between the mm spaces was considered in Zhang et al. (2022b). The fused GW distance (Vayer et al., 2019) enables comparing both feature and structural properties of structured data. Although most of these relaxations can be shown to converge to the GW distance (in terms of function values) under certain regimes, they involve solving non-convex problems, which limits their utility for numerical resolution of the GW problem. The recent work of Chen et al. (2023) proposes a semidefinite relaxation of the GW problem along with a certificate of optimality which, upon obtaining a solution to the relaxed problem, establishes if it is optimal for the original problem.

While these methods offer certain advantages, it is the approach based on entropic regularization (Peyré et al., 2016; Solomon et al., 2016) that is most frequently used in application. A low-rank variant of the EGW problem was proposed in Scetbon et al. (2022), where the distance distortion cost is only optimized over coupling admitting a certain low-rank structure. They arrive at a linear time algorithm for this problem by adapting the mirror descent method of Peyré et al. (2016). As an intermediate step of their analysis, they show that if μ_0 and μ_1 are supported on N distinct points, then the $O(N^3)$ complexity of mirror descent (see, e.g., Remark 1 in Peyré et al. 2016) can be reduced to $O(N^2)$ by assuming that the matrices of pairwise costs admit a low-rank decomposition (without imposing any structure on the couplings). This decomposition holds, for instance, when the cost is the squared Euclidean distance and the sample size dominates the ambient dimension. Although mirror descent seems to solve the EGW problem quite well in practice, formal guarantees concerning convergence rates or local optimality are lacking.

Other related work explores structural properties of GW distances, on which some of our findings also reflect. The existence of Monge maps for the GW problem was studied in Dumont et al. (2022) and they show that optimal couplings are induced by a bimap (viz. two-way map) under general conditions. Delon et al. (2022) focused on the GW distance

between Gaussian distributions, deriving upper and lower bounds on the optimal cost. A closed-form expression under the Gaussian setting was derived in Le et al. (2022) for the EGW distance with inner product cost.

As we establish stability results for the EGW problem by utilizing its connection to the EOT problem with a parametrized cost, we mention that different notions of stability for EOT have been studied. For instance, Ghosal et al. (2022) concerns stability of the EOT cost for varying marginals, cost function, and regularization parameter. The related works (Carlier and Laborde, 2020; Eckstein and Nutz, 2022; Nutz and Wiesel, 2023) concern stability of the EOT cost and related objects for weakly convergent marginal distributions.

1.3 Organization

This paper is organized as follows. In Section 2 we compile background material on EOT and the EGW problems. In Section 3, we describe the smoothness and convexity of the variational EGW problem. In Section 4, we analyze and test two algorithms for solving the EGW problem. We compile the proofs for Sections 3 and 4 in Section 5. Section 6 contains some concluding remarks.

1.4 Notation

Denote by $\mathcal{P}(\mathbb{R}^d)$ the collection of all Borel probability measures on $\mathbb{R}^d$, and by $\mathcal{P}_p(\mathbb{R}^d)$ the set of all $\mu \in \mathcal{P}(\mathbb{R}^d)$ with finite p-th moment ($p > 0$). The pushforward of $\mu \in \mathcal{P}(\mathbb{R}^{d_0})$ through a measurable map $T : \mathbb{R}^{d_0} \to \mathbb{R}^{d_1}$ is denoted by $T_\sharp \mu := \mu \circ T^{-1}$. The Frobenius inner product on $\mathbb{R}^{d_0 \times d_1}$ is defined by $\langle \boldsymbol{A}, \boldsymbol{B} \rangle_F = \operatorname{tr}(\boldsymbol{A}^\intercal \boldsymbol{B})$; the associated norm is denoted by $\|\cdot\|_F$. For a nonemtpy set $S \subset \mathbb{R}^d$, $\mathcal{C}(S)$ is the set of continuous functions on S. We adopt the shorthands $a \wedge b = \min\{a, b\}$ and $a \vee b = \max\{a, b\}$.

2 Background and Preliminaries

We first establish notation and review standard definitions and results underpinning our analysis of the EGW distance.

2.1 Entropic Optimal Transport

Entropic regularization transforms the linear OT problem into a strictly convex one. Given distributions $\mu_i \in \mathcal{P}(\mathbb{R}^{d_i})$, $i = 0, 1$, and a Borel cost function $c : \mathbb{R}^{d_0} \times \mathbb{R}^{d_1} \to \mathbb{R}$ that is bounded from below on $\operatorname{spt}(\mu_0) \times \operatorname{spt}(\mu_1)$, the primal EOT problem is obtained by regularizing the standard OT problem via the Kullback-Leibler (KL) divergence,

$$\mathsf{OT}_\varepsilon(\mu_0, \mu_1) = \inf_{\pi \in \Pi(\mu_0, \mu_1)} \int c \, d\pi + \varepsilon \mathsf{D}_{\mathsf{KL}}(\pi \| \mu_0 \otimes \mu_1),$$

where $\varepsilon > 0$ is a regularization parameter and

$$\mathsf{D}_{\mathsf{KL}}(\mu_0 \| \mu_1) = \begin{cases} \int \log\left(\frac{d\mu_0}{d\mu_1}\right) d\mu_0, & \text{if } \mu_0 \ll \mu_1, \\ +\infty, & \text{otherwise.} \end{cases}$$

Classical OT is obtained from the above by setting $\varepsilon = 0$. When $c \in L^1(\mu_0 \otimes \mu_1)$, EOT admits the following dual formulation,

$$\mathsf{OT}_\varepsilon(\mu_0, \mu_1) = \sup_{(\varphi_0,\varphi_1)\in L^1(\mu_0)\times L^1(\mu_1)} \int \varphi_0 d\mu_0 + \int \varphi_1 d\mu_1 - \varepsilon \int e^{\frac{\varphi_0 \oplus \varphi_1 - c}{\varepsilon}} d\mu_0 \otimes \mu_1 + \varepsilon,$$

where $\varphi_0 \oplus \varphi_1(x, y) = \varphi_0(x) + \varphi_1(y)$. For $\varepsilon > 0$, the set of solutions to the dual problem coincides with the set of solutions to the so-called Schrödinger system,

$$\begin{aligned} \int e^{\frac{\varphi_0(x)+\varphi_1(\cdot)-c(x,\cdot)}{\varepsilon}} d\mu_1 = 1, \quad \mu_0\text{-a.e. } x \in \mathbb{R}^{d_0}, \\ \int e^{\frac{\varphi_0(\cdot)+\varphi_1(y)-c(\cdot,y)}{\varepsilon}} d\mu_0 = 1, \quad \mu_1\text{-a.e. } y \in \mathbb{R}^{d_1}, \end{aligned} \tag{3}$$

for $(\varphi_0, \varphi_1) \in L^1(\mu_0) \times L^1(\mu_1)$. A pair $(\varphi_0, \varphi_1) \in L^1(\mu_0) \times L^1(\mu_1)$ solving (3) is known to be a.s. unique up to additive constants in the sense that any other solution $(\bar{\varphi}_0, \bar{\varphi}_1)$ satisfies $\bar{\varphi}_0 = \varphi_0 + a$ μ_0-a.s. and $\bar{\varphi}_1 = \varphi_1 - a$ μ_1-a.s. for some $a \in \mathbb{R}$. Moreover, the unique EOT coupling π_ε is characterized by

$$\frac{d\pi_\varepsilon}{d\mu_0 \otimes \mu_1}(x, y) = e^{\frac{\varphi_0(x)+\varphi_1(y)-c(x,y)}{\varepsilon}}, \tag{4}$$

and, under some additional conditions on the cost and marginals which hold throughout this paper, (3) admits a pair of continuous solutions which is unique up to additive constants and satisfies the system at all points $(x, y) \in \mathbb{R}^{d_0} \times \mathbb{R}^{d_1}$. We call such continuous solutions EOT potentials. The reader is referred to Nutz (2021) for a comprehensive overview of EOT.

2.2 Entropic Gromov-Wasserstein Distance

This work studies stability and computational aspects of the entropically regularized GW distance under the quadratic cost. By analogy to OT, EGW serves as a proxy of the standard (p, q)-GW distance, which quantifies discrepancy between complete and separable mm spaces $(\mathcal{X}_0, \mathsf{d}_0, \mu_0)$ and $(\mathcal{X}_1, \mathsf{d}_1, \mu_1)$ as (Mémoli, 2011; Sturm, 2012)

$$\mathsf{D}_{p,q}(\mu_0, \mu_1) \coloneqq \inf_{\pi\in\Pi(\mu_0,\mu_1)} \|\Gamma_q\|_{L^p(\pi\otimes\pi)},$$

where $\Gamma_q(x, y, x', y') = \left|\mathsf{d}_0^q(x, x') - \mathsf{d}_1^q(y, y')\right|$ is the distance distortion cost. This definition is the L^p-relaxation of the Gromov-Hausdorff distance between metric spaces,[2] and gives rise to a metric on the collection of all isomorphism classes of mm spaces with finite pq-size, i.e., such that $\int \mathsf{d}(x, x')^{pq}\, d\mu \otimes \mu(x, x') < \infty$.

From here on out, we instantiate the mm spaces as the Euclidean spaces $(\mathbb{R}^{d_i}, \|\cdot\|, \mu_i)$, for $i = 0, 1$, and focus on the EGW distance for the quadratic cost.

2. The Gromov-Hausdorff distance between $(\mathcal{X}_0, \mathsf{d}_0)$ and $(\mathcal{X}_1, \mathsf{d}_1)$ is given by $\frac{1}{2}\inf_{R\in\mathcal{R}(\mathcal{X}_0,\mathcal{X}_1)} \|\Gamma_1\|_{L^\infty(R)}$, where $\mathcal{R}(\mathcal{X}_0, \mathcal{X}_1)$ is the collection of all correspondence sets of $\mathcal{X}_0$ and $\mathcal{X}_1$, i.e., subsets $R \subset \mathcal{X}_0 \times \mathcal{X}_1$ such that the coordinate projection maps are surjective when restricted to R. The correspondence set can be thought of as $\mathrm{spt}(\pi)$ in the GW formulation.

Quadratic Cost The quadratic EGW distance, which corresponds to the $p = q = 2$ case, is defined as

$$\mathsf{S}_\varepsilon(\mu_0, \mu_1) = \inf_{\pi \in \Pi(\mu_0,\mu_1)} \int \left| \|x - x'\|^2 - \|y - y'\|^2 \right|^2 d\pi \otimes \pi(x, y, x', y') + \varepsilon \mathsf{D}_{\mathsf{KL}}(\pi \| \mu_0 \otimes \mu_1). \quad (5)$$

One readily verifies that, like the standard GW distance, EGW is invariant to isometric actions on the marginal spaces such as orthogonal rotations and translations. In addition, note that $\mathsf{S}_\varepsilon(\mu_0, \mu_1) = \varepsilon \mathsf{S}_1(\mu_0^\varepsilon, \mu_1^\varepsilon)$, where $\mu_i^\varepsilon = (\varepsilon^{-1/4} \operatorname{Id})_\sharp \mu_i$. In general, (5) is a non-convex quadratic program. Non-convexity can easily be discerned from the representation (6).

When μ_0, μ_1 are centered, which we may assume without loss of generality, the EGW distance decomposes[3] as (cf. Section 5.3 in Zhang et al. 2022a)

$$\mathsf{S}_\varepsilon(\mu_0, \mu_1) = \mathsf{S}^1(\mu_0, \mu_1) + \mathsf{S}^2_\varepsilon(\mu_0, \mu_1), \quad (6)$$

where

$$\mathsf{S}^1(\mu_0, \mu_1) = \int \|x - x'\|^4 d\mu_0 \otimes \mu_0(x, x') + \int \|y - y'\|^4 d\mu_1 \otimes \mu_1(y, y') - 4 \int \|x\|^2 \|y\|^2 d\mu_0 \otimes \mu_1(x, y),$$

$$\mathsf{S}^2_\varepsilon(\mu_0, \mu_1) = \inf_{\pi \in \Pi(\mu_0,\mu_1)} \int -4\|x\|^2\|y\|^2 d\pi(x, y) - 8 \sum_{\substack{1 \le i \le d_0 \\ 1 \le j \le d_1}} \left(\int x_i y_j \, d\pi(x, y) \right)^2 + \varepsilon \mathsf{D}_{\mathsf{KL}}(\pi \| \mu_0 \otimes \mu_1).$$

Evidently, S^1 depends only on the moments of the marginal distributions μ_0, μ_1, while S^2_ε captures the dependence on the coupling. A key observation in Zhang et al. (2022a) is that S^2_ε admits a variational form that ties it to the well understood EOT problem.

Lemma 1 (EGW duality; Theorem 1 in Zhang et al. 2022a)**.** *Fix $\varepsilon > 0$, $(\mu_0, \mu_1) \in \mathcal{P}_4(\mathbb{R}^{d_0}) \times \mathcal{P}_4(\mathbb{R}^{d_1})$, and let $M_{\mu_0,\mu_1} \coloneqq \sqrt{M_2(\mu_0) M_2(\mu_1)}$. Then, for any $M \ge M_{\mu_0,\mu_1}$,*

$$\mathsf{S}^2_\varepsilon(\mu_0, \mu_1) = \inf_{\boldsymbol{A} \in \mathcal{D}_M} 32\|\boldsymbol{A}\|_F^2 + \mathsf{OT}_{\boldsymbol{A},\varepsilon}(\mu_0, \mu_1), \quad (7)$$

where $\mathcal{D}_M \coloneqq \{\boldsymbol{A} \in \mathbb{R}^{d_0 \times d_1} : \|\boldsymbol{A}\|_F \le M/2\}$ and $\mathsf{OT}_{\boldsymbol{A},\varepsilon}(\mu_0, \mu_1)$ is the EOT problem with the cost function $c_{\boldsymbol{A}} : (x, y) \in \mathbb{R}^{d_0} \times \mathbb{R}^{d_1} \mapsto -4\|x\|^2\|y\|^2 - 32 x^\intercal \boldsymbol{A} y$ and regularization parameter ε. Moreover, the infimum is achieved at some $\boldsymbol{A}^\star \in \mathcal{D}_{M_{\mu_0,\mu_1}}$.

An analogous result holds in the unregularized ($\varepsilon = 0$) case, see Corollary 1 in Zhang et al. (2022a). The proof of Theorem 1 in Zhang et al. (2022a) demonstrates that if μ_0 and μ_1 are centered and $\pi_\star$ is optimal for the original EGW formulation, then $\boldsymbol{A}^\star = \frac{1}{2} \int x y^\intercal \, d\pi_\star(x, y)$ is optimal for (7) and $\pi_\star = \pi_{\boldsymbol{A}^\star}$, where $\pi_{\boldsymbol{A}^\star}$ is the unique EOT coupling for $\mathsf{OT}_{\boldsymbol{A}^\star,\varepsilon}(\mu_0, \mu_1)$. It follows from Jensen's inequality and the Cauchy-Schwarz inequality that $\boldsymbol{A}^\star \in \mathcal{D}_M$. Corollary 4 ahead expands on this connection by establishing a one-to-one correspondence between solutions of S_ε and S^2_ε and shows that all solutions of (7) lie in $\mathcal{D}_M$.

Although (7) illustrates a connection between the EGW and EOT problems, the outer minimization over $\mathcal{D}_M$ necessitates studying EOT with a parametrized cost function $c_{\boldsymbol{A}}$.

3. A similar decomposition holds for the inner product cost, obtained by replacing the squared Euclidean distances in Equation (5) by inner products. As such, all results derived in this manuscript apply to the inner product cost with minor modifications.

3 Stability of Entropic Gromov-Wasserstein Distances

We analyze the stability of the EGW problems with respect to (w.r.t.) the matrix $\boldsymbol{A}$ appearing in the dual formulation (7). Specifically, we characterize the first and second derivatives of the objective function whose optimization defines S_ε^2. These, in turn, elucidates its smoothness and convexity properties. Our stability analysis is later used to (i) gain insight into the structure of optimal couplings for the EGW problem; and (ii) devise novel approaches for computing the EGW distance with formal convergence guarantees.

Throughout this section, we restrict attention to compactly supported distributions, as some of the technical details do not directly translate to the unbounded setting (e.g., the proof of Lemma 18). For a Fréchet differentiable map $F : U \to V$ between normed vector spaces U and V,[4] we denote the derivative of F at the point $u \in U$ evaluated at $v \in V$ by $DF_{[u]}(v)$.

Fix compactly supported distributions $(\mu_0, \mu_1) \in \mathcal{P}(\mathbb{R}^{d_0}) \times \mathcal{P}(\mathbb{R}^{d_1})$ and some $\varepsilon > 0$. Let

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

denote the objective in (7). We first characterize the derivatives of Φ and then prove that this map is weakly convex and L-smooth.[5]

Proposition 2 (First and second derivatives)**.** *The map $\Phi : \boldsymbol{A} \in \mathbb{R}^{d_0 \times d_1} \mapsto 32\|\boldsymbol{A}\|_F^2 + \mathsf{OT}_{\boldsymbol{A},\varepsilon}(\mu_0, \mu_1)$ is smooth, coercive, and has first and second-order Fréchet derivatives at $\boldsymbol{A} \in \mathbb{R}^{d_0 \times d_1}$ given by*

$$D\Phi_{[\boldsymbol{A}]}(\boldsymbol{B}) = 64\,\mathrm{tr}(\boldsymbol{A}^\intercal \boldsymbol{B}) - 32 \int x^\intercal \boldsymbol{B} y \, d\pi_{\boldsymbol{A}}(x, y),$$

$$D^2\Phi_{[\boldsymbol{A}]}(\boldsymbol{B}, \boldsymbol{C}) = 64\,\mathrm{tr}(\boldsymbol{B}^\intercal \boldsymbol{C}) + 32\varepsilon^{-1} \int x^\intercal \boldsymbol{B} y \left(h_0^{\boldsymbol{A},\boldsymbol{C}}(x) + h_1^{\boldsymbol{A},\boldsymbol{C}}(y) - 32 x^\intercal \boldsymbol{C} y \right) d\pi_{\boldsymbol{A}}(x, y),$$

where $\boldsymbol{B}, \boldsymbol{C} \in \mathbb{R}^{d_0 \times d_1}$, $\pi_{\boldsymbol{A}}$ is the unique EOT coupling for $\mathsf{OT}_{\boldsymbol{A},\varepsilon}(\mu_0, \mu_1)$, and $\left(h_0^{\boldsymbol{A},\boldsymbol{C}}, h_1^{\boldsymbol{A},\boldsymbol{C}}\right)$ is the unique (up to additive constants) pair of functions in $\mathcal{C}(\mathrm{spt}(\mu_0)) \times \mathcal{C}(\mathrm{spt}(\mu_1))$ satisfying

$$\begin{aligned} &\int \left(h_0^{\boldsymbol{A},\boldsymbol{C}}(x) + h_1^{\boldsymbol{A},\boldsymbol{C}}(y) - 32 x^\intercal \boldsymbol{C} y \right) e^{\frac{\varphi_0^{\boldsymbol{A}}(x) + \varphi_1^{\boldsymbol{A}}(y) - c_{\boldsymbol{A}}(x,y)}{\varepsilon}} d\mu_1(y) = 0, \quad \forall\, x \in \mathrm{spt}(\mu_0), \\ &\int \left(h_0^{\boldsymbol{A},\boldsymbol{C}}(x) + h_1^{\boldsymbol{A},\boldsymbol{C}}(y) - 32 x^\intercal \boldsymbol{C} y \right) e^{\frac{\varphi_0^{\boldsymbol{A}}(x) + \varphi_1^{\boldsymbol{A}}(y) - c_{\boldsymbol{A}}(x,y)}{\varepsilon}} d\mu_0(x) = 0, \quad \forall\, y \in \mathrm{spt}(\mu_1). \end{aligned} \tag{8}$$

Here, $(\varphi_0^{\boldsymbol{A}}, \varphi_1^{\boldsymbol{A}})$ is any pair of EOT potentials for $\mathsf{OT}_{\boldsymbol{A},\varepsilon}(\mu_0, \mu_1)$.

Proposition 2 essentially follows from the implicit mapping theorem, which enables us to compute the Fréchet derivative of the EOT potentials for $\mathsf{OT}_{(\cdot),\varepsilon}(\mu_0, \mu_1)$ using the Schrödinger system (3). The derivative of $\mathsf{OT}_{(\cdot),\varepsilon}(\mu_0, \mu_1)$, which is a simple function of the

4. A map $F : U \to V$ is Fréchet differentiable at $u \in U$ if there exists a bounded linear operator $A : U \to V$ for which $F(u + h) = F(u) + Ah + o(h)$ as $h \to 0$. If such an A exists, it is called the Fréchet derivative to F at u.
5. A function $f : \mathbb{R}^d \to \mathbb{R}$ is ρ-weakly convex if $f + \frac{\rho}{2}\|\cdot\|^2$ is convex; f is L-smooth if its gradient is L-Lipschitz, i.e., $\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|$, for all $x, y \in \mathbb{R}^d$.

EOT potentials, is then readily obtained. By differentiating the Frobenius norm, this further yields the derivative of Φ. See Section 5.1 for details.

The following remark clarifies that the first and second Fréchet derivatives of Φ can be identified with the gradient and Hessian of a function on $\mathbb{R}^{d_0 d_1}$. Recall that the Frobenius inner product, $\langle \boldsymbol{A}, \boldsymbol{B} \rangle_F = \operatorname{tr}(\boldsymbol{A}^\intercal \boldsymbol{B}) = \sum_{\substack{1 \le i \le d_0 \\ 1 \le j \le d_1}} \boldsymbol{A}_{ij} \boldsymbol{B}_{ij}$, is simply the Euclidean inner product between the vectorized matrices $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{d_0 \times d_1}$.

Remark 3 (Interpreting derivatives as gradient/Hessian)**.** *The first derivative of Φ from Proposition 2 can be written as*

$$D\Phi_{[\boldsymbol{A}]}(\boldsymbol{B}) = \left\langle 64\boldsymbol{A} - 32 \int xy^\intercal \, d\pi_{\boldsymbol{A}}(x,y), \boldsymbol{B} \right\rangle_F .$$

Recall that if f is a continuously differentiable function f on $\mathbb{R}^d$, its directional derivative at x along the direction y is $Df_{[x]}(y) = \langle \nabla f(x), y \rangle$, for $x, y \in \mathbb{R}^d$. By analogy, we may think of $D\Phi_{[\boldsymbol{A}]}$ as $64\boldsymbol{A} - 32 \int xy^\intercal \, d\pi_{\boldsymbol{A}}(x,y)$. This perspective is utilized in Section 4 when studying computational guarantees for the EGW distance, as it is simpler to view iterates as matrices rather than abstract linear operators. By the same token, the second derivative of Φ at $\boldsymbol{A} \in \mathbb{R}^{d_0 \times d_1}$ is a bilinear form on $\mathbb{R}^{d_0 \times d_1}$ and hence can be identified with a $d_0 d_1 \times d_0 d_1$ matrix by analogy with the Hessian.

As a direct corollary to Proposition 2, we provide an (implicit) characterization of the stationary points of Φ and connect its minimizers to solutions of S_ε. Details are provided in Section 5.2.

Corollary 4 (Stationary points and correspondence between S_ε and S_ε^2)**.**

(i) A matrix $\boldsymbol{A} \in \mathbb{R}^{d_0 \times d_1}$ is a stationary point of Φ if and only if $\boldsymbol{A} = \frac{1}{2} \int xy^\intercal \, d\pi_{\boldsymbol{A}}(x,y)$. As Φ is coercive, all minimizers of Φ are stationary points and hence contained in $\mathcal{D}_{M_{\mu_0,\mu_1}}$.

(ii) If μ_0 and μ_1 are centered, then a given matrix $\boldsymbol{A}$ minimizes Φ if and only if $\pi_{\boldsymbol{A}}$ is optimal for S_ε and satisfies $\frac{1}{2} \int xy^\intercal \, d\pi_{\boldsymbol{A}}(x,y) = \boldsymbol{A}$.

(iii) Suppose μ_0 and μ_1 are centered. If S_ε admits a unique optimal coupling $\pi_\star$, then Φ admits a unique minimizer $\boldsymbol{A}^\star$ and $\pi_\star = \pi_{\boldsymbol{A}^\star}$. Conversely, if Φ admits a unique minimizer $\boldsymbol{A}^\star$, then $\pi_{\boldsymbol{A}^\star}$ is a unique optimal coupling for S_ε.

Although the second derivative of Φ involves the implicitly defined functions $(h_0^{\boldsymbol{A},\boldsymbol{C}}, h_1^{\boldsymbol{A},\boldsymbol{C}})$, its maximal and minimal eigenvalues admit the following explicit bounds.

Corollary 5 (Hessian eigenvalue bounds)**.** *The following hold for any $\boldsymbol{A} \in \mathbb{R}^{d_0 \times d_1}$:*

(i) The minimal eigenvalue of $D^2\Phi_{[\boldsymbol{A}]}$ satisfies

$$\begin{aligned} \lambda_{\min}\left(D^2\Phi_{[\boldsymbol{A}]}\right) &= 64 + \varepsilon^{-1} \inf_{\|\boldsymbol{C}\|_F = 1} \int \left[\left(h_0^{\boldsymbol{A},\boldsymbol{C}}(x) + h_1^{\boldsymbol{A},\boldsymbol{C}}(y) \right)^2 - 32^2 (x^\intercal \boldsymbol{C} y)^2 \right] d\pi_{\boldsymbol{A}}(x,y) \\ &\ge 64 - 32^2 \varepsilon^{-1} \sup_{\|\boldsymbol{C}\|_F = 1} \operatorname{Var}_{\pi_{\boldsymbol{A}}}(X^\intercal \boldsymbol{C} Y), \end{aligned}$$

where the variance term admits the uniform bound $\sup_{\|\boldsymbol{C}\|_F = 1} \operatorname{Var}_{\pi_{\boldsymbol{A}}}(X^\intercal \boldsymbol{C} Y) \le \sqrt{M_4(\mu_0) M_4(\mu_1)}$.

(ii) The maximal eigenvalue of $D^2\Phi_{[\boldsymbol{A}]}$ *satisfies* $\lambda_{\max}\left(D^2\Phi_{[\boldsymbol{A}]}\right) \leq 64$.

Corollary 5 follows from Proposition 2 by considering the variational form of the maximal and minimal eigenvalues; see Section 5.3 for details. We note that, in general, the variance bound in Item (i) is sharp up to constants in arbitrary dimensions. For example, it is attained up to a factor of 2 by $\mu_0 = \frac{1}{2}(\delta_0 + \delta_a)$ and $\mu_1 = \frac{1}{2}(\delta_0 + \delta_b)$ for $a \in \mathbb{R}^{d_0}$ and $b \in \mathbb{R}^{d_1}$; see Appendix A for the full variance computation.

Armed with the eigenvalue bounds, we now state the main result of this section addressing convexity and smoothness of Φ.

Theorem 6 (Convexity and L-smoothness)**.** *The map* Φ *is weakly convex with parameter at most* $32^2\varepsilon^{-1}\sqrt{M_4(\mu_0)M_4(\mu_1)} - 64$ *and, if* $\sqrt{M_4(\mu_0)M_4(\mu_1)} < \frac{\varepsilon}{16}$*, then it is strictly convex and admits a unique minimizer. Moreover, for any* $M > 0$*,* Φ *is* L*-smooth on* $\mathcal{D}_M$ *with*

$$L = \max_{\boldsymbol{A}\in\mathcal{D}_M} \lambda_{\max}\left(D^2\Phi_{[\boldsymbol{A}]}\right) \vee \left(-\lambda_{\min}\left(D^2\Phi_{[\boldsymbol{A}]}\right)\right) \leq 64 \vee \left(32^2\varepsilon^{-1}\sqrt{M_4(\mu_0)M_4(\mu_1)} - 64\right).$$

Theorem 6 shows that Φ is amenable to optimization by accelerated gradient methods with step size L and establishes sufficient conditions to guarantee convergence of these algorithms to a global minimizer (i.e., convexity of Φ). In general, optimal EGW couplings may not be unique. Theorem 6 provides sufficient conditions for uniqueness of solutions to both (7) and the EGW problem by the connection discussed in Corollary 4 when the marginals are centered. When the optimal EGW coupling is unique, symmetries in the marginal spaces result in certain structural properties for the optimal $\boldsymbol{A}^\star$ in (7). The following remark expands on these connections.

Remark 7 (Symmetries and uniqueness of couplings)**.** *Fix* $\varepsilon > 0$ *and a pair of centered distributions* $(\mu_0, \mu_1) \in \mathcal{P}_4(\mathbb{R}^{d_0}) \times \mathcal{P}_4(\mathbb{R}^{d_1})$*. Assume that* Φ *admits a unique minimizer* $\boldsymbol{A}^\star$ *and let* $\pi_{\boldsymbol{A}^\star}$ *be the associated EOT/EGW coupling (e.g., under the conditions of Theorem 6, given that* μ_0, μ_1 *are compactly supported). If, for* $i = 0, 1$*,* μ_i *is invariant under the action of the orthogonal transformation* $U_i : \mathbb{R}^{d_i} \to \mathbb{R}^{d_i}$ *in the sense that* $(U_i)_\sharp\mu_i = \mu_i$ *it follows that* $(U_0, U_1)_\sharp\pi_{\boldsymbol{A}^\star}$ *is also an optimal EGW coupling, whence* $(U_0, U_1)_\sharp\pi_{\boldsymbol{A}^\star} = \pi_{\boldsymbol{A}^\star}$ *by uniqueness. Thus, by Corollary 4,* $\boldsymbol{U}_0^\intercal\boldsymbol{A}^\star\boldsymbol{U}_1 = \boldsymbol{A}^\star$ *where* $\boldsymbol{U}_i$ *is the matrix associated with* U_i*, for* $i = 0, 1$*. The previous equality holds for any pair of orthogonal transformations leaving the marginals invariant, so the rows of* $\boldsymbol{A}^\star$ *are left eigenvectors of* $\boldsymbol{U}_1$ *with eigenvalue* 1 *and its columns are right eigenvectors of* $\boldsymbol{U}_0^\intercal$ *with eigenvalue* 1 *for every* $\boldsymbol{U}_i$ *such that* $(U_i)_\sharp\mu_i = \mu_i$*. Thus, we see that symmetries of the marginals dictate the structure of* $\boldsymbol{A}^\star$*. For example, if* $\mu_1 = (-\mathrm{Id})_\sharp\mu_1$*, we have that* $\boldsymbol{A}^\star = -\boldsymbol{A}^\star$*, so* $\boldsymbol{A}^\star = \boldsymbol{0}$.

4 Computational Guarantees

Building on the stability theory from Section 3, we now study computation of the EGW problem. The goal is to compute the distance between two discrete distributions $\mu_0 \in \mathcal{P}(\mathbb{R}^{d_0})$ and $\mu_1 \in \mathcal{P}(\mathbb{R}^{d_1})$ supported on N_0 and N_1 atoms $(x^{(i)})_{i=1}^{N_0}$ and $(y^{(j)})_{j=1}^{N_1}$, respectively. In light of the decomposition (6), we focus on S_ε^2, which is given by a smooth optimization problem whose convexity depends on the value of ε (cf. Section 3). Throughout, we adopt the perspective of Remark 3 and treat $D\Phi_{[\boldsymbol{A}]}$, for $\boldsymbol{A} \in \mathbb{R}^{d_0\times d_1}$, as the matrix $64\boldsymbol{A} - 32\int xy^\intercal\, d\pi_{\boldsymbol{A}}(x, y)$.

4.1 Inexact Oracle Methods

As these problems are already $d_0 d_1$-dimensional and computing the second Fréchet derivative of Φ may be infeasible (in particular, it requires solving (8)), we focus on first-order methods. Given the regularity of the S^2_ε optimization problem, standard out-of-the-box numerical routines are likely to yield good results in practice. However, to provide meaningful formal guarantees one must account for the fact that evaluation of Φ and its gradient, for $\boldsymbol{A} \in \mathcal{D}_M$, requires computing the corresponding EOT plan, which often entails an approximation. Indeed, an explicit characterizations of the EOT plan between arbitrary distributions is unknown and algorithms typically rely on a fast numerical proxy of the coupling. We model this under the scope of gradient methods with inexact gradient oracles (d'Aspremont, 2008; Devolder et al., 2014; Dvurechensky, 2017).

For a fixed $\varepsilon > 0$ and μ_0, μ_1 as above, our goal is thus to solve

$$\min_{\boldsymbol{A} \in \mathcal{D}_M} 32\|\boldsymbol{A}\|_F^2 + \mathsf{OT}_{\boldsymbol{A},\varepsilon}(\mu_0, \mu_1),$$

where $M > M_{\mu_0,\mu_1}$, which guarantees that all the optimizers are within the optimization domain (cf. Corollary 4). As we are in the discrete setting, the EOT coupling $\pi^{\boldsymbol{A}}$ for $\mathsf{OT}_{\boldsymbol{A},\varepsilon}(\mu_0, \mu_1)$, $\boldsymbol{A} \in \mathcal{D}_M$, is represented by $\boldsymbol{\Pi}^{\boldsymbol{A}} \in \mathbb{R}^{N_0 \times N_1}$, where $\boldsymbol{\Pi}^{\boldsymbol{A}}_{ij} = \pi^{\boldsymbol{A}}(x^{(i)}, y^{(j)})$. The inexact oracle paradigm assumes that, for any $\boldsymbol{A} \in \mathcal{D}_M$, we have access to a δ-oracle approximation $\widetilde{\boldsymbol{\Pi}}^{\boldsymbol{A}}$ of $\boldsymbol{\Pi}^{\boldsymbol{A}}$ with $\|\widetilde{\boldsymbol{\Pi}}^{\boldsymbol{A}} - \boldsymbol{\Pi}^{\boldsymbol{A}}\|_\infty < \delta$. Such oracles can be obtained, for instance, by numerical resolution of the EOT problem. To this end, Sinkhorn's algorithm (Sinkhorn, 1967; Cuturi, 2013) serves as the canonical approach.

Proposition 8 (Inexact oracle via Sinkhorn iterations)**.** *Fix $\delta > 0$. Then, Sinkhorn's algorithm (Algorithm 3) returns an $(e^\delta - 1)$-oracle approximation $\widetilde{\boldsymbol{\Pi}}^{\boldsymbol{A}}$ of $\boldsymbol{\Pi}^{\boldsymbol{A}}$ in at most $\tilde{k}$ iterations, where $\tilde{k}$ depends only on $\mu_0, \mu_1, \boldsymbol{A}, \delta$, and ε, and is given explicitly in* (28).

The proof of Proposition 8 follows by combining a number of known results. Complete details can be found in Appendix B. To our knowledge, the majority of the literature concerning the use of Sinkhorn's algorithm for EOT focuses on approximating unregularized OT and treats the KL divergence term as a bias. Here we quantify the accuracy of estimating the true EOT plan, which may be of an independent interest.

With these preparations, we first discuss the case where Φ is known to be convex on $\mathcal{D}_M$.

4.2 Convex Case

Assume that Φ is convex on $\mathcal{D}_M$, e.g., under the setting of Theorem 6. As convexity implies that the minimal eigenvalue of $D^2\Phi_{[\boldsymbol{A}]}$ is positive for any $\boldsymbol{A} \in \mathcal{D}_M$, Theorem 6 further yields that Φ is 64-smooth. With that, we can the apply inexact oracle first-order method from d'Aspremont (2008). To describe the approach, assume that we are given a δ-oracle $\widetilde{\boldsymbol{\Pi}}^{\boldsymbol{A}}$ for the EOT plan $\boldsymbol{\Pi}^{\boldsymbol{A}}$ for $\mathsf{OT}_{\boldsymbol{A},\varepsilon}(\mu_0, \mu_1)$, and define the corresponding gradient approximation

$$\widetilde{D}\Phi_{[\boldsymbol{A}]} = 64\boldsymbol{A} - 32 \sum_{\substack{1 \le i \le N_0 \\ 1 \le j \le N_1}} x^{(i)} (y^{(j)})^\intercal \widetilde{\boldsymbol{\Pi}}^{\boldsymbol{A}}_{ij}. \tag{9}$$

We now present the algorithm and follow it with formal convergence guarantees.

Algorithm 1 Fast gradient method with inexact oracle

Fix $L = 64$ and let $\alpha_k = \frac{k+1}{2}$, and $\tau_k = \frac{2}{k+3}$
$k \leftarrow 0$
$\boldsymbol{A}_0 \leftarrow \mathbf{0}$
$\boldsymbol{G}_0 \leftarrow \widetilde{D}\Phi_{[\boldsymbol{A}_0]}$
$\boldsymbol{W}_0 \leftarrow \alpha_0 \boldsymbol{G}_0$
while stopping condition is not met **do**
  $\boldsymbol{B}_k \leftarrow \min\left(1, \frac{M}{2\|\boldsymbol{A}_k - L^{-1}\boldsymbol{G}_k\|_F}\right)(\boldsymbol{A}_k - L^{-1}\boldsymbol{G}_k)$
  $\boldsymbol{C}_k \leftarrow -\min\left(1, \frac{M}{2\|L^{-1}\boldsymbol{W}_k\|_F}\right) L^{-1}\boldsymbol{W}_k$
  $\boldsymbol{A}_{k+1} \leftarrow \tau_k \boldsymbol{C}_k + (1-\tau_k)\boldsymbol{B}_k$
  $\boldsymbol{G}_{k+1} \leftarrow \widetilde{D}\Phi_{[\boldsymbol{A}_{k+1}]}$
  $\boldsymbol{W}_{k+1} \leftarrow \boldsymbol{W}_k + \alpha_{k+1}\boldsymbol{G}_{k+1}$
  $k \leftarrow k+1$
return $\boldsymbol{B}_k$

The multiplication operations in Algorithm 1 are applied entrywise and it is understood that $\min(1, M/0) = 1$. Due to inexactness, stopping conditions based on insufficient progress of functions values or setting a threshold on the norm of the gradient require care. A condition based on the number of iterations is discussed in Remark 10.

We now provide formal convergence guarantees for Algorithm 1.

Theorem 9 (Fast convergence rates)**.** *Assume that Φ is convex and L-smooth on $\mathcal{D}_M$ and that $\widetilde{\mathbf{\Pi}}^{\boldsymbol{A}}$ is a δ-oracle for $\mathbf{\Pi}^{\boldsymbol{A}}$. Then, the iterates $\boldsymbol{B}_k$ in Algorithm 1 with $\widetilde{D}\Phi_{[\boldsymbol{A}_k]}$ given by* (9) *satisfy*

$$\Phi(\boldsymbol{B}_k) - \Phi(\boldsymbol{B}^\star) \leq \frac{2L\|\boldsymbol{B}^\star\|_F^2}{(k+1)(k+2)} + 3\delta', \tag{10}$$

where $\boldsymbol{B}^\star$ is a global minimizer of Φ and $\delta' = 32M\delta \sum_{\substack{1\leq i\leq N_0 \\ 1\leq j\leq N_1}} \left\|x^{(i)}\right\| \left\|y^{(j)}\right\|$. Moreover, for any $\eta > 3\delta'$, Algorithm 1 requires at most

$$k = \left\lceil -\frac{3}{2} + \frac{1}{2}\sqrt{1 + \frac{8L\|\boldsymbol{B}^\star\|_F^2}{\eta - 3\delta'}} \right\rceil \leq \left\lceil -\frac{3}{2} + \frac{1}{2}\sqrt{1 + \frac{128M^2}{\eta - 3\delta'}} \right\rceil \tag{11}$$

iterations to achieve an η-approximate solution.

The proof of Theorem 9, given in Section 5.5, follows from Theorem 2.2 in d'Aspremont (2008) after casting our problem as an instance of their setting. Some implications of Theorem 9 are discussed in the following remark.

Remark 10 (Optimal rates and stopping conditions)**.** *First, consider the convergence rate of the function values in* (10)*. The first term on the right-hand side exhibits the optimal complexity bound for smooth constrained optimization of $O(1/k^2)$ (cf., e.g., Nesterov 2003). The second term accounts for the underlying oracle error. Notably, the progress of the optimization procedure and the oracle error are completely decoupled in this bound.*

Next, we describe a stopping condition based on the number of iterations. Observe that all terms involved in the upper bound in (11) *are explicit as soon as a desired precision η is chosen since the oracle error δ can be fixed according to Proposition 8. Consequently,* (11) *can be used as an explicit stopping condition for Algorithm 1.*

4.3 General Case

We now discuss an optimization procedure which does not require convexity of the objective. This accounts for the fact that outside the sufficient conditions of Theorem 6, convexity of Φ is generally unknown. However, the same result shows that Φ is L-smooth with $L = 64 \vee \left(32^2\varepsilon^{-1}\sqrt{M_4(\mu_0)M_4(\mu_1)} - 64\right)$ and $\mathsf{OT}_{(\cdot),\varepsilon}$ is L'-smooth with $L' = 32^2\varepsilon^{-1}\sqrt{M_4(\mu_0)M_4(\mu_1)}$. Hence, we adapt the smooth non-convex optimization routine of Ghadimi and Lan (2016) to account for our inexact oracle. Notably, their method adapts to the convexity of Φ as described in Theorem 11.

We now present the algorithm and describe its convergence rate.

Algorithm 2 Adaptive gradient method with inexact oracle

Given $\boldsymbol{C}_0 \in \mathcal{D}_M$, fix the step sequences $\beta_k = \frac{1}{2L}$, $\gamma_k = \frac{k}{4L}$, and $\tau_k = \frac{2}{k+2}$.

$k \leftarrow 1$
$\boldsymbol{A}_1 \leftarrow \boldsymbol{C}_0$
$\boldsymbol{G}_1 \leftarrow \widetilde{D}\Phi_{[\boldsymbol{A}_1]}$
while stopping condition is not met **do**
  $\boldsymbol{B}_k \leftarrow \min\left(1, \frac{M}{2\|\boldsymbol{A}_k - \beta_k \boldsymbol{G}_k\|_F}\right)(\boldsymbol{A}_k - \beta_k \boldsymbol{G}_k)$
  $\boldsymbol{C}_k \leftarrow \min\left(1, \frac{M}{2\|\boldsymbol{C}_{k-1} - \gamma_k \boldsymbol{G}_k\|_F}\right)(\boldsymbol{C}_{k-1} - \gamma_k \boldsymbol{G}_k)$
  $\boldsymbol{B}_k \leftarrow \frac{M}{2}\operatorname{sign}(\boldsymbol{A}_k - \beta_k \boldsymbol{G}_k)\min\left(\frac{2}{M}|\boldsymbol{A}_k - \beta_k \boldsymbol{G}_k|, 1\right)$
  $\boldsymbol{C}_k \leftarrow \frac{M}{2}\operatorname{sign}(\boldsymbol{C}_{k-1} - \gamma_k \boldsymbol{G}_k)\min\left(\frac{2}{M}|\boldsymbol{C}_{k-1} - \gamma_k \boldsymbol{G}_k|, 1\right)$
  $\boldsymbol{A}_{k+1} \leftarrow \tau_k \boldsymbol{C}_k + (1 - \tau_k)\boldsymbol{B}_k$
  $\boldsymbol{G}_{k+1} \leftarrow \widetilde{D}\Phi_{[\boldsymbol{A}_{k+1}]}$
  $k \leftarrow k + 1$
return $\boldsymbol{B}_k$

Unlike Algorithm 1, which can be initialized at any fixed $\boldsymbol{A}_0$, the starting point in Algorithm 2 should be chosen according to some selection rule that avoids initializing at a stationary point (e.g., random initialization). Indeed, if $\boldsymbol{A}_1$ is set to a stationary point of Φ, then $D\Phi_{[\boldsymbol{A}_1]} = \boldsymbol{0}$ and, consequently $\widetilde{D}\Phi_{[\boldsymbol{A}_1]} \approx \boldsymbol{0}$ (given that the approximate gradient is reasonably accurate), which may result in premature and undesirable termination. Clearly, this is not a concern for Algorithm 1 since it assumes convexity of Φ, whereby any stationary point is a global optimum.

The following result follows by adapting the proofs of Theorem 2 and Corollary 2 in Ghadimi and Lan (2016). For completeness, we provide a self-contained argument in Appendix C along with a discussion of how this problem fits in the framework of Ghadimi and Lan (2016).

Theorem 11 (Adaptive convergence rate)**.** *Assume that Φ is L-smooth on $\mathcal{D}_M$ and that $\widetilde{\mathbf{\Pi}}^{\boldsymbol{A}}$ is a δ-oracle for $\mathbf{\Pi}^{\boldsymbol{A}}$. Then, the iterates $\boldsymbol{A}_k, \boldsymbol{B}_k$ in Algorithm 2 with $\widetilde{D}\Phi_{[\boldsymbol{A}_k]}$ given by (9) satisfy*

1. *If Φ is non-convex and $\mathsf{OT}_{(\cdot),\varepsilon}(\mu_0, \mu_1)$ is L'-smooth, then*

$$\min_{1\le i\le k} \left\|\beta_i^{-1}(\boldsymbol{B}_i - \boldsymbol{A}_i)\right\|_F^2 \le \frac{96L^2}{k(k+1)(k+2)}\|\boldsymbol{C}_0 - \boldsymbol{B}^\star\|_F^2 + \frac{24LL'}{k}\left(\|\boldsymbol{B}^\star\|_F^2 + \frac{5M^2}{16}\right) + 8L\delta',$$

   where $\boldsymbol{B}^\star$ is a global minimizer of Φ, and $\delta' = 32M\delta \sum_{\substack{1\le i\le N_0\\ 1\le j\le N_1}} \left\|x^{(i)}\right\| \left\|y^{(j)}\right\|$.

2. *If Φ is convex, then*

$$\min_{1\le i\le k} \left\|\beta_i^{-1}\left(\boldsymbol{B}_i - \boldsymbol{A}_i\right)\right\|_F^2 \le \frac{96L^2}{k(k+1)(k+2)}\|\boldsymbol{C}_0 - \boldsymbol{B}^\star\|_F^2 + 8L\delta'.$$

To better motivate this result, we show that when $\left\|\beta_k^{-1}\left(\boldsymbol{B}_k - \boldsymbol{A}_k\right)\right\|_F$ is small, $D\Phi_{[\boldsymbol{A}_k]}$ is approximately stationary.

Corollary 12 (Approximate stationarity)**.** *Let $\boldsymbol{A}_k, \boldsymbol{B}_k$ be iterates from Algorithm 2 and assume that $\boldsymbol{B}_k \in \operatorname{int}(\mathcal{D}_M)$. Then,*

$$\|D\Phi_{[\boldsymbol{A}_k]}\|_F < 32\delta \sum_{\substack{1\le i\le N_0\\ 1\le j\le N_1}} \|x^{(i)}\|\|y^{(j)}\| + \left\|\beta_k^{-1}(\boldsymbol{B}_k - \boldsymbol{A}_k)\right\|_F.$$

The proof of Corollary 12 follows from the δ-oracle assumption and the fact that when $\boldsymbol{B}_k$ is an interior point of $\mathcal{D}_M$, we have $\boldsymbol{B}_k = \boldsymbol{A}_k - \beta_k \boldsymbol{G}_k$. See Section 5.6 for the full details. When $\boldsymbol{B}_k$ is not an interior point of $\mathcal{D}_M$, the interpretation of $\|\beta_k^{-1}(\boldsymbol{B}_k - \boldsymbol{A}_k)\|_F$ is less straightforward. However, as all stationary points of Φ are contained in $\mathcal{D}_{M_{\mu_0,\mu_1}}$, it is expected that Algorithm 2 will converge to an interior point when $M > M_{\mu_0,\mu_1}$. By analogy with Remark 10, when all iterates are interior points Algorithm 2 yields a bound on the total number of iterations required to achieve an approximate stationary point.

The following remark addresses the distinctions between the convex and non-convex settings in Theorem 11.

Remark 13 (Adaptivity of Algorithm 2)**.** *As in Theorem 9, the convergence rates in Theorem 11 are decoupled into a term related to the progress of the optimization procedure and a term related to the oracle error.*

In the case where Φ is non-convex, the dominant term in the optimization error is $O(1/k)$, which coincides with the best known rates for solving general unconstrained nonlinear programs (Ghadimi and Lan, 2016). On the other hand, when Φ is convex, the rate of convergence improves to $O(1/k^3)$ which essentially matches the best known rates for the norm of the gradient in the unconstrained accelerated gradient method applied to a convex L-smooth function (see Theorem 6 in Shi et al. (2021) and Theorem 3.1 in Chen et al. (2022)[6]). This adaptivity is beneficial, as Φ may be convex beyond the conditions derived in Theorem 6.

6. More precisely, Chen et al. (2022) show that the iterates $(y_i)_{i=1}^k$ generated by the accelerated gradient method applied to a convex L-smooth function f are such that $\min_{0\le i\le k}\|\nabla f(y_i)\|^2 = o(1/k^3)$.

An empirical comparison of Algorithms 1 and 2 in the convex setting is included in Section 4.5. In particular, Algorithm 1 is seen to outperform Algorithm 2 in terms of average runtime despite having the same per iteration complexity when the inexact gradient is computed using standard Sinkhorn iterations.

Remark 14 (Computational complexity of Algorithms 1 and 2)**.** *As Sinkhorn's algorithm is known to have a complexity of $O(N_0N_1)$ (cf. e.g. Scetbon et al. 2022), the gradient approximation* (9) *can be computed in $O(N_0N_1)$ time. It follows that Algorithms 1 and 2 admit a computational complexity of $O(N_0N_1)$.*

4.4 Approximating Unregularized Gromov-Wasserstein Distances

The EGW distance can approximate unregularized GW to an arbitrary precision, with error $|\mathsf{S}_\varepsilon(\mu_0,\mu_1) - \mathsf{S}_0(\mu_0,\mu_1)| = O\left(\varepsilon \log\left(\frac{1}{\varepsilon}\right)\right)$ as $\varepsilon \downarrow 0$ (see Proposition 1 in Zhang et al. (2022a)). It is therefore natural to ask if the proposed algorithms can be used to approximate the unregularized GW distance between finitely supported marginals. Note, however, that for $\varepsilon > 0$ sufficiently small, Φ may fail to be convex (see Theorem 6), whence Algorithm 2 may only converge to an approximate stationary point of Φ. As such, it is not guaranteed that itself $\mathsf{S}_\varepsilon(\mu_0,\mu_1)$ (i.e., the global minimum of the entropic problem) can be approximated to within a desired accuracy. Nevertheless, we show that these approximate stationary points can be used to capture a notion of local optimality for $\mathsf{S}_0^2(\mu_0,\mu_1)$, keeping in mind that the quadratic GW distance decomposes as $\mathsf{S}_0(\mu_0,\mu_1) = \mathsf{S}^1(\mu_0,\mu_1) + \mathsf{S}_0^2(\mu_0,\mu_1)$ for centered distributions (see Corollary 1 in Zhang et al. (2022a)).

As opposed to EOT, optimal solutions to unregularized OT may fail to be unique, which entails that $\Phi_0 \coloneqq 32\|\cdot\|_F^2 + \mathsf{OT}_{(\cdot),0}(\mu_0,\mu_1)$ may be non-differentiable. Remarkably, the following analogue of Corollary 4 on stationary points still holds.

Proposition 15 (Optimality conditions for Φ_0)**.** *Let $(\mu_0,\mu_1) \in \mathcal{P}(\mathbb{R}^{d_0})\times\mathcal{P}(\mathbb{R}^{d_1})$ be compactly supported. If $\bar{\boldsymbol{A}}$ is a local minimizer of Φ_0, then there exists a solution $\bar\pi$ to $\mathsf{OT}_{\bar{\boldsymbol{A}},0}(\mu_0,\mu_1)$ with $\bar{\boldsymbol{A}} = \frac{1}{2}\int xy^\intercal d\bar\pi(x,y) \in \mathcal{D}_M$ for any $M > M_{\mu_0,\mu_1}$. If $\bar{\boldsymbol{A}}$ is globally optimal and μ_0,μ_1 are centered, then $\bar\pi$ solves $\mathsf{S}_0(\mu_0,\mu_1)$.*

The proof of Proposition 15, which will appear in Section 5.7, follows similar lines to the proofs provided in the regularized case with the caveat that Φ_0 is merely locally Lipschitz. To adapt these results, we characterize the Clarke subdifferential of Φ_0 (Clarke, 1975).

Proposition 15 shows that any minimizer, $\bar{\boldsymbol{A}}_0 \in \mathcal{D}_M$, of Φ_0 satisfies an analogous criterion to the stationary points, $\bar{\boldsymbol{A}}_\varepsilon \in \mathcal{D}_M$, of $\Phi_\varepsilon \coloneqq 32\|\cdot\|_F^2 + \mathsf{OT}_{(\cdot),\varepsilon}(\mu_0,\mu_1)$ for $\varepsilon > 0$ (namely, that $\bar{\boldsymbol{A}}_\varepsilon = \frac{1}{2}\int xy^\intercal d\bar\pi_\varepsilon(x,y)$ for some $\bar\pi_\varepsilon$ solving $\mathsf{OT}_{\bar{\boldsymbol{A}}_\varepsilon,0}(\mu_0,\mu_1)$ for $\varepsilon \geq 0$). As noted previously, when ε is small, we can only guarantee that Algorithm 2 will converge to an approximate stationary point, $\boldsymbol{A}_\varepsilon^\star \in \mathcal{D}_M$, satisfying $\|D(\Phi_\varepsilon)_{[\boldsymbol{A}_\varepsilon^\star]}\|_F = \|64\boldsymbol{A}_\varepsilon^\star - 32\int xy^\intercal d\pi_\varepsilon^\star(x,y)\|_F \leq \delta$, for some $\delta > 0$, where $\pi_\varepsilon^\star$ solves $\mathsf{OT}_{\boldsymbol{A}_\varepsilon^\star,\varepsilon}(\mu_0,\mu_1)$. We now show that the limit points of $\boldsymbol{A}_\varepsilon^\star$, as $\varepsilon \downarrow 0$, are approximately stationary for Φ_0. Furthermore, if the matrices along the sequence are globally/locally optimal (in an appropriate sense), we show that global/local optimality is inherited at the limit.

Theorem 16 (Convergence of approximate stationary points)**.** *Let $(\mu_0,\mu_1) \in \mathcal{P}(\mathbb{R}^{d_0})\times\mathcal{P}(\mathbb{R}^{d_1})$ be compactly supported and fix $\delta \geq 0$. For $\varepsilon > 0$, let $\boldsymbol{A}_\varepsilon^\star \in \mathcal{D}_M$ be such that $\|D(\Phi_\varepsilon)_{[\boldsymbol{A}_\varepsilon^\star]}\|_F \leq \delta$ and let $\boldsymbol{A}_0^\star$ be a cluster point of $(\boldsymbol{A}_\varepsilon^\star)_{\varepsilon>0}$ (as $\varepsilon \downarrow 0$). Then,*

1. $\|64\boldsymbol{A}_0^\star - 32\int xy^\intercal d\pi_{\boldsymbol{A}_0^\star}\|_F \leq \delta$, *where* $\pi_{\boldsymbol{A}_0^\star}$ *is a solution of* $\mathsf{OT}_{\boldsymbol{A}_0^\star,0}(\mu_0,\mu_1)$.

2. *if* $\boldsymbol{A}_\varepsilon^\star$ *is a global minimizer of* Φ_ε *for all* $\varepsilon > 0$ *sufficiently small, then* $\boldsymbol{A}_0^\star$ *minimizes* Φ_0.

3. *if, up to a subsequence* $\varepsilon_n \downarrow 0$ *along which* $\boldsymbol{A}_{\varepsilon_n}^\star \to \boldsymbol{A}_0^\star$, $\boldsymbol{A}_{\varepsilon_n}^\star$ *minimizes* Φ_{ε_n} *on a ball of fixed radius* $r > 0$ *centred at* $\boldsymbol{A}_0^\star$, *then* $\boldsymbol{A}_0^\star$ *is a local minimizer of* Φ_0.

Recall that $\mathcal{D}_M$ is compact such that a cluster point $\boldsymbol{A}_0^\star$ always exists in Theorem 16. In light of Proposition 15, these limit points can be thought of as approximate stationary points for Φ_0. This enables approximating local solutions to the unregularized variational GW problem by stationary points obtained using Algorithm 2 for small ε. The proof of Theorem 16 uses the notion of Γ-convergence (see e.g. Braides 2014) to show that the solutions of $\mathsf{OT}_{\boldsymbol{A}_\varepsilon^\star,\varepsilon}(\mu_0,\mu_1)$ converge to a solution of $\mathsf{OT}_{\boldsymbol{A}_0^\star,0}(\mu_0,\mu_1)$ up to a subsequence and that the local/global minimizers of Φ_ε admit local/global minimizers of Φ_0 as cluster points, see Section 5.8 for details.

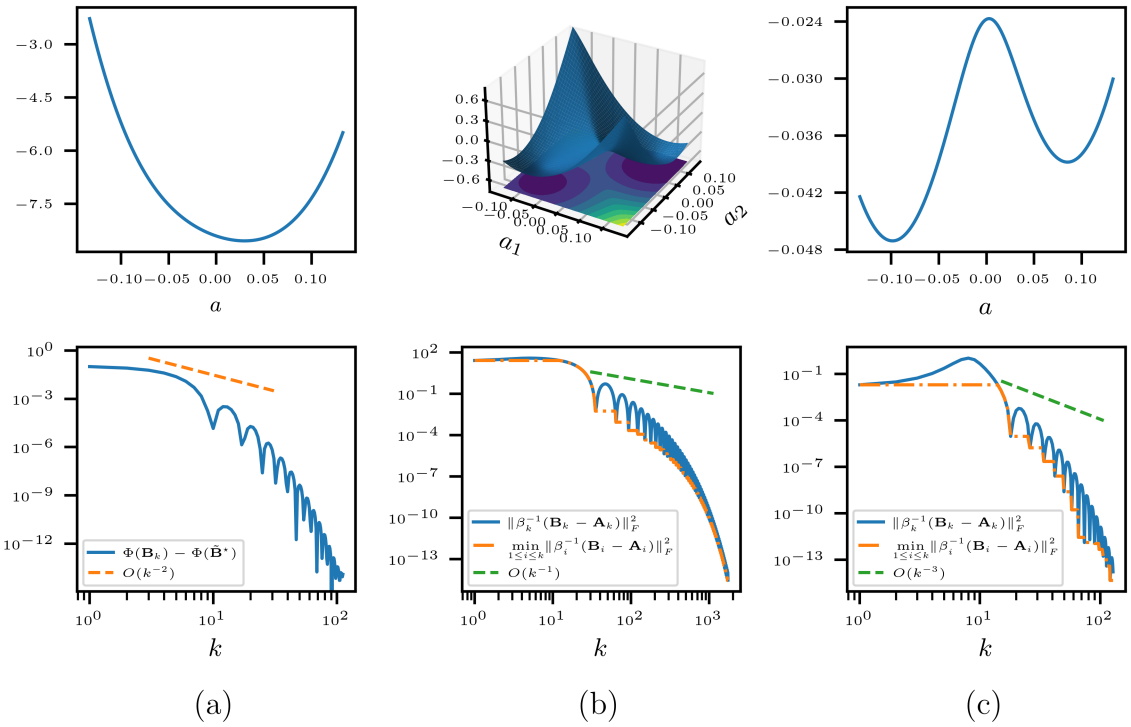
4.5 Numerical Experiments

We conclude this section with some experiments that empirically validate the rates obtained in Theorems 9 and 11 and the computational complexity discussed in Remark 14. All experiments were performed on a desktop computer with 16 GB of RAM and an Intel i5-10600k CPU using the Python programming language. The considered marginal distributions described below, μ_0, μ_1 were randomly generated.

Convergence rates. Figure 1 (a) presents an example of applying Algorithm 1 to a convex Φ, where the marginals are $\mu_0 = 0.4\delta_{-1.4} + 0.6\delta_{1.2}$ and $\mu_1 = 0.4\delta_{-1.01} + 0.6\delta_{1.31}$, with ε chosen large enough to guarantee convexity. The theoretical rate of $O(k^{-2})$ from Theorem 9 on the optimality gap $\Phi(\boldsymbol{B}_k) - \Phi(\boldsymbol{B}^\star)$ is seen to hold.[7] Figure 1 (b) illustrates the progress of Algorithm 2 applied to a non-convex Φ, for $\mu_0 = \frac{1}{3}(\delta_{0.3} + \delta_{-0.8} + \delta_{-0.5})$ and $\mu_1 = \frac{1}{3}\left(\delta_{(0.1,0.6)} + \delta_{(-0.5,0.3)} + \delta_{(0.4,-0.3)}\right)$, with $\varepsilon = 0.07$ which makes Φ non-convex. The $O(k^{-1})$ rate for $\min_{1\leq i\leq k}\left\|\beta_i^{-1}(\boldsymbol{B}_i - \boldsymbol{A}_i)\right\|_F^2$ in the non-convex case from Theorem 11 is well reflected in this example. Figure 1 (c) shows that Algorithm 2 can match the theoretical rate of $O(k^{-3})$ in the convex regime when initialized in a region of local convexity. In this example, the generated marginals are $\mu_0 = \frac{1}{5}(\delta_{-0.1} + \delta_{-0.2} + \delta_{0.2} + \delta_{-0.3} + \delta_{0.3})$ and $\mu_1 = \frac{1}{5}(\delta_{0.2} + \delta_{-0.3} + \delta_{0.3} + \delta_{-0.4} + \delta_{0.4})$ and $\varepsilon = 0.03$. The stopping condition used in all these example is $\|\boldsymbol{G}_k\|_F < 5 \times 10^{-8}$ and the approximate gradient (9) is computed using the standard implementation of Sinkhorn's algorithm from the Python Optimal Transport package (Flamary et al., 2021).

Time complexity. To study the time complexity of Algorithms 1 and 2, we first choose the dimension $d \in \{1, 16, 64, 128\}$ and let $\mu_0, \mu_1 \in \mathcal{P}(\mathbb{R}^d)$ be supported on $N \in \{16, 32, 64, 128, 256, 512, 1024, 2048, 4096, 8192, 16384\}$ samples of a mean-zero normal distribution with standard deviation 0.05 for μ_0 and 0.1 for μ_1. The weights are chosen uniformly at random from $[0, 1)$ and normalized so as to sum to 1. This procedure is repeated to generate a collection of pairs of random distributions $\{(\mu_{0,i}, \mu_{1,i})\}_{i=1}^{50}$. In the sequel, a *single experiment* refers to the

7. The plot shows the approximate gap $\Phi(\boldsymbol{B}_k) - \Phi(\bar{\boldsymbol{B}}^\star)$, where $\bar{\boldsymbol{B}}^\star$ is the iterate attaining the minimal objective value.



(a) (b) (c)

Figure 1: The top row compiles plots of Φ for the different examples described in the text. The bottom row consists of plots tracking the progress of the iterates. In (b) and (c), Algorithm 2 is initialized at $\boldsymbol{C}_0 = (1,1) \times 10^{-5}$ and $\boldsymbol{C}_0 = 1 \times 10^{-5}$, respectively.

process of timing the computation of $\mathsf{S}_\varepsilon(\mu_{0,i}, \mu_{1,i})$ for some fixed d, N and all $i = 1, \ldots, 50$. For practical reasons, we choose to abort an experiment before all 50 EGW distances have been computed if the total runtime for this experiment exceeds 1 hour. The average runtime is then computed among all completed calculations in a single experiment.

The convex case: First, ε is chosen as $1.05 \times 16\sqrt{M_4(\mu_0)M_4(\mu_1)}$ so as to guarantee convexity of Φ for each instance by Theorem 6 and M is set to $\sqrt{M_2(\mu_0)M_2(\mu_1)} + 10^{-5}$. Figure 2 presents the average runtime of both algorithms in this setting with the stopping condition $\|\boldsymbol{G}_k\|_F < 10^{-6}$. We compare the performance of our methods with the two implementations of the $O(N^2)$ mirror descent algorithm provided in Scetbon et al. (2022).[8] The first implementation includes certain algorithmic tweaks when $d^2 \ll N$, whereas the second only requires $d \ll N$ to achieve the quadratic complexity. Our implementation of the mirror descent algorithm is based on the code provided in Scetbon et al. (2022) with some small modifications (e.g., EOT couplings are computed using Sinkhorn's algorithm from the Python Optimal Transport package (Flamary et al., 2021) and some extraneous logging features are removed) the algorithm is run until the generated couplings differ by less than 10^{-6} under the Frobenius norm. We note that the first version of the mirror descent algorithm encounters "out of memory" errors for $N = 16384$.

8. We do not compare with the original implementation of mirror descent Peyré et al. (2016) or the iterative algorithm from Solomon et al. (2016) due to their much slower $O(N^3)$ runtime.

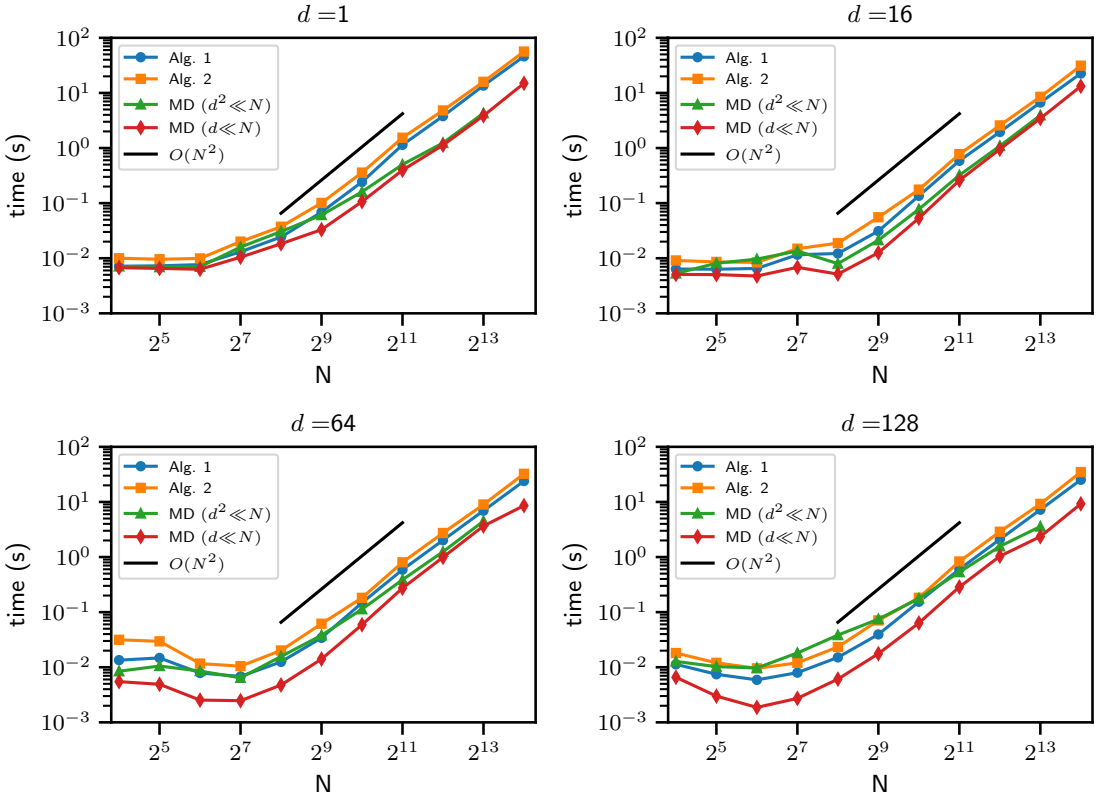


Figure 2: The various plots compile the average runtime of Algorithms 1 and 2, and two versions of the mirror descent algorithm in the convex regime for different combinations of d and N.

The plots show that the four algorithms perform similarly on the considered examples, and empirically validate the $O(N^2)$ computational complexity from Remark 14. To verify that the algorithms all converge to solutions with similar objective values, we evaluate the relative error[9] between all pairs of algorithms for each d, N. The largest relative error we observe is 3.3×10^{-6} for $d = 1$ and, for the other choices of d, is at most 7.9×10^{-13}. We conclude that the values obtained are in good agreement.

The non-convex case: To evaluate the performance of Algorithm 2 when convexity is unknown, we set ε to violate the condition of Theorem 6, but still be large enough so as to avoid numerical errors. If errors in running Algorithm 2 or the mirror descent methods occur, we double ε until all algorithms converge without errors. The initial point $\boldsymbol{C}_0$ for Algorithm 2 is taken as the matrix of all ones scaled by $\min\{M, 1\} \times 10^{-5}$. We consider two ways of choosing the smoothness parameter L, which effectively dictates the rate of

9. Relative error is measured by $\max_{i \in \mathcal{C}(d,N)} \left| \mathsf{S}_\varepsilon^{A1}(\mu_{0,i}, \mu_{1,i}) - \mathsf{S}_\varepsilon^{A2}(\mu_{0,i}, \mu_{1,i}) \right| / \left(\mathsf{S}_\varepsilon^{A1}(\mu_{0,i}, \mu_{1,i}) \wedge \mathsf{S}_\varepsilon^{A2}(\mu_{0,i}, \mu_{1,i}) \right)$, where $\mathsf{S}_\varepsilon^{A1}(\mu_{0,i}, \mu_{1,i})$ and $\mathsf{S}_\varepsilon^{A2}(\mu_{0,i}, \mu_{1,i})$ denote the objective values achieved by the first and second algorithm of the pair, and $\mathcal{C}(d, N)$ is the collection of completed runs from a given experiment.

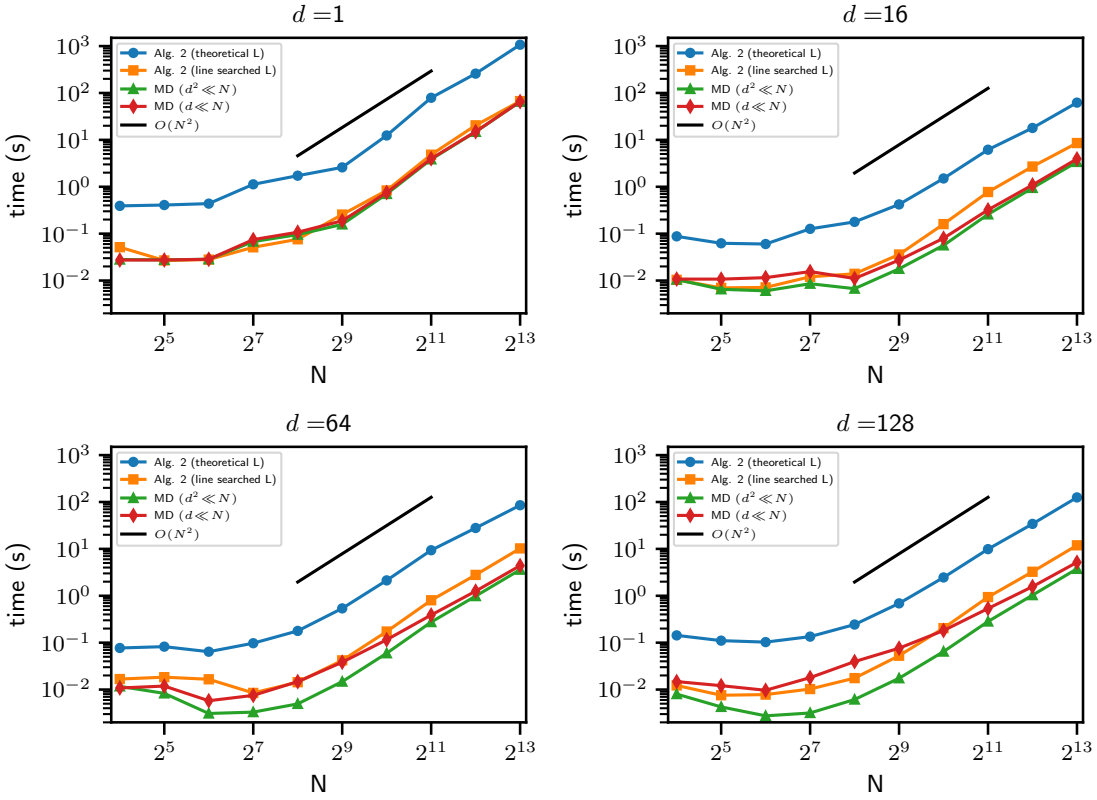


Figure 3: The various plots compile the average runtime of Algorithm 2 with the two methods for choosing L, and two versions of the mirror descent algorithm in the non-convex regime for different combinations of d and N.

convergence. The first is to set L equals to the theoretical upper bound from Theorem 6, i.e., $L = 64 \vee \left(32^2 \varepsilon^{-1} \sqrt{M_4(\mu_0) M_4(\mu_1)} - 64\right)$. As this choice may be too conservative in practice, we also consider setting L via a line search. Namely, we fix a value for L (e.g., the theoretical upper bound or an arbitrary value) and check if the algorithm converges for a given choice of $d, N, \mu_{0,i}, \mu_{1,i}$. If so, we multiply L by 0.99 and repeat this procedure until the algorithm no longer converges. For each d and N, we choose the value of L that attains the fastest convergence, and repeat this procedure for 5 pairs of distributions. For Algorithm 2 with the choice of L that comes from the theoretical bound and the two versions of mirror descent we follow the same methodology as in the convex case, i.e., averaging over 50 pairs and stopping an experiment after 1 hour. The average runtimes of all methods are reported in Figure 3. The restriction to 5 runs in the line search case is only out of convenience and we note that all algorithms yield similar results if we restrict to 5 runs throughout.

The plots again validate the $O(N^2)$ time complexity for all four approaches. However, we see that choosing L in Algorithm 2 according to the theoretical upper bound may indeed be too conservative, as it results in a $10\times$ or larger slowdown compared to the other methods. By setting L via the line search, on the other hand, Algorithm 2 and mirror descent

exhibit similar performance. This suggests that the longer runtime of Algorithm 2 with the theoretical L value can be attributed to this being an overly conservative choice as opposed to a fundamental limitation of this method. Optimization routines that update L at each iteration have been proposed in Tseng (2008); Becker et al. (2011); Nesterov (2013), but require solving an additional EOT problem at each step for our application. As such, these approaches may reduce the number of iterations required for convergence, at the cost of increasing the per iteration complexity.

Real-world data. We next assess the performance of our algorithms on real-world data from the Fashion-MNIST dataset (Xiao et al., 2017). Since the ground truth EGW value is unknown, we test the performance of the algorithm in capturing the invariance of the EGW distance to isometries. As the EGW distance generally does not nullify between isomorphic mm spaces,[10] we consider a centered/debiased version, inspired by debiased EOT (also known as Sinkhorn divergence, Feydy et al., 2018; Genevay et al., 2018). Define the debiased quadratic EGW distance between $(\mu_0, \mu_1) \in \mathcal{P}(\mathbb{R}^{d_0}) \times \mathcal{P}(\mathbb{R}^{d_1})$ as

$$\bar{\mathsf{S}}_\varepsilon(\mu_0, \mu_1) \coloneqq \mathsf{S}_\varepsilon(\mu_0, \mu_1) - \frac{1}{2}\big(\mathsf{S}_\varepsilon(\mu_0, \mu_0) + \mathsf{S}_\varepsilon(\mu_1, \mu_1)\big).$$

The recentering guarantees that $\bar{\mathsf{S}}_\varepsilon$ nullifies whenever $(\mathbb{R}^{d_0}, \|\cdot\|, \mu_0)$ and $(\mathbb{R}^{d_1}, \|\cdot\|, \mu_1)$ are isomorphic, as desired.

Having that, our experiment consists of comparing a fixed image from the Fashion MNIST dataset to rotated versions of itself and other images from the dataset, by computing the debiased EGW distance between them. By doing so, we empirically validate that computation of $\bar{\mathsf{S}}_\varepsilon$ using Algorithm 2 is invariant to isometric transformations on a real-world dataset. Precisely, we pad the 28×28 pixel images from the dataset with zeros on all sides such that the effective image size is 40×40 pixels (this guarantees that no nonzero pixels are lost upon rotation) and treat these padded images as probability distributions on a 40×40 grid of points in $\mathbb{R}^2$ with weights proportional to the pixel intensity value. We then compare these distributions using $\bar{\mathsf{S}}_\varepsilon$ upon removing the points with zero mass from each distribution, the values thus obtained are included in Figure 4.

We note that rotating an image by an angle which is not a multiple of 90° does not correspond to an isometric action on $\mathbb{R}^2$, as it requires interpolating the pixel back onto the 40×40 grid of points. Nevertheless, we see from Figure 4, that the images subject to random rotations and the unrotated images achieve similar values of $\bar{\mathsf{S}}_\varepsilon$ relative to the fixed reference image. The discrepancy between these two values can be thought of as a quantification of the distortion to the image structure caused by the interpolation procedure. When the rotation is a multiple of 90°, we see that the values obtained are identical, as no interpolation is performed on the image.

Figure 4 marks in red values corresponding to images that are closest to the reference image in the debiased EGW distance. These images have a notable structural similarity to the reference in their overall shape and/or features (e.g., pleats on the dress). We also observe that quite naturally the images with largest discrepancy are those of shoes, which have a

10. Indeed, $\mathsf{D}_{\mathsf{KL}}(\pi \| \mu_0 \otimes \mu_1) \geq 0$ with equality if and only if $\pi = \mu_0 \otimes \mu_1$ whereas the integral of the distance distortion cost vanishes if and only if the coupling is induced by some isometry $T : \mathbb{R}^{d_0} \to \mathbb{R}^{d_1}$. By Corollary 4, an optimal coupling $\pi^\star$ for $\mathsf{S}_\varepsilon(\mu_0, \mu_1)$ is equivalent to $\mu_0 \otimes \mu_1$ (in the sense that $\pi^\star \ll \mu_0 \otimes \mu_1$ and $\mu_0 \otimes \mu_1 \ll \pi^\star$), so $\pi^\star$ cannot be induced by an isometry unless μ_0, μ_1 are supported on one point.

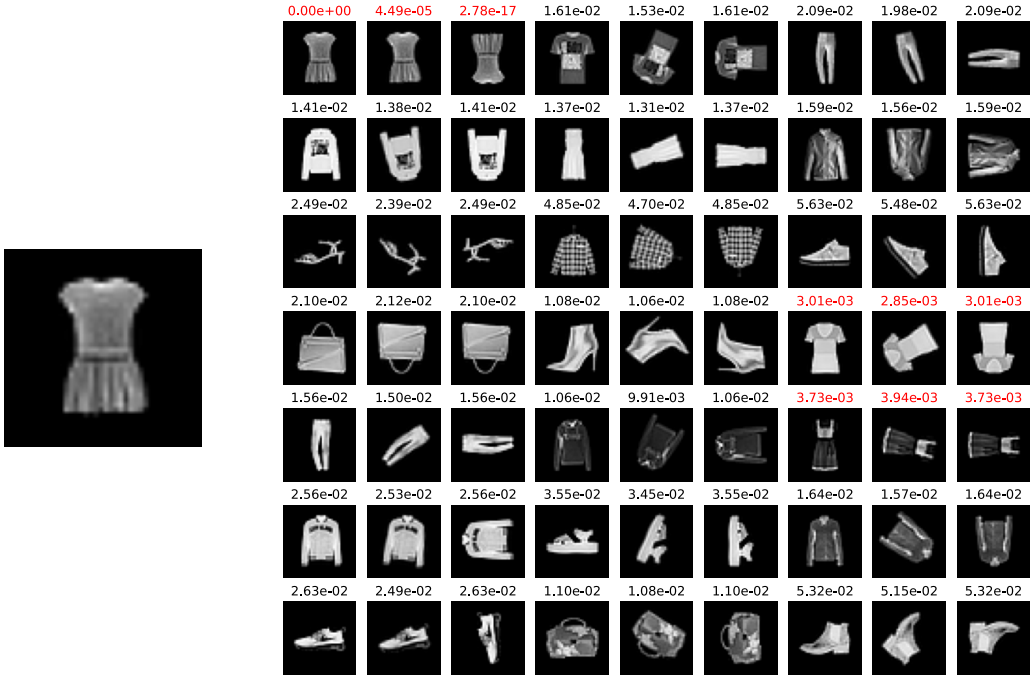


Figure 4: We compare the image on the left to the images on the right using debiased EGW with $\varepsilon = 0.1$ as a figure of merit. The corresponding value of $\bar{\mathsf{S}}_\varepsilon$ is included on top of the corresponding images. The images are presented in groups of three, where the leftmost image in the group is the original image, the middle image is obtained via a random rotation of the original image, and the rightmost image is a rotation by a multiple of 90°. Distance values smaller than 5×10^{-3} are in red.

distinct structure, and the plaid shirt which has a different pattern. Interestingly, between these two extremes, there are images which have comparable values, but are structurally dissimilar (for instance the images in the center column except for the plaid shirt and the sandal). This behaviour demonstrates that the debiased EGW distance does not simply compare the shapes of the images, but rather takes into account the intricate interplay between the intensity values of the images.

5 Proofs

5.1 Proof of Proposition 2

We first fix some notation. Let $S_i = \mathrm{spt}(\mu_i)$ for $i = 0, 1$ and define the Banach spaces

$$\mathfrak{E} = \left\{ (f_0, f_1) \in \mathcal{C}(S_0) \times \mathcal{C}(S_1) : \int f_0 d\mu_0 = 0 \right\},$$
$$\mathfrak{F} = \left\{ (f_0, f_1) \in \mathcal{C}(S_0) \times \mathcal{C}(S_1) : \int f_0 d\mu_0 = \int f_1 d\mu_1 \right\}.$$

Consider the map $\Upsilon : \mathbb{R}^{d_0 \times d_1} \times \mathfrak{E} \to \mathcal{C}(S_0) \times \mathcal{C}(S_1)$ given by

$$\Upsilon : (\boldsymbol{A}, \varphi_0, \varphi_1) \mapsto \left(\int e^{\frac{\varphi_0(\cdot)+\varphi_1(y)-c_{\boldsymbol{A}}(\cdot,y)}{\varepsilon}} d\mu_1(y) - 1, \int e^{\frac{\varphi_0(x)+\varphi_1(\cdot)-c_{\boldsymbol{A}}(x,\cdot)}{\varepsilon}} d\mu_0(x) - 1 \right).$$

For fixed $\boldsymbol{A} \in \mathbb{R}^{d_0 \times d_1}$, the solution to the equation $\Upsilon(\boldsymbol{A}, \cdot, \cdot) = 0$ is the unique pair of EOT potentials $(\varphi_0^{\boldsymbol{A}}, \varphi_1^{\boldsymbol{A}})$ for μ_0, μ_1 with the cost $c_{\boldsymbol{A}}$ satisfying the normalization from $\mathfrak{E}$. Observe that, by compactness of S_0 and S_1, the potentials are bounded.

The following lemmas verify the conditions to apply the implicit mapping theorem to Υ in order to obtain the Fréchet derivative of the map $\boldsymbol{A} \in \mathbb{R}^{d_0 \times d_1} \mapsto (\varphi_0^{\boldsymbol{A}}, \varphi_1^{\boldsymbol{A}})$. Given that $\mathsf{OT}_{\boldsymbol{A},\varepsilon}(\mu_0, \mu_1) = \int \varphi_0^{\boldsymbol{A}} d\mu_0 + \int \varphi_1^{\boldsymbol{A}} d\mu_1$, the derivative of the map $\boldsymbol{A} \mapsto \mathsf{OT}_{\boldsymbol{A},\varepsilon}(\mu_0, \mu_1)$ and that of Φ itself will readily follow.

Lemma 17. *The map Υ is smooth with first derivative at $(\boldsymbol{A}, \varphi_0, \varphi_1) \in \mathbb{R}^{d_0 \times d_1} \times \mathfrak{E}$ given by,*

$$D\Upsilon_{[\boldsymbol{A},\varphi_0,\varphi_1]}(\boldsymbol{B}, h_0, h_1) = \varepsilon^{-1} \Bigg(\int (h_0(\cdot) + h_1(y) + 32(\cdot)^\intercal \boldsymbol{B} y) e^{\frac{\varphi_0(\cdot)+\varphi_1(y)-c_{\boldsymbol{A}}(\cdot,y)}{\varepsilon}} d\mu_1(y), \\ \int (h_0(x) + h_1(\cdot) + 32x^\intercal \boldsymbol{B}(\cdot)) e^{\frac{\varphi_0(x)+\varphi_1(\cdot)-c_{\boldsymbol{A}}(x,\cdot)}{\varepsilon}} d\mu_0(x) \Bigg),$$

where $(\boldsymbol{B}, h_0, h_1) \in \mathbb{R}^{d_0 \times d_1} \times \mathfrak{E}$.

The proof of this result is straightforward, but included in Appendix D.1 for completeness. Now, define $\xi_{\boldsymbol{A}} \coloneqq \varepsilon D\Upsilon_{[\boldsymbol{A},\varphi_0^{\boldsymbol{A}},\varphi_1^{\boldsymbol{A}}]}(0, \cdot, \cdot)$ and let $\pi_{\boldsymbol{A}}$ be the EOT coupling for $\mathsf{OT}_{\boldsymbol{A},\varepsilon}(\mu_0, \mu_1)$. Note that for any $(h_0, h_1) \in \mathfrak{E}$, we have $\xi_{\boldsymbol{A}}(h_0, h_1) \in \mathfrak{F}$, which follows by recalling that $\frac{d\pi_{\boldsymbol{A}}}{d\mu_0 \otimes \mu_1}(x, y) = e^{\frac{\varphi_0^{\boldsymbol{A}}(x)+\varphi_1^{\boldsymbol{A}}(y)-c_{\boldsymbol{A}}(x,y)}{\varepsilon}}$ and observing

$$\int \left(\xi_{\boldsymbol{A}}(h_0, h_1)\right)_1 d\mu_0 = \int h_0 d\mu_0 + \int h_1 d\pi_{\boldsymbol{A}} = \int h_0 d\mu_0 + \int h_1 d\mu_1$$
$$\int \left(\xi_{\boldsymbol{A}}(h_0, h_1)\right)_2 d\mu_1 = \int h_0 d\pi_{\boldsymbol{A}} + \int h_1 d\mu_1 = \int h_0 d\mu_0 + \int h_1 d\mu_1.$$

We next prove that $\xi_{\boldsymbol{A}}$ is an isomorphism between $\mathfrak{E}$ and $\mathfrak{F}$ by following the proof of Proposition 3.1 in Carlier and Laborde (2020).

Lemma 18. *The map $\xi_{\boldsymbol{A}}$ is an isomorphism between $\mathfrak{E}$ and $\mathfrak{F}$.*

Proof Observe that $\xi_{\boldsymbol{A}}$ extends naturally to a map on $L^2(\mu_0) \times L^2(\mu_1)$ and admits the representation

$$\xi_{\boldsymbol{A}} : (f_0, f_1) \in L^2(\mu_0) \times L^2(\mu_1) \mapsto (f_0, f_1) + \mathcal{L}(f_0, f_1) \in L^2(\mu_0) \times L^2(\mu_1),$$

where

$$\mathcal{L}(f_0, f_1) = \left(\int f_1(y) e^{\frac{\varphi_0^{\boldsymbol{A}}(\cdot)+\varphi_1^{\boldsymbol{A}}(y)-c_{\boldsymbol{A}}(\cdot,y)}{\varepsilon}} d\mu_1(y), \int f_0(x) e^{\frac{\varphi_0^{\boldsymbol{A}}(x)+\varphi_1^{\boldsymbol{A}}(\cdot)-c_{\boldsymbol{A}}(x,\cdot)}{\varepsilon}} d\mu_0(x) \right).$$

Lemma 33 in Appendix D.2 demonstrates that $\mathcal{L}$ is a compact linear self-map of $L^2(\mu_0) \times L^2(\mu_1)$.

We first show that $\xi_{\boldsymbol{A}}$ is injective on $E \coloneqq \{(f_0, f_1) \in L^2(\mu_0) \times L^2(\mu_1) : \int f_0 d\mu_0 = 0\}$. Suppose that $(\bar{f}_0, \bar{f}_1)$ satisfies $\xi_{\boldsymbol{A}}(\bar{f}_0, \bar{f}_1) = 0$. Multiplying $(\xi_{\boldsymbol{A}}(\bar{f}_0, \bar{f}_1))_1$ by $\bar{f}_0$ and $(\xi_{\boldsymbol{A}}(\bar{f}_0, \bar{f}_1))_2$ by $\bar{f}_1$, we have

$$\int \left(\bar{f}_0^2(\cdot) + \bar{f}_0(\cdot)\bar{f}_1(y)\right) e^{\frac{\varphi_0^{\boldsymbol{A}}(\cdot)+\varphi_1^{\boldsymbol{A}}(y)-c_{\boldsymbol{A}}(\cdot,y)}{\varepsilon}} d\mu_1(y) = 0,$$
$$\int \left(\bar{f}_0(x)f_1(\cdot) + \bar{f}_1^2(\cdot)\right) e^{\frac{\varphi_0^{\boldsymbol{A}}(x)+\varphi_1^{\boldsymbol{A}}(\cdot)-c_{\boldsymbol{A}}(x,\cdot)}{\varepsilon}} d\mu_0(x) = 0,$$

and summing these equations gives $\int (\bar{f}_0 + \bar{f}_1)^2 d\pi_{\boldsymbol{A}} = 0$. As $\pi_{\boldsymbol{A}}$ is equivalent to $\mu_0 \otimes \mu_1$, we have $\bar{f}_0 + \bar{f}_1 = 0$ $\mu_0 \otimes \mu_1$-a.e., which further implies that $(\bar{f}_0, \bar{f}_1) = (a, -a)$ $\mu_0 \otimes \mu_1$-a.e. for some $a \in \mathbb{R}$. Consequently, $\ker(\xi_{\boldsymbol{A}})$ is 1-dimensional and $\xi_{\boldsymbol{A}}$ is injective on E.

Next, we show that $\xi_{\boldsymbol{A}}$ is onto $F \coloneqq \{(f_0, f_1) \in L^2(\mu_0) \times L^2(\mu_1) : \int f_0 d\mu_0 = \int f_1 d\mu_1\}$. As in the lead-up to this lemma, $\xi_{\boldsymbol{A}}(E) \subset F$. By the Fredholm alternative (cf. Theorem 6.6 in Brézis (2011)), $(\mathrm{Id} + \mathcal{L})(L^2(\mu_0) \times L^2(\mu_1))$ has codimension 1 and, as F has codimension 1, we must have $\xi_{\boldsymbol{A}}(E) = F$.

As such, for any $(g_0, g_1) \in \mathfrak{F} \subset F$, there exists $(h_0, h_1) \in E$ for which

$$\xi_{\boldsymbol{A}}(h_0, h_1) = (h_0, h_1) + \mathcal{L}(h_0, h_1) = (g_0, g_1).$$

As $\mathcal{L}(h_0, h_1) \in \mathcal{C}(S_0) \times \mathcal{C}(S_1)$, $(h_0, h_1) = (g_0, g_1) - \mathcal{L}(h_0, h_1) \in \mathcal{C}(S_0) \times \mathcal{C}(S_1)$ with $\int h_0 d\mu_0 = 0$, and thus $(h_0, h_1) \in \mathfrak{E}$. This implies that $\xi_{\boldsymbol{A}}(\mathfrak{E}) \supset \mathfrak{F}$ and from before we have $\xi_{\boldsymbol{A}}(\mathfrak{E}) \subset \mathfrak{F}$, yielding $\xi_{\boldsymbol{A}}(\mathfrak{E}) = \mathfrak{F}$. We have shown that $\xi_A : \mathfrak{E} \to \mathfrak{F}$ is a continuous linear bijection and hence an isomorphism by the open mapping theorem (cf. Corollary 2.7 in Brézis 2011). ■

We now apply the implicit mapping theorem to obtain the Fréchet derivative of $(\varphi_0^{(\cdot)}, \varphi_1^{(\cdot)})$.

Lemma 19. *The map $\boldsymbol{A} \in \mathbb{R}^{d_0 \times d_1} \mapsto (\varphi_0^{\boldsymbol{A}}, \varphi_1^{\boldsymbol{A}}) \in \mathfrak{E}$ is smooth with Fréchet derivative*

$$D\left(\varphi_0^{(\cdot)}, \varphi_1^{(\cdot)}\right)_{[\boldsymbol{A}]}(\boldsymbol{B}) = -\left(h_0^{\boldsymbol{A},\boldsymbol{B}}, h_1^{\boldsymbol{A},\boldsymbol{B}}\right),$$

where $\left(h_0^{\boldsymbol{A},\boldsymbol{B}}, h_1^{\boldsymbol{A},\boldsymbol{B}}\right) \in \mathfrak{E}$ satisfies

$$\begin{aligned}
&\int \left(h_0^{\boldsymbol{A},\boldsymbol{B}}(x) + h_1^{\boldsymbol{A},\boldsymbol{B}}(y) - 32x^\intercal \boldsymbol{B} y\right) e^{\frac{\varphi_0^{\boldsymbol{A}}(x)+\varphi_1^{\boldsymbol{A}}(y)-c_{\boldsymbol{A}}(x,y)}{\varepsilon}} d\mu_1(y) = 0, \quad \forall\, x \in \mathrm{spt}(\mu_0), \\
&\int \left(h_0^{\boldsymbol{A},\boldsymbol{B}}(x) + h_1^{\boldsymbol{A},\boldsymbol{B}}(y) - 32x^\intercal \boldsymbol{B} y\right) e^{\frac{\varphi_0^{\boldsymbol{A}}(x)+\varphi_1^{\boldsymbol{A}}(y)-c_{\boldsymbol{A}}(x,y)}{\varepsilon}} d\mu_0(x) = 0, \quad \forall\, y \in \mathrm{spt}(\mu_1),
\end{aligned} \tag{12}$$

with $(\varphi_0^{\boldsymbol{A}}, \varphi_1^{\boldsymbol{A}})$ being any pair of EOT potentials for (μ_0, μ_1) with the cost $c_{\boldsymbol{A}}$.

Proof Fix $\boldsymbol{A} \in \mathbb{R}^{d_0 \times d_1}$ with corresponding EOT potentials $\left(\varphi_0^{\boldsymbol{A}}, \varphi_1^{\boldsymbol{A}}\right)$. For notational convenience, define the shorthands $D_1\Upsilon_{\boldsymbol{A}} = D\Upsilon_{[\boldsymbol{A},\varphi_0^{\boldsymbol{A}},\varphi_1^{\boldsymbol{A}}]}(\cdot, 0, 0)$ and $D_2\Upsilon_{\boldsymbol{A}} = D\Upsilon_{[\boldsymbol{A},\varphi_0^{\boldsymbol{A}},\varphi_1^{\boldsymbol{A}}]}(0, \cdot, \cdot)$ (cf. Lemma 17). By Lemma 18, $D_2\Upsilon_{\boldsymbol{A}}$ is an isomorphism and we may invoke the implicit mapping theorem (cf. Theorem 5.14 in Bonnans and Shapiro 2013). This implies that there

exists an open neighborhood $U \subset \mathbb{R}^{d_0 \times d_1}$ of $\boldsymbol{A}$ and a smooth map $g : U \to \mathfrak{E}$ for which $\Upsilon\big(\boldsymbol{B}, g(\boldsymbol{B})\big) = 0$ for every $\boldsymbol{B} \in U$ and

$$Dg_{[\boldsymbol{A}]}(\boldsymbol{B}) = -\left(D_2 \Upsilon_{\boldsymbol{A}}\right)^{-1} \left(D_1 \Upsilon_{\boldsymbol{A}}(\boldsymbol{B})\right),$$

i.e., $-Dg_{[\boldsymbol{A}]}(\boldsymbol{B})$ solves (12). By construction, $g(\boldsymbol{B}) = (\varphi_0^{\boldsymbol{B}}, \varphi_1^{\boldsymbol{B}})$ and by repeating this process at any $\boldsymbol{A} \in \mathbb{R}^{d_0 \times d_1}$, differentiability of the potentials is extended to the entire space $\mathbb{R}^{d_0 \times d_1}$. ■

Given the dual form of the EOT cost, Lemma 19 suffices to prove Proposition 2.

Proof of Proposition 2 As $\mathsf{OT}_{\boldsymbol{A},\varepsilon}(\mu_0, \mu_1) = \int \varphi_0^{\boldsymbol{A}} d\mu_0 + \int \varphi_1^{\boldsymbol{A}} d\mu_1$, Lemma 19 implies that $\mathsf{OT}_{(\cdot),\varepsilon}(\mu_0, \mu_1)$ is smooth with first derivative at $\boldsymbol{A} \in \mathbb{R}^{d_0 \times d_1}$ given by

$$D\big(\mathsf{OT}_{(\cdot),\varepsilon}(\mu_0, \mu_1)\big)_{[\boldsymbol{A}]}(\boldsymbol{B}) = -\int h_0^{\boldsymbol{A},\boldsymbol{B}} d\mu_0 - \int h_1^{\boldsymbol{A},\boldsymbol{B}} d\mu_1,$$

where $\boldsymbol{B} \in \mathbb{R}^{d_0 \times d_1}$. Integrating the first equation in (12) w.r.t. μ_0 while using $\frac{d\pi_{\boldsymbol{A}}}{\mu_0 \otimes \mu_1}(x, y) = e^{\frac{\varphi_0^{\boldsymbol{A}}(x) + \varphi_1^{\boldsymbol{A}}(y) - c_{\boldsymbol{A}}(x,y)}{\varepsilon}}$, yields

$$\int \Big(h_0^{\boldsymbol{A},\boldsymbol{B}}(x) + h_1^{\boldsymbol{A},\boldsymbol{B}}(y)\Big)\, d\pi_{\boldsymbol{A}}(x, y) = \int h_0^{\boldsymbol{A},\boldsymbol{B}} d\mu_0 + \int h_1^{\boldsymbol{A},\boldsymbol{B}} d\mu_1 = 32 \int x^\intercal \boldsymbol{B} y \, d\pi_{\boldsymbol{A}}(x, y), \tag{13}$$

whence

$$D\big(\mathsf{OT}_{(\cdot),\varepsilon}(\mu_0, \mu_1)\big)_{[\boldsymbol{A}]}(\boldsymbol{B}) = -32 \int x^\intercal \boldsymbol{B} y \, d\pi_{\boldsymbol{A}}(x, y).$$

As $\|\boldsymbol{A}\|_F^2 = \mathrm{tr}(\boldsymbol{A}^\intercal \boldsymbol{A})$, we have $D\big(32\|\cdot\|_F^2\big)_{[\boldsymbol{A}]}(\boldsymbol{B}) = 64\mathrm{tr}(\boldsymbol{A}^\intercal \boldsymbol{B})$, which together with the display above yields

$$D\Phi_{[\boldsymbol{A}]}(\boldsymbol{B}) = 64\,\mathrm{tr}(\boldsymbol{A}^\intercal \boldsymbol{B}) - 32 \int x^\intercal \boldsymbol{B} y \, d\pi_{\boldsymbol{A}}(x, y),$$

as desired.

For the second-order derivative, recall from (4) that $\frac{d\pi_{\boldsymbol{A}}}{d\mu_0 \otimes \mu_1}(x, y) = e^{\frac{\varphi_0^{\boldsymbol{A}}(x) + \varphi_1^{\boldsymbol{A}}(y) - c_{\boldsymbol{A}}(x,y)}{\varepsilon}}$. As in the proof of Lemma 17, as the map

$$\boldsymbol{A} \in \mathbb{R}^{d_0 \times d_1} \mapsto \big((x, y) \in S_0 \times S_1 \mapsto \varphi_0^{\boldsymbol{A}}(x) + \varphi_1^{\boldsymbol{A}}(y) - c_{\boldsymbol{A}}(x, y)\big) \in \mathcal{C}(S_0 \times S_1)$$

is differentiable at $\boldsymbol{A} \in \mathbb{R}^{d_0 \times d_1}$ with derivative

$$\boldsymbol{C} \in \mathbb{R}^{d_0 \times d_1} \mapsto \Big((x, y) \in S_0 \times S_1 \mapsto -\Big(h_0^{\boldsymbol{A},\boldsymbol{C}}(x) + h_1^{\boldsymbol{A},\boldsymbol{C}}(y) - 32 x^\intercal \boldsymbol{C} y\Big)\Big) \in \mathcal{C}(S_0 \times S_1),$$

the expansion

$$\frac{d\pi_{\boldsymbol{A}+\boldsymbol{C}}}{d\mu_0 \otimes \mu_1}(x, y) - \frac{d\pi_{\boldsymbol{A}}}{d\mu_0 \otimes \mu_1}(x, y) = -\varepsilon^{-1} z_{\boldsymbol{A},\boldsymbol{C}}(x, y) \frac{d\pi_{\boldsymbol{A}}}{d\mu_0 \otimes \mu_1}(x, y) + R_{\boldsymbol{C}}(x, y),$$

holds uniformly over $(x,y) \in S_0 \times S_1$, where $R_{\boldsymbol{C}}(x,y) = o(\boldsymbol{C})$ as $\|\boldsymbol{C}\|_F \to 0$ and $z_{\boldsymbol{A},\boldsymbol{C}}(x,y) = h_0^{\boldsymbol{A},\boldsymbol{C}}(x) + h_1^{\boldsymbol{A},\boldsymbol{C}}(y) - 32x^\intercal \boldsymbol{C} y$. Thus,

$$\begin{aligned}
&\sup_{\|\boldsymbol{B}\|_F=1} \frac{\left|\int x^\intercal \boldsymbol{B} y \, d\pi_{\boldsymbol{A}+\boldsymbol{C}}(x,y) - \int x^\intercal \boldsymbol{B} y \, d\pi_{\boldsymbol{A}}(x,y) - \varepsilon^{-1} \int x^\intercal \boldsymbol{B} y z_{\boldsymbol{A},\boldsymbol{C}}(x,y) d\pi_{\boldsymbol{A}}(x,y)\right|}{\|\boldsymbol{C}\|_F} \\
&= \sup_{\|\boldsymbol{B}\|_F=1} \left| \int x^\intercal \boldsymbol{B} y \|\boldsymbol{C}\|_F^{-1} R_{\boldsymbol{C}}(x,y) d\mu_0 \otimes \mu_1(x,y) \right| \\
&\leq \sup_{(x,y)\in S_1 \times S_2} \|x\|\|y\| \int \|\boldsymbol{C}\|_F^{-1} \left| R_{\boldsymbol{C}}(x,y) \right| d\mu_0 \otimes \mu_1(x,y).
\end{aligned}$$

As $R_{\boldsymbol{C}}(x,y) = o(\boldsymbol{C})$, this final term converges to 0 as $\|\boldsymbol{C}\|_F \to 0$, so

$$D^2\big(\mathsf{OT}_{(\cdot),\varepsilon}(\mu_0,\mu_1)\big)_{[\boldsymbol{A}]}(\boldsymbol{B},\boldsymbol{C}) = 32\varepsilon^{-1} \int x^\intercal \boldsymbol{B} y \left(h_0^{\boldsymbol{A},\boldsymbol{C}}(x) + h_1^{\boldsymbol{A},\boldsymbol{C}}(y) - 32x^\intercal \boldsymbol{C} y \right) d\pi_{\boldsymbol{A}}(x,y).$$

As $D\big(32\|\cdot\|_F^2\big)_{[\boldsymbol{A}]}(\boldsymbol{B}) = 64\mathrm{tr}(\boldsymbol{A}^\intercal \boldsymbol{B})$, $D^2\big(32\|\cdot\|_F^2\big)_{[\boldsymbol{A}]}(\boldsymbol{B},\boldsymbol{C}) = 64\mathrm{tr}(\boldsymbol{C}^\intercal \boldsymbol{B})$. Altogether,

$$D^2\Phi_{[\boldsymbol{A}]}(\boldsymbol{B},\boldsymbol{C}) = 64\,\mathrm{tr}(\boldsymbol{B}^\intercal \boldsymbol{C}) + 32\varepsilon^{-1} \int x^\intercal \boldsymbol{B} y \left(h_0^{\boldsymbol{A},\boldsymbol{C}}(x) + h_1^{\boldsymbol{A},\boldsymbol{C}}(y) - 32x^\intercal \boldsymbol{C} y \right) d\pi_{\boldsymbol{A}}(x,y).$$

Coercivity of Φ is due to nonnegativity of the KL divergence along with the Cauchy-Schwarz inequality,

$$\begin{aligned}
\mathsf{OT}_{\boldsymbol{A},\varepsilon}(\mu_0,\mu_1) &= \inf_{\pi\in\Pi(\mu_0,\mu_1)} \left\{ \int -4\|x\|^2\|y\|^2 - 32x^\intercal \boldsymbol{A} y \, d\pi(x,y) + \varepsilon \mathsf{D}_{\mathsf{KL}}(\pi \| \mu_0 \otimes \mu_1) \right\}, \\
&\geq \inf_{\pi\in\Pi(\mu_0,\mu_1)} \left\{ \int -4\|x\|^2\|y\|^2 - 32\|\boldsymbol{A}\|_F \|x\|\|y\| d\pi(x,y) \right\} \\
&\geq -4\sqrt{M_4(\mu_0)M_4(\mu_1)} - 32\|\boldsymbol{A}\|_F \sqrt{M_2(\mu_0)M_2(\mu_1)},
\end{aligned}$$

such that $\Phi(\boldsymbol{A}) = 32\|\boldsymbol{A}\|_F^2 + \mathsf{OT}_{\boldsymbol{A},\varepsilon}(\mu_0,\mu_1) \to +\infty$ as $\|\boldsymbol{A}\|_F \to \infty$. ■

5.2 Proof of Corollary 4

Item (i). The expression for the stationary points follows immediately from Proposition 2. To see that all stationary points are elements of $\mathcal{D}_{M_{\mu_0,\mu_1}}$, observe that if $\boldsymbol{A}$ is a stationary point, then

$$\|\boldsymbol{A}\|_F = \frac{1}{2} \left\| \int xy^\intercal \, d\pi_{\boldsymbol{A}}(x,y) \right\| \leq \frac{1}{2} \int \|x\|\|y\| d\pi_{\boldsymbol{A}}(x,y) \leq \frac{1}{2} \sqrt{M_2(\mu_0)M_2(\mu_1)},$$

where the first inequality is due to Jensen's inequality, and the second is due to the Cauchy-Schwarz inequality.

Item (ii). As discussed in Section 2.2, if $\pi_\star$ is optimal for S_ε then $\frac{1}{2}\int xy^\intercal \, d\pi_\star(x,y)$ minimizes Φ. On the other hand, if $\boldsymbol{A}$ minimizes Φ, then we have $\boldsymbol{A} = \frac{1}{2}\int xy^\intercal \, d\pi_{\boldsymbol{A}}$ and

hence

$$\begin{aligned}\mathsf{S}_\varepsilon^2(\mu_0,\mu_1) &= 8\left\|\int xy^\intercal\, d\pi_{\boldsymbol{A}}(x,y)\right\|_F^2 - 4\int \|x\|^2\|y\|^2 d\pi_{\boldsymbol{A}}(x,y)\\ &\quad - 32\left\langle \frac{1}{2}\int xy^\intercal\, d\pi_{\boldsymbol{A}}, \int xy^\intercal\, d\pi_{\boldsymbol{A}}\right\rangle_F + \varepsilon \mathsf{D}_{\mathsf{KL}}(\pi_{\boldsymbol{A}}||\mu_0\otimes\mu_1)\\ &= -4\int \|x\|^2\|y\|^2 d\pi_{\boldsymbol{A}(x,y)} - 8\left\|\int xy^\intercal\, d\pi_{\boldsymbol{A}}(x,y)\right\|_F^2 + \varepsilon \mathsf{D}_{\mathsf{KL}}(\pi_{\boldsymbol{A}}||\mu_0\otimes\mu_1).\end{aligned}$$

By (6),

$$\begin{aligned}\mathsf{S}_\varepsilon(\mu_0,\mu_1) &= \mathsf{S}_\varepsilon(\mu_0,\mu_1) + \mathsf{S}_\varepsilon^2(\mu_0,\mu_1)\\ &= \int \left|\|x-x'\|^2 - \|y-y'\|^2\right|^2 + 2\|x-x'\|^2\|y-y'\|^2 d\pi_{\boldsymbol{A}}\otimes\pi_{\boldsymbol{A}}(x,y,x',y')\\ &\quad - 4\int \|x\|^2\|y\|^2 d\mu_0\otimes\mu_1(x,y) - 4\int \|x\|^2\|y\|^2 d\pi_{\boldsymbol{A}}(x,y)\\ &\quad - 8\left\|\int xy^\intercal\, d\pi_{\boldsymbol{A}}(x,y)\right\|_F^2 + \varepsilon \mathsf{D}_{\mathsf{KL}}(\pi_{\boldsymbol{A}}||\mu_0\otimes\mu_1).\end{aligned} \tag{14}$$

As $\|x-x'\|^2\|y-y'\|^2 = \left(\|x\|^2 - 2x^\intercal x' + \|x'\|^2\right)\left(\|y\|^2 - 2y^\intercal y' + \|y'\|^2\right)$, we have

$$\begin{aligned}&\int \|x-x'\|^2\|y-y'\|^2 d\pi_{\boldsymbol{A}}\otimes\pi_{\boldsymbol{A}}(x,y,x',y')\\ &= 2\int \|x\|^2\|y\|^2 d\mu_0\otimes\mu_1(x,y) + 2\int \|x\|^2\|y\|^2 d\pi_{\boldsymbol{A}}(x,y)\\ &\quad + 4\int x^\intercal x' y^\intercal y'\, d\pi_{\boldsymbol{A}}\otimes\pi_{\boldsymbol{A}}(x,y,x',y'),\end{aligned}$$

which, together with (14) yields

$$\mathsf{S}_\varepsilon(\mu_0,\mu_1) = \int \left|\|x-x'\|^2 - \|y-y'\|^2\right|^2 d\pi_{\boldsymbol{A}}\otimes\pi_{\boldsymbol{A}}(x,y,x',y') + \varepsilon \mathsf{D}_{\mathsf{KL}}(\pi_{\boldsymbol{A}}||\mu_0\otimes\mu_1),$$

proving optimality of $\pi_{\boldsymbol{A}}$.

Item (iii). Suppose S_ε admits a unique optimal coupling. If two matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ minimize Φ, then $\pi_{\boldsymbol{A}} = \pi_{\boldsymbol{B}}$ by uniqueness, so $\boldsymbol{A} = \frac{1}{2}\int xy^\intercal\, d\pi_{\boldsymbol{A}}(x,y) = \frac{1}{2}\int xy^\intercal\, d\pi_{\boldsymbol{B}}(x,y) = \boldsymbol{B}$. Conversely, suppose Φ admits a unique minimizer $\boldsymbol{A}^\star$. If π is optimal for S_ε, then π solves the EOT problem $\mathsf{OT}_{\boldsymbol{A}^\star,\varepsilon}(\mu_0,\mu_1)$, so $\pi = \pi_{\boldsymbol{A}^\star}$. □

5.3 Proof of Corollary 5

We first prove Item (i). The minimal eigenvalue of $D^2\Phi_{[\boldsymbol{A}]}$ is given in variational form as

$$\begin{aligned}&\inf_{\|\boldsymbol{C}\|_F=1} D^2\Phi_{[\boldsymbol{A}]}(\boldsymbol{C},\boldsymbol{C})\\ &= \inf_{\|\boldsymbol{C}\|_F=1}\left\{64\|\boldsymbol{C}\|_F^2 + 32\varepsilon^{-1}\int x^\intercal \boldsymbol{C} y\left(h_0^{\boldsymbol{A},\boldsymbol{C}}(x) + h_1^{\boldsymbol{A},\boldsymbol{C}}(y) - 32x^\intercal \boldsymbol{C} y\right) d\pi_{\boldsymbol{A}}(x,y)\right\}\\ &\geq 64 + 32\varepsilon^{-1}\inf_{\|\boldsymbol{C}\|_F=1}\left\{\int x^\intercal \boldsymbol{C} y\left(h_0^{\boldsymbol{A},\boldsymbol{C}}(x) + h_1^{\boldsymbol{A},\boldsymbol{C}}(y) - 32x^\intercal \boldsymbol{C} y\right) d\pi_{\boldsymbol{A}}(x,y)\right\},\end{aligned}$$

using the formula for $D^2\Phi_{[\boldsymbol{A}]}$ from Proposition 2. Recall that $(h_0^{\boldsymbol{A},\boldsymbol{C}}, h_1^{\boldsymbol{A},\boldsymbol{C}})$ satisfy

$$\int \left(h_0^{\boldsymbol{A},\boldsymbol{C}}(x) + h_1^{\boldsymbol{A},\boldsymbol{C}}(y) - 32x^\intercal \boldsymbol{C} y\right) e^{\frac{\varphi_0^{\boldsymbol{A}}(x)+\varphi_1^{\boldsymbol{A}}(y)-c_{\boldsymbol{A}}(x,y)}{\varepsilon}} d\mu_1(y) = 0, \quad \forall\, x \in \mathrm{spt}(\mu_0),$$
$$\int \left(h_0^{\boldsymbol{A},\boldsymbol{C}}(x) + h_1^{\boldsymbol{A},\boldsymbol{C}}(y) - 32x^\intercal \boldsymbol{C} y\right) e^{\frac{\varphi_0^{\boldsymbol{A}}(x)+\varphi_1^{\boldsymbol{A}}(y)-c_{\boldsymbol{A}}(x,y)}{\varepsilon}} d\mu_0(x) = 0, \quad \forall\, y \in \mathrm{spt}(\mu_1),$$

such that, multiplying the top equation by $h_0^{\boldsymbol{A},\boldsymbol{C}}$ and integrating w.r.t. μ_0 and performing the same operations on the lower equation with $h_1^{\boldsymbol{A},\boldsymbol{C}}$ and μ_1 respectively,

$$\int \left[\left(h_0^{\boldsymbol{A},\boldsymbol{C}}(x)\right)^2 + h_1^{\boldsymbol{A},\boldsymbol{C}}(y) h_0^{\boldsymbol{A},\boldsymbol{C}}(x) - 32x^\intercal \boldsymbol{C} y h_0^{\boldsymbol{A},\boldsymbol{C}}(x)\right] d\pi_{\boldsymbol{A}}(x,y) = 0,$$
$$\int \left[h_0^{\boldsymbol{A},\boldsymbol{C}}(x) h_1^{\boldsymbol{A},\boldsymbol{C}}(y) + \left(h_1^{\boldsymbol{A},\boldsymbol{C}}(y)\right)^2 - 32x^\intercal \boldsymbol{C} y h_1^{\boldsymbol{A},\boldsymbol{C}}(y)\right] d\pi_{\boldsymbol{A}}(x,y) = 0.$$

Summing these equations gives

$$32\int x^\intercal \boldsymbol{C} y \left(h_0^{\boldsymbol{A},\boldsymbol{C}}(x) + h_1^{\boldsymbol{A},\boldsymbol{C}}(y)\right) d\pi_{\boldsymbol{A}}(x,y) = \int \left(h_0^{\boldsymbol{A},\boldsymbol{C}}(x) + h_1^{\boldsymbol{A},\boldsymbol{C}}(y)\right)^2 d\pi_{\boldsymbol{A}}(x,y),$$

such that

$$\begin{aligned} &32\int x^\intercal \boldsymbol{C} y \left(h_0^{\boldsymbol{A},\boldsymbol{C}}(x) + h_1^{\boldsymbol{A},\boldsymbol{C}}(y) - 32x^\intercal \boldsymbol{C} y\right) d\pi_{\boldsymbol{A}}(x,y) \\ &= \int \left(h_0^{\boldsymbol{A},\boldsymbol{C}}(x) + h_1^{\boldsymbol{A},\boldsymbol{C}}(y)\right)^2 d\pi_{\boldsymbol{A}}(x,y) - 32^2 \int (x^\intercal \boldsymbol{C} y)^2 d\pi_{\boldsymbol{A}}(x,y), \end{aligned}$$

which proves the first part of Item (i). It remains to show that

$$\int \left(h_0^{\boldsymbol{A},\boldsymbol{C}}(x) + h_1^{\boldsymbol{A},\boldsymbol{C}}(y)\right)^2 d\pi_{\boldsymbol{A}}(x,y) - 32^2 \int (x^\intercal \boldsymbol{C} y)^2 d\pi_{\boldsymbol{A}}(x,y) \geq -32^2 \mathrm{Var}_{\pi_{\boldsymbol{A}}}[X^\intercal \boldsymbol{C} Y].$$

By Jensen's inequality, we have

$$\begin{aligned} \int \left(h_0^{\boldsymbol{A},\boldsymbol{C}}(x) + h_1^{\boldsymbol{A},\boldsymbol{C}}(y)\right)^2 d\pi_{\boldsymbol{A}}(x,y) &\geq \left(\int h_0^{\boldsymbol{A},\boldsymbol{C}}(x) + h_1^{\boldsymbol{A},\boldsymbol{C}}(y) d\pi_{\boldsymbol{A}}(x,y)\right)^2 \\ &= 32^2 \left(\int x^\intercal \boldsymbol{C} y \, d\pi_{\boldsymbol{A}}(x,y)\right)^2, \end{aligned}$$

where the equality follows from (13), proving the desired inequality.

To prove the uniform bound on the variance in Item (i), observe that

$$\begin{aligned} \sup_{\|\boldsymbol{C}\|_F=1} \mathrm{Var}_{\pi_{\boldsymbol{A}}}[X^\intercal \boldsymbol{C} Y] &\leq \sup_{\|\boldsymbol{C}\|_F=1} \mathbb{E}_{\pi_{\boldsymbol{A}}}[(X^\intercal \boldsymbol{C} Y)^2] \\ &\leq \sup_{\|\boldsymbol{C}\|_F=1} \|\boldsymbol{C}\|_F^2 \int \|x\|^2 \|y\|^2 d\pi_{\boldsymbol{A}}(x,y), \\ &\leq \sqrt{M_4(\mu_0) M_4(\mu_1)} \end{aligned}$$

where the final two inequalities follow from the Cauchy-Schwarz inequality.

We now prove the upper bound on the maximum eigenvalue of $D^2\Phi_{[\boldsymbol{A}]}$ from Item (ii) again using its variational characterization,

$$\lambda_{\max}\left(D^2\Phi_{[\boldsymbol{A}]}\right) = \sup_{\|\boldsymbol{C}\|_F=1} D^2\Phi_{[\boldsymbol{A}]}(\boldsymbol{C},\boldsymbol{C}) = 64 + \lambda_{\max}\left(D^2\mathsf{OT}_{(\cdot),\varepsilon}(\mu_0,\mu_1)_{[\boldsymbol{A}]}\right).$$

Observe that $\mathsf{OT}_{\boldsymbol{A},\varepsilon}(\mu_0,\mu_1) = \inf_{\pi\in\Pi(\mu_0,\mu_1)} g(\boldsymbol{A},\pi,\mu_0,\mu_1,\varepsilon)$, where g depends on $\boldsymbol{A}$ only through the term $32\mathrm{tr}(\boldsymbol{A}^\intercal\int xy^\intercal\, d\pi(x,y))$ which is linear in $\boldsymbol{A}$. It follows from, e.g., Proposition 2.1.2 in Hiriart-Urruty and Lemaréchal (2004) that $\mathsf{OT}_{(\cdot),\varepsilon}(\mu_0,\mu_1)$ is concave. As such, $\lambda_{\max}\left(D^2\mathsf{OT}_{(\cdot),\varepsilon}(\mu_0,\mu_1)_{[\boldsymbol{A}]}\right) \leq 0$, so $\lambda_{\max}\left(D^2\Phi_{[\boldsymbol{A}]}\right) \leq 64$. □

5.4 Proof of Theorem 6

We first discuss the convexity properties of Φ. By Corollary 5, $\lambda_{\min}\left(D^2\Phi_{[\boldsymbol{A}]} + \frac{\rho}{2}\|\boldsymbol{A}\|_F^2\right) \geq 64 - 32^2\varepsilon^{-1}\sqrt{M_4(\mu_0)M_4(\mu_1)} + \rho$ for any $\boldsymbol{A}\in\mathbb{R}^{d_0\times d_1}$ and $\rho\geq 0$. When this lower bound is nonnegative, Φ is ρ-weakly convex on $\mathbb{R}^{d_0\times d_1}$ by definition; recall Section 3. It follows that Φ is always ρ-weakly convex for $\rho = 32^2\varepsilon^{-1}\sqrt{M_4(\mu_0)M_4(\mu_1)} - 64$. Moreover, if $\sqrt{M_4(\mu_0)M_4(\mu_1)} < \frac{\varepsilon}{16}$, then $\lambda_{\min}\left(D^2\Phi_{[\boldsymbol{A}]}\right) > 0$ such that Φ is strictly convex.

L-smoothness of Φ follows from the mean value inequality (see e.g. Example 2 p. 356 in Apostol 1974)

$$\begin{aligned}\|D\Phi_{[\boldsymbol{A}]} - D\Phi_{[\boldsymbol{B}]}\|_F &\leq \sup_{\boldsymbol{C}\in[\boldsymbol{A},\boldsymbol{B}]}\sup_{\|\boldsymbol{E}\|_F=1}\left|D^2\Phi_{[\boldsymbol{C}]}\left(\boldsymbol{A}-\boldsymbol{B},\boldsymbol{E}\right)\right|,\\ &\leq \sup_{\boldsymbol{C}\in[\boldsymbol{A},\boldsymbol{B}]}\left(\left|\lambda_{\min}\left(D^2\Phi_{[\boldsymbol{C}]}\right)\right| \vee \left|\lambda_{\max}\left(D^2\Phi_{[\boldsymbol{C}]}\right)\right|\right)\|\boldsymbol{A}-\boldsymbol{B}\|_F,\end{aligned}$$

for any $\boldsymbol{A},\boldsymbol{B}\in\mathbb{R}^{d_0\times d_1}$, where $[\boldsymbol{A},\boldsymbol{B}]$ denotes the line segment connecting $\boldsymbol{A}$ and $\boldsymbol{B}$. The claimed result then follows by noting that, for any $\boldsymbol{A},\boldsymbol{B}\in\mathcal{D}_M$, $[\boldsymbol{A},\boldsymbol{B}]\subset\mathcal{D}_M$ by convexity and the supremum over $\mathcal{D}_M$ is achieved by compactness and the fact that the objective is continuous. Indeed, the maps $\lambda_{\max}(\cdot),\lambda_{\min}(\cdot)$ are continuous on the space of symmetric matrices, and $D^2\Phi_{[\cdot]}$ is continuous as Φ is smooth. □

5.5 Proof of Theorem 9

In this section, we show that Theorem 2.2 in d'Aspremont (2008) on the convergence rate of Algorithm 1 is applicable in our setting. We particularize their result to a fixed prox-function $d = \frac{1}{2}\|\cdot\|_F^2$ which is smooth and 1-strongly convex.

First, we justify the expressions for the iterates $\boldsymbol{B}_k,\boldsymbol{C}_k$ in Algorithm 1, which are defined in d'Aspremont (2008) as the proximal operators

$$\begin{aligned}\boldsymbol{B}_k &= \operatorname*{argmin}_{\boldsymbol{V}\in\mathcal{D}_M}\left\{\mathrm{tr}\left(\boldsymbol{G}_k^\intercal\boldsymbol{V}\right) + \frac{L}{2}\|\boldsymbol{V}-\boldsymbol{A}_k\|_F^2\right\},\\ \boldsymbol{C}_k &= \operatorname*{argmin}_{\boldsymbol{V}\in\mathcal{D}_M}\left\{\mathrm{tr}\left(\boldsymbol{W}_k^\intercal\boldsymbol{V}\right) + \frac{L}{2}\|\boldsymbol{V}\|_F^2\right\}.\end{aligned}$$

Rearranging terms, both problems can be written, equivalently, as

$$\operatorname*{argmin}_{\boldsymbol{V}\in\mathcal{D}_M}\left\{\|\boldsymbol{V}-\boldsymbol{U}\|_F^2\right\}, \tag{15}$$

for $\boldsymbol{U} = \boldsymbol{A}_k - L^{-1}\boldsymbol{G}_k$ and $\boldsymbol{U} = -L^{-1}\boldsymbol{W}_k$ for the $\boldsymbol{B}_k$ and $\boldsymbol{C}_k$ iterations respectively. The solution of (15) is given by $\boldsymbol{V} = \boldsymbol{U}$ if $\boldsymbol{V} = \boldsymbol{U} \in \mathcal{D}_M$, and $\frac{M}{2}\boldsymbol{U}/\|\boldsymbol{U}\|_F$ otherwise.

Next, we show that our notion of δ-oracle yields a δ'-approximate gradient in the sense of Equation (2.3) in d'Aspremont (2008). Precisely, we prove that

$$\left|\operatorname{tr}\left(\left(\widetilde{D}\Phi_{[\boldsymbol{A}]} - D\Phi_{[\boldsymbol{A}]}\right)^{\intercal}(\boldsymbol{B}-\boldsymbol{C})\right)\right| \leq \delta', \tag{16}$$

for any $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C} \in \mathcal{D}_M$. By the Cauchy-Schwarz inequality,

$$\left|\operatorname{tr}\left(\left(\widetilde{D}\Phi_{[\boldsymbol{A}]} - D\Phi_{[\boldsymbol{A}]}\right)^{\intercal}(\boldsymbol{B}-\boldsymbol{C})\right)\right| \leq M\left\|\widetilde{D}\Phi_{[\boldsymbol{A}]} - D\Phi_{[\boldsymbol{A}]}\right\|_F.$$

Recall that

$$\widetilde{D}\Phi_{[\boldsymbol{A}]} - D\Phi_{[\boldsymbol{A}]} = 32\sum_{\substack{1\leq i\leq N_0\\1\leq j\leq N_1}} x^{(i)}\left(y^{(j)}\right)^{\intercal}\left(\widetilde{\boldsymbol{\Pi}}^{\boldsymbol{A}}_{ij} - \boldsymbol{\Pi}^{\boldsymbol{A}}_{ij}\right),$$

where $\left\|\widetilde{\boldsymbol{\Pi}}^{\boldsymbol{A}} - \boldsymbol{\Pi}^{\boldsymbol{A}}\right\|_\infty < \delta$ uniformly in $\boldsymbol{A} \in \mathcal{D}_M$ by the δ-oracle assumption such that

$$\left\|\widetilde{D}\Phi_{[\boldsymbol{A}]} - D\Phi_{[\boldsymbol{A}]}\right\|_F \leq 32\left\|\widetilde{\boldsymbol{\Pi}}^{\boldsymbol{A}} - \boldsymbol{\Pi}^{\boldsymbol{A}}\right\|_\infty \sum_{\substack{1\leq i\leq N_0\\1\leq j\leq N_1}} \left\|x^{(i)}\left(y^{(j)}\right)^{\intercal}\right\|_F < 32\delta \sum_{\substack{1\leq i\leq N_0\\1\leq j\leq N_1}} \left\|x^{(i)}\right\|\left\|y^{(j)}\right\|. \tag{17}$$

Combining the displayed equations yields

$$\left|\operatorname{tr}\left(\left(\widetilde{D}\Phi_{[\boldsymbol{A}]} - D\Phi_{[\boldsymbol{A}]}\right)^{\intercal}(\boldsymbol{B}-\boldsymbol{C})\right)\right| \leq 32M\delta \sum_{\substack{1\leq i\leq N_0\\1\leq j\leq N_1}} \left\|x^{(i)}\right\|\left\|y^{(j)}\right\| = \delta',$$

proving (16).

With these preparations Theorem 9 follows from Theorem 2.2 in d'Aspremont (2008) and the discussion following its proof, noting that $\sum_{i=0}^{k}\frac{i+1}{2} = \frac{(k+1)(k+2)}{4}$. □

5.6 Proof of Corollary 12

As $\boldsymbol{A}_k, \boldsymbol{B}_k$ be iterates from Algorithm 2 with $\boldsymbol{B}_k \in \operatorname{int}(\mathcal{D}_M)$ such that $\boldsymbol{B}_k = \boldsymbol{A}_k - \beta_k\widetilde{D}\Phi_{[\boldsymbol{A}_k]}$ by definition. By the triangle inequality,

$$\|D\Phi_{[\boldsymbol{A}_k]}\|_F \leq \|D\Phi_{[\boldsymbol{A}_k]} - \widetilde{D}\Phi_{[\boldsymbol{A}_k]}\|_F + \|\widetilde{D}\Phi_{[\boldsymbol{A}_k]}\|_F = \|D\Phi_{[\boldsymbol{A}_k]} - \widetilde{D}\Phi_{[\boldsymbol{A}_k]}\|_F + \|\beta_k^{-1}(\boldsymbol{B}_k - \boldsymbol{A}_k)\|_F.$$

(17) further yields

$$\|D\Phi_{[\boldsymbol{A}_k]} - \widetilde{D}\Phi_{[\boldsymbol{A}_k]}\|_F < 32\delta \sum_{\substack{1\leq i\leq N_0\\1\leq j\leq N_1}} \|x^{(i)}\|\|y^{(j)}\|,$$

proving the claim. □

5.7 Proof of Proposition 15

Let $S_i = \mathrm{spt}(\mu_i)$ for $i = 0, 1$. Recall that S_0, S_1 are compact by assumption. The proof of Proposition 15 follows by verifying optimality conditions for minimizing locally Lipschitz functions using the Clarke subdifferential (Clarke, 1990). To this end, we first verify that the objective Φ_0 is locally Lipschitz and prove several auxiliary results to characterize the Clarke subdifferential $\partial\Phi_0$, recalling that the Clarke subdifferential of a locally Lipschitz function $f : \mathbb{R}^{d_0 \times d_1} \to \mathbb{R}$ at a point $\boldsymbol{A} \in \mathbb{R}^{d_0 \times d_1}$ is given by (cf. e.g. Theorem 2.5.1 in Clarke 1990)

$$\partial f(x) = \mathrm{conv}\left(\left\{\lim_{n\to\infty} Df_{[\boldsymbol{A}_n]} : U \not\ni \boldsymbol{A}_n \to \boldsymbol{A}\right\}\right), \tag{18}$$

where $\mathrm{conv}(A)$ denotes the convex hull of the set A, $U \subset \mathbb{R}^{d_0 \times d_1}$ denotes a set of full measure on which f is differentiable, the existence of which is guaranteed by Rademacher's theorem, and we tacitly restrict this definition to convergent sequences of derivatives.

Lemma 20. *The function $\boldsymbol{A} \in \mathbb{R}^{d_0 \times d_1} \mapsto \Phi_0(\boldsymbol{A})$ is locally Lipschitz continuous and coercive.*

Proof We start by proving that Φ_0 is locally Lipschitz. Fix a compact set $\boldsymbol{K} \subset \mathbb{R}^{d_0 \times d_1}$ and observe that, for any $\boldsymbol{A}, \boldsymbol{A'} \in \boldsymbol{K}$,

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

due to the reverse triangle inequality. This shows that $\|\cdot\|_F^2$ is locally Lipschitz. As for $\mathsf{OT}_{(\cdot),0}(\mu_0, \mu_1)$, let $\pi_{\boldsymbol{A}}, \pi_{\boldsymbol{A'}}$ be solutions of $\mathsf{OT}_{\boldsymbol{A},0}(\mu_0, \mu_1), \mathsf{OT}_{\boldsymbol{A'},0}(\mu_0, \mu_1)$ respectively, then

$$\mathsf{OT}_{\boldsymbol{A},0}(\mu_0, \mu_1) - \mathsf{OT}_{\boldsymbol{A'},0}(\mu_0, \mu_1) \geq \int c_{\boldsymbol{A}} d\pi_{\boldsymbol{A}} - \int c_{\boldsymbol{A'}} d\pi_{\boldsymbol{A}} = -32\left\langle \boldsymbol{A} - \boldsymbol{A'}, \int xy^\intercal d\pi_{\boldsymbol{A}}(x, y)\right\rangle_F, \tag{19}$$

and similarly $\mathsf{OT}_{\boldsymbol{A},0}(\mu_0, \mu_1) - \mathsf{OT}_{\boldsymbol{A'},0}(\mu_0, \mu_1) \leq -32\langle \boldsymbol{A} - \boldsymbol{A'}, \int xy^\intercal d\pi_{\boldsymbol{A'}}(x, y)\rangle_F$. Thus,

$$\left|\mathsf{OT}_{\boldsymbol{A},0}(\mu_0, \mu_1) - \mathsf{OT}_{\boldsymbol{A'},0}(\mu_0, \mu_1)\right| \leq 16M\|\boldsymbol{A} - \boldsymbol{A'}\|_F,$$

recalling that, for any $\pi \in \Pi(\mu_0, \mu_1)$, $\int xy^\intercal d\pi \in \mathcal{D}_M$. This proves that Φ_0 is locally Lipschitz. Coercivity follows by adapting the proof of Proposition 2. ■

By Lemma 20 and Rademacher's theorem, Φ_0 is differentiable almost everywhere on $\mathbb{R}^{d_0 \times d_1}$. As the squared Frobenius norm is smooth, we study differentiability of $\mathsf{OT}_{\boldsymbol{A},0}(\mu_0, \mu_1)$. To simplify notation, let $\Pi^\star_{\boldsymbol{A},0}$ denote the set of optimal solutions to $\mathsf{OT}_{\boldsymbol{A},0}(\mu_0, \mu_1)$.

Lemma 21. *The function $\boldsymbol{A} \in \mathbb{R}^{d_0 \times d_1} \mapsto -\mathsf{OT}_{\boldsymbol{A},0}(\mu_0, \mu_1)$ is convex; its subdifferential*[11] *at $\boldsymbol{A}$ contains $\left\{32\int xy^\intercal d\pi : \pi \in \Pi^\star_{\boldsymbol{A},0}\right\}$.*

Proof As $\mathsf{OT}_{(\cdot),0}(\mu_0, \mu_1)$ is the infimum of a family of affine functions, it is concave (see Theorem 5.5 in Rockafellar 1997). The second claim follows directly from (19). ■

With Lemma 21, it is easy to classify the points at which $\mathsf{OT}_{(\cdot),0}(\mu_0, \mu_1)$ is differentiable.

11. The subdifferential of a convex function $f : \mathbb{R}^{d_0 \times d_1} \mapsto \mathbb{R}$ at $\boldsymbol{A}$ consists of all $\boldsymbol{\Xi} \in \mathbb{R}^{d_0 \times d_1}$ for which $f(\boldsymbol{A'}) - f(\boldsymbol{A}) \geq \langle \boldsymbol{A'} - \boldsymbol{A}, \boldsymbol{\Xi}\rangle_F$ for every $\boldsymbol{A'} \in \mathbb{R}^{d_0 \times d_1}$.

Lemma 22. *$-\mathsf{OT}_{(\cdot),0}(\mu_0,\mu_1)$ is differentiable at $\boldsymbol{A}$ with derivative $-D(\mathsf{OT}_{(\cdot),0}(\mu_0,\mu_1))_{[\boldsymbol{A}]} = 32\int xy^\intercal d\pi(x,y)$ for any $\pi \in \Pi^\star_{\boldsymbol{A},0}$ if and only if all couplings in $\Pi^\star_{\boldsymbol{A},0}$ admit the same cross-correlation matrix.*

Proof We recall that a convex function is differentiable at $\boldsymbol{A}$ precisely when its subdifferential at $\boldsymbol{A}$ is a singleton, see Theorem 25.1 in Rockafellar (1997). If the proposed condition on the cross-correlation matrices fails, Lemma 22 implies that the subdifferential is not a singleton, so differentiability fails.

To prove the other direction, assume that all couplings in $\Pi^\star_{\boldsymbol{A},0}$ admit the same cross-correlation matrix. Fix a sequence $\boldsymbol{H}_n \in \mathbb{R}^{d_0\times d_1}\backslash\{\boldsymbol{0}\}$ with $\|\boldsymbol{H}_n\|_F \downarrow 0$. From (19), we have, for any $\pi_{\boldsymbol{A}+\boldsymbol{H}_n} \in \Pi^\star_{\boldsymbol{A}+\boldsymbol{H}_n,0}$ and $\pi_{\boldsymbol{A}} \in \Pi^\star_{\boldsymbol{A},0}$,

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

As $c_{\boldsymbol{A}+\boldsymbol{H}_n}$ converges uniformly to $c_{\boldsymbol{A}}$ on $S_0\times S_1$, for any subsequence n' of n there exists a further subsequence n'' along which $\pi_{\boldsymbol{A}+\boldsymbol{H}_{n''}} \overset{w}{\to} \pi \in \Pi^\star_{\boldsymbol{A},0}$ by Theorem 5.20 in Villani (2008). Since $(x,y)\in S_0\times S_1 \mapsto x_i y_j$ is continuous and bounded for any $1\leq i\leq d_0, 1\leq j\leq d_1$, $\left\|\int xy^\intercal d\pi_{\boldsymbol{A}+\boldsymbol{H}_{n''}}(x,y) - \int xy^\intercal d\pi(x,y)\right\|_F = \left\|\int xy^\intercal d\pi_{\boldsymbol{A}+\boldsymbol{H}_{n''}}(x,y) - \int xy^\intercal d\pi_{\boldsymbol{A}}(x,y)\right\|_F \to 0$. As this limit is independent of the choice of subsequence by assumption, it holds along the original sequence $\boldsymbol{H}_n$ such that $D\left(\mathsf{OT}_{(\cdot),0}(\mu_0,\mu_1)\right)_{[\boldsymbol{A}]} = -32\int xy^\intercal d\pi_{\boldsymbol{A}}(x,y)$. ■

Those auxiliary results lead to finding the Clarke subdifferential $\partial\Phi_0$, as given below.

Lemma 23. *The Clarke subdifferential of Φ_0 at $\boldsymbol{A}\in\mathbb{R}^{d_0\times d_1}$ is given by*

$$\partial\Phi_0(\boldsymbol{A}) = \left\{64\boldsymbol{A} - 32\int xy^\intercal d\pi : \pi\in\Pi^\star_{\boldsymbol{A},0}\right\}.$$

Proof We first characterize the Clarke subdifferential of $\mathsf{OT}_{(\cdot),0}(\mu_0,\mu_1)$ at $\boldsymbol{A}$. By Propositions 2.2.7 and 2.3.1 in Clarke (1990) along with Lemma 21,

$$\partial\left(\mathsf{OT}_{(\cdot),0}(\mu_0,\mu_1)\right)(\boldsymbol{A}) \supset \left\{-32\int xy^\intercal d\pi(x,y) : \pi\in\Pi^\star_{\boldsymbol{A},0}\right\}. \tag{20}$$

As Lemma 22 establishes the set $U\subset\mathbb{R}^{d_0\times d_1}$ on which Φ_0 is differentiable, we study the Clarke subdifferential of Φ_0 through the lens of (18). Observe that if $U\not\ni \boldsymbol{A}_n\to\boldsymbol{A}$ and $\pi_{\boldsymbol{A}_n}\in\Pi^\star_{\boldsymbol{A}_n,0}$, then $\pi_{\boldsymbol{A}_n}\overset{w}{\to}\pi_{\boldsymbol{A}}\in\Pi^\star_{\boldsymbol{A},0}$ as in the proof of Lemma 22 (up to a subsequence). Moreover,

$$\lim_{n\to\infty} D\left(\mathsf{OT}_{(\cdot),0}(\mu_0,\mu_1)\right)_{[\boldsymbol{A}_n]} = \lim_{n\to\infty} -32\int xy^\intercal d\pi_{\boldsymbol{A}_n}(x,y) = -32\int xy^\intercal d\pi_{\boldsymbol{A}}(x,y)$$

along this subsequence. Thus, if $\boldsymbol{A}'_n\notin U$ converges to $\boldsymbol{A}$ and $\lim_{n\to\infty} D\left(\mathsf{OT}_{(\cdot),0}(\mu_0,\mu_1)\right)_{[\boldsymbol{A}'_n]}$ exists, then the limit is given by $-32\int xy^\intercal d\pi(x,y)$ for some $\pi\in\Pi^\star_{\boldsymbol{A},0}$. It is easy to see that

$\Pi^\star_{\boldsymbol{A},0}$ is convex, so (20) and (18) together imply that

$$\partial\left(\mathsf{OT}_{(\cdot),0}(\mu_0,\mu_1)\right)(\boldsymbol{A}) = \left\{-32\int xy^\intercal d\pi(x,y) : \pi \in \Pi^\star_{\boldsymbol{A},0}\right\}.$$

Conclude by applying the subdifferential sum rule (Corollary 1 on p. 39 of Clarke 1990). ■

Proof of Proposition 15 By Proposition 2.3.2 in Clarke (1990), if Φ_0 attains a local minimum at $\boldsymbol{A}^\star$, then $0 \in \partial\Phi_0(\boldsymbol{A}^\star)$, i.e., there exists $\pi \in \Pi^\star_{\boldsymbol{A}^\star,0}$ for which $\boldsymbol{A}^\star = \frac{1}{2}\int xy^\intercal d\pi(x,y)$ by Lemma 23. The second result on optimality for the original GW problem follows directly from the proof of Corollary 4. ■

5.8 Proof of Theorem 16

Throughout, let $\mathcal{X}_i \subset \mathbb{R}^{d_i}$ be a closed ball with finite radius $r > 0$ for which $\mathrm{spt}(\mu_i) \subset \mathcal{X}_i$ for $i = 0, 1$. By the Bolzano-Weierstrass theorem, $\boldsymbol{A}^\star_\varepsilon$ admits a limit point $\boldsymbol{A}^\star_0$ as $\varepsilon \downarrow 0$. Let $\varepsilon_k \downarrow 0$ be a sequence along which $\boldsymbol{A}^\star_{\varepsilon_k} \to \boldsymbol{A}^\star_0$ and define

$$F_k : \pi \in \mathcal{P}(\mathcal{X}_0 \times \mathcal{X}_1) \mapsto \begin{cases} \int c_{\boldsymbol{A}^\star_{\varepsilon_k}} d\pi + \varepsilon_k \mathsf{D}_{\mathsf{KL}}(\pi \| \mu_0 \otimes \mu_1), & \text{if } \pi \in \Pi(\mu_0,\mu_1), \\ +\infty, & \text{otherwise,} \end{cases}$$

$$F : \pi \in \mathcal{P}(\mathcal{X}_0 \times \mathcal{X}_1) \mapsto \begin{cases} \int c_{\boldsymbol{A}^\star_0} d\pi, & \text{if } \pi \in \Pi(\mu_0,\mu_1), \\ +\infty, & \text{otherwise.} \end{cases}$$

Throughout, we endow $\mathcal{P}(\mathcal{X}_0 \times \mathcal{X}_1)$ with the topology induced by the weak convergence of probability measures which is metrized by the 2-Wasserstein distance, W_2, for instance; $(\mathcal{P}(\mathcal{X}_0 \times \mathcal{X}_1), \mathsf{W}_2)$ is notably a separable metric space, see Theorem 6.18 in Villani (2008).

We first show that F_k Γ-converges to F according to Definition 2.3 in Braides (2014). To this end, we show that

$$F(\pi) \leq \liminf_{k\to\infty} F_k(\pi_k), \text{ for every } \pi_k \xrightarrow{w} \pi. \tag{21}$$

and exhibit a sequence $\pi_k \xrightarrow{w} \pi$ satisfying

$$F(\pi) \geq \limsup_{k\to\infty} F_k(\pi_k). \tag{22}$$

To prove (21), first assume that $\pi \in \Pi(\mu_0,\mu_1)$. Then, if $\pi_k \xrightarrow{w} \pi$, $F_k(\pi_k) \geq \int c_{\boldsymbol{A}^\star_{\varepsilon_k}} d\pi_k$, by nonnegativity of the KL divergence. As $c_{\boldsymbol{A}^\star_{\varepsilon_k}}$ converges to $c_{\boldsymbol{A}^\star_0}$ uniformly on $\mathcal{X}_0 \times \mathcal{X}_1$ and $c_{\boldsymbol{A}^\star_0}$ is continuous and bounded,

$$\left|\int c_{\boldsymbol{A}^\star_{\varepsilon_k}} d\pi_k - \int c_{\boldsymbol{A}^\star_0} d\pi\right| \leq \left|\int c_{\boldsymbol{A}^\star_{\varepsilon_k}} - c_{\boldsymbol{A}^\star_0} d\pi_k\right| + \left|\int c_{\boldsymbol{A}^\star_0} d\pi - \int c_{\boldsymbol{A}^\star_0} d\pi_k\right| \to 0. \tag{23}$$

This shows that, if $\pi \in \Pi(\mu_0,\mu_1)$, then

$$\liminf_{k\to\infty} F_k(\pi_k) \geq \lim_{k\to\infty} \int c_{\boldsymbol{A}^\star_{\varepsilon_k}} d\pi_k = F(\pi).$$

If $\pi \notin \Pi(\mu_0, \mu_1)$, then $\pi_k \notin \Pi(\mu_0, \mu_1)$ for all k sufficiently large as $\Pi(\mu_0, \mu_1)$ is compact for the weak convergence on $\mathcal{P}(\mathcal{X}_0 \times \mathcal{X}_1)$ (cf. e.g. the proof of Theorem 1.4 in Santambrogio 2015) so the bound holds vacuously, proving (21).

As for (22), if $\pi \notin \Pi(\mu_0, \mu_1)$, then there is nothing to show. For $\pi \in \Pi(\mu_0, \mu_1)$, we consider a block approximation; cf. Carlier et al. (2017); Genevay et al. (2019). First, for $i = 0, 1$, partition $\mathbb{R}^{d_i}$ by hypercubes of length $\ell > 0$,

$$\{H_{i,q,\ell} = [k_1\ell, (q_1+1)\ell) \times \cdots \times [q_{d_i}\ell, (q_{d_i}+1)\ell) : q = (q_1, \ldots, q_{d_i}) \in \mathbb{Z}^{d_i}\},$$

and define the block approximation, π_ℓ, of π by

$$\pi_\ell = \sum_{(q,q') \in \mathbb{Z}^{d_0} \times \mathbb{Z}^{d_1}} \pi_\ell|_{H_{0,q,\ell} \times H_{1,q',\ell}},$$
$$\pi_\ell|_{H_{0,q,\ell} \times H_{1,q',\ell}} = \frac{\pi(H_{0,q,\ell} \times H_{1,q',\ell})}{\mu_0 \otimes \mu_1(H_{0,q,\ell} \times H_{1,q',\ell})} (\mu_0|_{H_{0,q,\ell}} \otimes \mu_1|_{H_{1,q',\ell}}),$$

for every $(q, q') \in \mathbb{Z}^{d_0} \times \mathbb{Z}^{d_1}$ with the convention $\frac{0}{0} = 0$. Here, $\mu_0|_{H_{0,q,\ell}}(A) = \mu_0(A \cap H_{0,q,\ell})$ and similarly for the other restrictions. Of note is that $\pi_\ell \ll \mu_0 \otimes \mu_1$ and that $\pi_\ell \in \Pi(\mu_0, \mu_1)$ (see the discussion surrounding Definition 1 in Genevay et al. 2019).

Lemma 24. *The block approximation π_ℓ of π converges weakly to π as $\ell \downarrow 0$.*

Proof Let $Q = \left\{(q, q') \in \mathbb{Z}^{d_0} \times \mathbb{Z}^{d_1} : \pi(H_{0,q,\ell} \times H_{1,q',\ell}) > 0\right\}$ and $\gamma = \sum_{(q,q') \in Q} \pi(H_{0,q,\ell} \times H_{1,q',\ell}) \gamma_{q,q'}$ where $\gamma_{q,q'}$ is any coupling of the measures $\left(\frac{\pi|_{H_{0,q,\ell} \times H_{1,q',\ell}}}{\pi(H_{0,q,\ell} \times H_{1,q',\ell})}, \frac{\pi_\ell|_{H_{0,q,\ell} \times H_{1,q',\ell}}}{\pi_\ell(H_{0,q,\ell} \times H_{1,q',\ell})}\right)$ with support in $\overline{H_{0,q,\ell} \times H_{1,q',\ell}}$ (the closure of $H_{0,q,\ell} \times H_{1,q',\ell}$). As $\pi_\ell(H_{0,q,\ell} \times H_{1,q',\ell}) = \pi(H_{0,q,\ell} \times H_{1,q',\ell})$ by construction, it is readily seen that $\gamma \in \Pi(\pi, \pi_\ell)$. Thus,

$$\begin{aligned}\mathsf{W}_2^2(\pi, \pi_\ell) \leq \int \|x - y\|^2 d\gamma(x, y) &= \sum_{(q,q') \in Q} \pi(H_{0,q,\ell} \times H_{1,q',\ell}) \int \|x - y\|^2 d\gamma_{q,q'}(x, y) \\ &\leq (d_0 + d_1)\ell^2,\end{aligned}$$

noting that $\operatorname{diam}(H_{0,q,\ell} \times H_{1,q',\ell}) = \sqrt{d_0 + d_1}\ell$ is the diameter of a hypercube in $\mathbb{R}^{d_0+d_1}$ of length ℓ. Conclude that $\mathsf{W}_2(\pi, \pi_\ell) \to 0$ as $\ell \downarrow 0$ such that $\pi_\ell \xrightarrow{w} \pi$. ■

Setting $\ell_k = \varepsilon_k$, Lemma 24 yields

$$\lim_{k \to \infty} \int c_{\boldsymbol{A}^\star_{\varepsilon_k}} d\pi_{\varepsilon_k} = \int c_{\boldsymbol{A}^\star_0} d\pi,$$

by a simple adaption of (23). It remains to show that $\lim_{k\to\infty} \varepsilon_k \mathsf{D}_{\mathsf{KL}}(\pi_{\varepsilon_k} \| \mu_0 \otimes \mu_1) = 0$ such that

$$\lim_{k \to \infty} F_k(\pi_{\varepsilon_k}) \to F(\pi).$$

Lemma 25. *The block approximation π_ℓ of π satisfies $\varepsilon_k \mathsf{D}_{\mathsf{KL}}(\pi_{\varepsilon_k} \| \mu_0 \otimes \mu_1) \to 0$ as $k \to \infty$.*

Proof As $\pi(H_{0,q,\ell} \times H_{1,q',\ell}) \leq 1$,

$$\begin{aligned}
\mathsf{D}_{\mathsf{KL}}(\pi^\ell \| \mu_0 \otimes \mu_1) &= \sum_{(q,q') \in \mathbb{Z}^{d_0} \times \mathbb{Z}^{d_1}} \int \log\left(\frac{\pi\left(H_{0,q,\ell} \times H_{1,q',\ell}\right)}{\mu_0 \otimes \mu_1 (H_{0,q,\ell} \times H_{1,q',\ell})} \right) d\pi_\ell|_{H_{0,q,\ell} \times H_{1,q',\ell}} \\
&\leq \sum_{(q,q') \in \mathbb{Z}^{d_0} \times \mathbb{Z}^{d_1}} \int -\log(\mu_0(H_{0,q,\ell})) - \log(\mu_1(H_{1,q',\ell})) d\pi_\ell|_{H_{0,q,\ell} \times H_{1,q',\ell}} \\
&= \sum_{(q,q') \in \mathbb{Z}^{d_0} \times \mathbb{Z}^{d_1}} \left(-\log(\mu_0(H_{0,q,\ell})) - \log(\mu_1(H_{1,q',\ell}))\right) \pi(H_{0,q,\ell} \times H_{1,q',\ell}) \\
&= -\sum_{q \in \mathbb{Z}^{d_0}} \log(\mu_0(H_{0,q,\ell})) \mu_0(H_{0,q,\ell}) - \sum_{q' \in \mathbb{Z}^{d_1}} \log(\mu_1(H_{1,q',\ell})) \mu_1(H_{1,q',\ell}).
\end{aligned} \tag{24}$$

Observe that

$$\begin{aligned}
&-\sum_{q \in \mathbb{Z}^{d_0}} \log(\mu_0(H_{0,q,\ell})) \mu_0(H_{0,q,\ell}) \\
&= -\int \log\left(\frac{\mu_0(H_{0,q,\ell})}{\ell^{d_0}} \mathbb{1}_{H_{0,q,\ell}} \right) \frac{\mu_0(H_{0,q,\ell})}{\ell^{d_0}} \mathbb{1}_{H_{0,q,\ell}} d\lambda - d_0 \log(\ell),
\end{aligned}$$

where λ denotes the Lebesgue measure. The first term on the last line is the differential entropy of a probability distribution supported on $\mathcal{X}_{0,\ell} = \cup_{q \in I} H_{0,q,\ell}$, where $I = \left\{q \in \mathbb{Z}^{d_0} : H_{0,q,\ell} \cap \mathcal{X}_0 \neq \emptyset\right\}$. This quantity is maximized among all probability distributions supported on $\mathcal{X}_{0,\ell}$ which are absolutely continuous w.r.t. the Lebesgue measure by the uniform distribution on $\mathcal{X}_{0,\ell}$ with value $d_0 \log(\lambda(\mathcal{X}_{0,\ell}))$. With this and (24), we obtain

$$\mathsf{D}_{\mathsf{KL}}(\pi^\ell \| \mu_0 \otimes \mu_1) \leq d_0 \log\left(\frac{\lambda(\mathcal{X}_{0,\ell})}{\ell} \right) + d_1 \log\left(\frac{\lambda(\mathcal{X}_{1,\ell})}{\ell} \right).$$

Conclude that

$$0 \leq \varepsilon_k \mathsf{D}_{\mathsf{KL}}(\pi_{\varepsilon_k} \| \mu_0 \otimes \mu_1) \leq \varepsilon_k \left(d_0 \log\left(\frac{\lambda(\mathcal{X}_{0,\varepsilon_k})}{\varepsilon_k} \right) + d_1 \log\left(\frac{\lambda(\mathcal{X}_{1,\varepsilon_k})}{\varepsilon_k} \right) \right) \to 0,$$

as $\lambda(\mathcal{X}_{i,\varepsilon_k}) \geq \lambda(\mathcal{X}_i) > 0$ for $i = 0, 1$. ■

With this, we have shown that F_k Γ-converges to F. Now, let π_{ε_k} minimize F_k. As $\Pi(\mu_0, \mu_1)$ is compact, π_{ε_k} admits a cluster point π_0 which minimizes F by Theorem 2.1 in Braides (2014). Thus, π_0 is an optimal solution of $\mathsf{OT}_{\boldsymbol{A}_0^\star}(\mu_0, \mu_1)$ and, along the subsequence where $\pi_{\varepsilon_k} \xrightarrow{w} \pi_0$,

$$\left\| 64\boldsymbol{A}_0^\star - 32 \int xy^\intercal d\pi_0 \right\|_F = \lim_{k \to \infty} \left\| 64\boldsymbol{A}_{\varepsilon_k}^\star - 32 \int xy^\intercal d\pi_{\varepsilon_k} \right\|_F \leq \delta.$$

This concludes the proof of the first result.

For the results pertaining to local/global optimality, let $\varepsilon_k \downarrow 0$ be arbitrary and define

$$G_k : \boldsymbol{A} \in \mathbb{R}^{d_0 \times d_1} \mapsto \begin{cases} \Phi_{\varepsilon_k}, & \text{if } \boldsymbol{A} \in \mathcal{D}_M, \\ +\infty, & \text{otherwise,} \end{cases}, \quad G : \boldsymbol{A} \in \mathbb{R}^{d_0 \times d_1} \mapsto \begin{cases} \Phi_0, & \text{if } \boldsymbol{A} \in \mathcal{D}_M, \\ +\infty, & \text{otherwise.} \end{cases}$$

We now show that G_k Γ-converges to G. Observe that, for any $\boldsymbol{A}_k \to \boldsymbol{A} \in \mathcal{D}_M$,

$$\Phi_{\varepsilon_k}(\boldsymbol{A}_k) = 32\|\boldsymbol{A}_k\|_F^2 + \mathsf{OT}_{\boldsymbol{A}_k,\varepsilon_k}(\mu_0,\mu_1) \geq 32\|\boldsymbol{A}_k\|_F^2 + \mathsf{OT}_{\boldsymbol{A}_k,0}(\mu_0,\mu_1),$$

where the inequality is due to nonnegativity of the KL divergence. As $c_{\boldsymbol{A}_k} \to c_{\boldsymbol{A}}$ uniformly, $32\|\boldsymbol{A}_k\|_F^2 + \mathsf{OT}_{\boldsymbol{A}_k,0}(\mu_0,\mu_1) \to 32\|\boldsymbol{A}\|_F^2 + \mathsf{OT}_{\boldsymbol{A},0}(\mu_0,\mu_1) = \Phi_0(\boldsymbol{A})$. If $\boldsymbol{A} \notin \mathcal{D}_M$, then $\boldsymbol{A}_k \notin \mathcal{D}_M$ for every k sufficiently large, so the lim inf condition (21) holds.

For the lim sup condition (22), if $\boldsymbol{A} \notin \mathcal{D}_M$, then the bound is vacuous. If $\boldsymbol{A} \in \mathcal{D}_M$, then let $\pi_{\boldsymbol{A}}$ be an optimal solution of $\mathsf{OT}_{\boldsymbol{A},0}(\mu_0,\mu_1)$ and π_{ε_k} be the block approximation of $\pi_{\boldsymbol{A}}$ with $\ell = \varepsilon_k$. Then,

$$0 \leq \mathsf{OT}_{\boldsymbol{A},\varepsilon_k}(\mu_0,\mu_1) - \mathsf{OT}_{\boldsymbol{A},0}(\mu_0,\mu_1) \leq \int c_{\boldsymbol{A}} d\pi_{\varepsilon_k} + \varepsilon_k \mathsf{D}_{\mathsf{KL}}(\pi_{\varepsilon_k} \| \mu_0 \otimes \mu_1) - \int c_{\boldsymbol{A}} d\pi_{\boldsymbol{A}}.$$

We have shown previously that the rightmost term converges to 0 as $k \to \infty$, so $\Phi_{\varepsilon_k}(\boldsymbol{A}) \to \Phi_0(\boldsymbol{A})$, such that G_k Γ-converges to G. By Theorem 2.1 in Braides (2014), any cluster point of a sequence of minizers of G_k minimizes G, proving the claim.

Finally, consider the case where $(\boldsymbol{A}_{\varepsilon_k})_{k\in\mathbb{N}}$ satisfies the conditions of part 3 of Theorem 16, i.e., $\boldsymbol{A}_{\varepsilon_k} \to \boldsymbol{A}^\star$ and $\boldsymbol{A}_{\varepsilon_k}$ minimizes Φ_{ε_k} on a closed ball of fixed radius $r > 0$ centred at $\boldsymbol{A}^\star$, say $B_r^\star$. A simple adaptation of the previous arguments yields that G_k restricted to $B_r^\star$ Γ-converges to G restricted to $B_r^\star$. As such, $\boldsymbol{A}^\star$ minimizes Φ_0 over $B_r^\star$ and is thus locally minimal for Φ_0.

6 Concluding Remarks

This work studied efficient computation of the quadratic EGW alignment problem between Euclidean spaces subject to non-asymptotic convergence guarantees. Despite the availability of various heuristic methods for computing EGW, formal guarantees beyond asymptotic convergence to a stationary point were absent until now. To develop our algorithms, we leveraged the variational form of the EGW distance that ties it to the well-understood EOT problem with a certain parametrized cost function. By analyzing the stability of the variational problem, its convexity and smoothness properties were established, which led to two new efficient algorithms for computing the EGW distance. The complexity of our algorithms agree with the best known complexity of $O(N^2)$ for computing the quadratic EGW distance directly, but unlike previous approaches, our methods are subject to non-asymptotic convergence rate guarantees to global/local solutions in the convex/non-convex regime. As the first derivative of the objective function depends on the solution to an EOT problem which must be solved numerically, we quantify both the error incurred by Sinkhorn's algorithm and the resulting effect on the convergence of both algorithms. Moreover, we establish a suitable notion of convergence of solutions to the variational EGW problem to those of the variational GW problem in the limit of vanishing regularization. Below, we discuss possible extensions and future research direction stemming from this work.

Algorithmic improvements. The stability analysis of the variational problem lays the groundwork for solving the EGW problem via smooth optimization methods. Consequently, improvements or alternatives to the proposed accelerated gradient methods are of great practical interest. For instance, marked improvements can be attained by analyzing the

tradeoff between the per iteration cost associated with updating the step size parameter and the resulting decrease in the number of iterations required for convergence. Similarly, establishing sharper bounds on the eigenvalues of the Hessian would improve our characterization of the smoothness and convexity properties of the objective.

Expanding duality theory. To the best of our knowledge, the current duality theory for the EGW problem is limited to the quadratic and inner product costs over Euclidean spaces. As the present work makes heavy use of this duality theory, we anticipate that these results could be extended to the EGW problem with other costs and/or spaces once an adequate duality theory has been established. Non-Euclidean spaces are not only of theoretical, but also of practical interest under the GW paradigm as they allow comparing/aligning important examples such as manifold or graph data.

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

D. Alvarez-Melis and T. Jaakkola. Gromov-wasserstein alignment of word embedding spaces. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1881–1890. Association for Computational Linguistics, 2018.

T. M. Apostol. *Mathematical analysis*. Addison-Wesley, 5 edition, 1974.

S. R. Becker, E. J. Candès, and M. C. Grant. Templates for convex cone problems with applications to sparse signal recovery. *Mathematical programming computation*, 3:165–218, 2011.

R. Beinert, C. Heiss, and G. Steidl. On assignment problems related to Gromov-Wasserstein distances on the real line. *arXiv preprint arXiv:2205.09006*, 2022.

G. Birkhoff. Extensions of Jentzsch's theorem. *Transactions of the American Mathematical Society*, 85(1):219–227, 1957.

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

J. F. Bonnans and A. Shapiro. *Perturbation analysis of optimization problems*. Springer Science & Business Media, 2013.

A. Braides. *Local minimization, variational evolution and* Γ*-convergence*, volume 2094. Springer, 2014.

H. Brézis. *Functional analysis, Sobolev spaces and partial differential equations*, volume 2. Springer, 2011.

C. Bunne, D. Alvarez-Melis, A. Krause, and S. Jegelka. Learning generative models across incomparable spaces. In *International conference on machine learning*, pages 851–861. PMLR, 2019.

G. Carlier and M. Laborde. A differential approach to the multi-marginal Schrödinger system. *SIAM Journal on Mathematical Analysis*, 52(1):709–717, 2020.

G. Carlier, V. Duval, G. Peyré, and B. Schmitzer. Convergence of entropic schemes for optimal transport and gradient flows. *SIAM Journal on Mathematical Analysis*, 49(2): 1385–1418, 2017.

J. Chen, B. T. Nguyen, and Y. S. Soh. Semidefinite relaxations of the gromov-wasserstein distance. *arXiv preprint arXiv:2312.14572*, 2023.

S. Chen, B. Shi, and Y. Yuan. Gradient norm minimization of Nesterov acceleration: $o(1/k^3)$. *arXiv preprint arXiv:2209.08862*, 2022.

F. H. Clarke. Generalized gradients and applications. *Transactions of the American Mathematical Society*, 205:247–262, 1975.

F. H. Clarke. *Optimization and nonsmooth analysis*. SIAM, 1990.

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

M. Cuturi. Sinkhorn distances: lightspeed computation of optimal transport. In *Proceedings of the 26th International Conference on Neural Information Processing Systems*, pages 2292–2300, 2013.

A. d'Aspremont. Smooth optimization with approximate gradient. *SIAM Journal on Optimization*, 19(3):1171–1183, 2008.

J. Delon, A. Desolneux, and A. Salmona. Gromov-Wasserstein distances between Gaussian distributions. *Journal of Applied Probability*, pages 1–21, 2022.

P. Demetci, R. Santorella, B. Sandstede, W. S. Noble, and R. Singh. SCOT: single-cell multi-omics alignment with optimal transport. *Journal of Computational Biology*, 29(1): 3–18, 2022.

O. Devolder, F. Glineur, and Y. Nesterov. First-order methods of smooth convex optimization with inexact oracle. *Mathematical Programming*, 146:37–75, 2014.

T. Dumont, T. Lacombe, and F.-X. Vialard. On the existence of Monge maps for the Gromov-Wasserstein distance. *arXiv preprint arXiv:2210.11945*, 2022.

P. Dvurechensky. Gradient method with inexact oracle for composite non-convex optimization. *arXiv preprint arXiv:1703.09180*, 2017.

P. Dvurechensky, A. Gasnikov, and A. Kroshnin. Computational optimal transport: complexity by accelerated gradient descent is better than by Sinkhorn's algorithm. In *International conference on machine learning*, pages 1367–1376. PMLR, 2018.

S. Eckstein and M. Nutz. Quantitative stability of regularized optimal transport and convergence of Sinkhorn's algorithm. *SIAM Journal on Mathematical Analysis*, 54(6): 5922–5948, 2022.

J. Feydy, T. Séjourné, F.-X. Vialard, S.-I. Amari, A. Trouvé, and G. Peyré. Interpolating between optimal transport and mmd using sinkhorn divergences. *arXiv preprint arXiv:1810.08278*, Oct. 2018.

R. Flamary, N. Courty, A. Gramfort, M. Z. Alaya, A. Boisbunon, S. Chambon, L. Chapel, A. Corenflos, K. Fatras, N. Fournier, L. Gautheron, N. T. Gayraud, H. Janati, A. Rakotomamonjy, I. Redko, A. Rolet, A. Schutz, V. Seguy, D. J. Sutherland, R. Tavenard, A. Tong, and T. Vayer. POT: Python optimal transport. *Journal of Machine Learning Research*, 22(78):1–8, 2021.

J. Franklin and J. Lorenz. On the scaling of multidimensional matrices. *Linear Algebra and its applications*, 114:717–735, 1989.

A. Genevay, G. Peyré, and M. Cuturi. Learning generative models with Sinkhorn divergences. In *International Conference on Artificial Intelligence and Statistics*, pages 1608–1617. PMLR, 2018.

A. Genevay, L. Chizat, F. Bach, M. Cuturi, and G. Peyré. Sample complexity of sinkhorn divergences. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*, pages 1574–1583, 2019.

S. Ghadimi and G. Lan. Accelerated gradient methods for nonconvex nonlinear and stochastic programming. *Mathematical Programming*, 156(1-2):59–99, 2016.

P. Ghosal, M. Nutz, and E. Bernton. Stability of entropic optimal transport and Schrödinger bridges. *Journal of Functional Analysis*, 283(9):109622, 2022.

J.-B. Hiriart-Urruty and C. Lemaréchal. *Fundamentals of convex analysis*. Springer Science & Business Media, 2004.

L. V. Kantorovich. On the translocation of masses. In *Doklady Akademii Nauk USSR*, volume 37, pages 199–201, 1942.

P. Koehl, M. Delarue, and H. Orland. Computing the Gromov-Wasserstein distance between two surface meshes using optimal transport. *Algorithms*, 16(3):131, 2023.

K. Le, D. Q. Le, H. Nguyen, D. Do, T. Pham, and N. Ho. Entropic Gromov-Wasserstein between Gaussian distributions. In *International Conference on Machine Learning*, pages 12164–12203. PMLR, 2022.

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

F. Mémoli. Gromov-Wasserstein distances and the metric approach to object matching. *Found. Comput. Math.*, 11(4):417–487, 2011.

Y. Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2003.

Y. Nesterov. Gradient methods for minimizing composite functions. *Mathematical programming*, 140(1):125–161, 2013.

M. Nutz. Introduction to entropic optimal transport. *Lecture notes, Columbia University*, 2021.

M. Nutz and J. Wiesel. Stability of Schrödinger potentials and convergence of Sinkhorn's algorithm. *The Annals of Probability*, 51(2):699–722, 2023.

G. Peyré, M. Cuturi, and J. Solomon. Gromov-Wasserstein averaging of kernel and distance matrices. In *International Conference on Machine Learning*, pages 2664–2672. PMLR, 2016.

R. T. Rockafellar. *Convex analysis*, volume 11. Princeton university press, 1997.

H. Samelson. On the Perron-Frobenius theorem. *Michigan Mathematical Journal*, 4(1):57 – 59, 1957.

F. Santambrogio. *Optimal Transport for Applied Mathematicians*. Birkhäuser, 2015.

M. Scetbon, G. Peyré, and M. Cuturi. Linear-time Gromov- Wasserstein distances using low rank couplings and costs. In *International Conference on Machine Learning*, pages 19347–19365. PMLR, 2022.

T. Séjourné, F.-X. Vialard, and G. Peyré. The unbalanced Gromov-Wasserstein distance: conic formulation and relaxation. *Advances in Neural Information Processing Systems*, 34: 8766–8779, 2021.

B. Shi, S. S. Du, M. I. Jordan, and W. J. Su. Understanding the acceleration phenomenon via high-resolution differential equations. *Mathematical Programming*, pages 1–70, 2021.

R. Sinkhorn. Diagonal equivalence to matrices with prescribed row and column sums. *The American Mathematical Monthly*, 74(4):402–405, 1967.

J. Solomon, G. Peyré, V. G. Kim, and S. Sra. Entropic metric alignment for correspondence problems. *ACM Transactions on Graphics (ToG)*, 35(4):1–13, 2016.

K.-T. Sturm. The space of spaces: curvature bounds and gradient flows on the space of metric measure spaces. *arXiv preprint arXiv:1208.0434*, 2012.

P. Tseng. On accelerated proximal gradient methods for convex-concave optimization. `https://www.mit.edu/~dimitrib/PTseng/papers/apgm.pdf`, 2008.

T. Vayer, L. Chapel, R. Flamary, R. Tavenard, and N. Courty. Optimal transport for structured data with application on graphs. In *International Conference on Machine Learning (ICML)*, 2019.

T. Vayer, R. Flamary, R. Tavenard, L. Chapel, and N. Courty. Sliced Gromov-Wasserstein. *arXiv preprint arXiv:1905.10124*, 2020.

C. Villani. *Optimal Transport: Old and New*. Springer, 2008.

C. Vincent-Cuaz, R. Flamary, M. Corneli, T. Vayer, and N. Courty. Semi-relaxed Gromov Wasserstein divergence with applications on graphs. In *International Conference on Learning Representations*, 2022.

H. Xiao, K. Rasul, and R. Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

H. Xu, D. Luo, and L. Carin. Scalable Gromov-Wasserstein learning for graph partitioning and matching. *Advances in neural information processing systems*, 32, 2019a.

H. Xu, D. Luo, H. Zha, and L. C. Duke. Gromov-Wasserstein learning for graph matching and node embedding. In *International conference on machine learning*, pages 6932–6941. PMLR, 2019b.

Y. Yan, W. Li, H. Wu, H. Min, M. Tan, and Q. Wu. Semi-supervised optimal transport for heterogeneous domain adaptation. In *IJCAI*, volume 7, pages 2969–2975, 2018.

K. Yosida. *Functional analysis*. Springer Science & Business Media, 1995.

Z. Zhang, Z. Goldfeld, Y. Mroueh, and B. K. Sriperumbudur. Gromov-Wasserstein distances: entropic regularization, duality, and sample complexity. *arXiv preprint arXiv:2212.12848*, 2022a.

Z. Zhang, Y. Mroueh, Z. Goldfeld, and B. Sriperumbudur. Cycle consistent probability divergences across different spaces. In *International Conference on Artificial Intelligence and Statistics*, pages 7257–7285. PMLR, 2022b.

Appendix A. Sharpness of variance bound from Corollary 5

Let $\mu_0 = \frac{1}{2}(\delta_0 + \delta_a)$ and $\mu_1 = \frac{1}{2}(\delta_0 + \delta_b)$ for $a \in \mathbb{R}^{d_0}$ and $b \in \mathbb{R}^{d_1}$. In this case, any coupling $\pi \in \Pi(\mu_0, \mu_1)$ is of the form $\pi_{00}\delta_{(0,0)} + \pi_{0b}\delta_{(0,b)} + \pi_{a0}\delta_{(a,0)} + \pi_{ab}\delta_{(a,b)}$ with the constraint that $\pi_{00} = \pi_{ab}$ and $\pi_{0b} = \pi_{a0} = \frac{1}{2} - \pi_{ab}$. For any $\boldsymbol{A} \in \mathcal{D}_M$, $\mathsf{OT}_{\boldsymbol{A},\varepsilon}(\mu_0, \mu_1)$ is given by

$$\begin{aligned}&\inf_{\pi \in \Pi(\mu_0,\mu_1)} \left\{ \int -4\|x\|^2\|y\|^2 - 32x^\intercal \boldsymbol{A} y \, d\pi(x,y) + \varepsilon \mathsf{D}_{\mathsf{KL}}(\pi \| \mu_0 \otimes \mu_1) \right\} \\ &= \inf_{\pi_{ab} \in (0,1/2)} \left\{ -\pi_{ab}(4\|a\|^2\|b\|^2 + 32a^\intercal \boldsymbol{A} b) + 2\varepsilon \pi_{ab} \log(4\pi_{ab}) + (1 - 2\pi_{ab})\, \varepsilon \log(2 - 4\pi_{ab}) \right\},\end{aligned}$$

the objective is a sum of convex functions and the first-order optimality condition reads

$$4\|a\|^2\|b\|^2 + 32a^\intercal \boldsymbol{A} b = 2\varepsilon \log(4\pi_{ab}) - 2\varepsilon \log(2 - 4\pi_{ab}) \iff \pi_{ab} = \frac{e^z}{2(1 + e^z)},$$

for $z = \left(2\|a\|^2\|b\|^2 + 16a^\intercal \boldsymbol{A} b\right)/\varepsilon$. Let $\pi^\star$ be the corresponding EOT coupling for $\mathsf{OT}_{\boldsymbol{A},\varepsilon}(\mu_0, \mu_1)$. For any $\boldsymbol{C} \in \mathbb{R}^{d_0 \times d_1}$,

$$\mathrm{Var}_{\pi^\star}[X^\intercal \boldsymbol{C} Y] = \pi^\star_{ab}(1 - \pi^\star_{ab})(a^\intercal \boldsymbol{C} b)^2 \leq \pi^\star_{ab}(1 - \pi^\star_{ab})\|\boldsymbol{C}\|_F^2 \|a\|^2\|b\|^2,$$

with equality for $\boldsymbol{C} = Cab^\intercal$ with $C \in \mathbb{R}$. Hence,

$$\sup_{\|\boldsymbol{C}\|_F = 1} \{\mathrm{Var}_{\pi^\star}[X^\intercal \boldsymbol{C} Y]\} = \pi^\star_{ab}(1 - \pi^\star_{ab})\|a\|^2\|b\|^2,$$

which can be made arbitrarily close to $\frac{1}{4}\|a\|^2\|b\|^2$ for fixed a, b by choosing $\boldsymbol{A} \in \mathcal{D}_M$ and $\varepsilon > 0$ as to make z sufficiently large. On the other hand, $\sqrt{M_4(\mu_0)M_4(\mu_1)} = \frac{1}{2}\|a\|^2\|b\|^2$, such that the variance bound in Corollary 5 is tight up to a constant factor.

Appendix B. Sinkhorn's Algorithm as an inexact oracle

Given $\mu_0 = \sum_{i=1}^{N_0} a_i \delta_{x^{(i)}} \in \mathcal{P}(\mathbb{R}^{d_0})$ and $\mu_1 = \sum_{j=1}^{N_1} b_j \delta_{y^{(j)}} \in \mathcal{P}(\mathbb{R}^{d_1})$, let a, b denote the corresponding (positive) probability vectors. Fix an underlying cost function $c : \mathbb{R}^{d_0} \times \mathbb{R}^{d_1} \to \mathbb{R}$ and $\varepsilon > 0$, and let $\boldsymbol{K} \in \mathbb{R}^{N_0 \times N_1}$ with $\boldsymbol{K}_{ij} = e^{-\frac{c\left(x^{(i)}, y^{(j)}\right)}{\varepsilon}}$. Consider the standard implementation of Sinkhorn's algorithm (cf. e.g. Cuturi 2013; Flamary et al. 2021).

In Algorithm 3 and the remainder of this section, the division of two vectors is understood componentwise. The stopping condition is based only on one of the marginal constraints, as $\boldsymbol{\Pi}^k \mathbb{1}_{N_1} = a$ by construction.

The following definitions enable describing the convergence properties of Algorithm 3; we follow the approach of Franklin and Lorenz (1989) with only minor modifications. Let $\mathbb{R}^d_+$ denote the set of vectors with positive entries and, for $x, y \in \mathbb{R}^d_+$, let

$$\mathsf{d}_H(x, y) = \log \max_{1 \leq i,j \leq d} \frac{x_i y_j}{y_i x_j},$$

denote Hilbert's projective metric[12] on $\mathbb{R}^d_+$. By definition,

$$\mathsf{d}_H(x, y) = \mathsf{d}_H(x/y, \mathbb{1}_d), \tag{25}$$

12. $\mathsf{d}_H(x, y) = 0$ if and only if $x = \alpha y$ for $\alpha > 0$, d_H is symmetric and satisfies the triangle inequality.

Algorithm 3 Sinkhorn Algorithm

Fix a threshold γ and a maximum iteration number $k_{\max}$.
$u_0 \leftarrow \mathbb{1}_{N_0}/N_0$
$k \leftarrow 1$
repeat
    $v_k \leftarrow b/(\boldsymbol{K}^\intercal u_{k-1})$
    $u_k \leftarrow a/(\boldsymbol{K} v_k)$
    $\boldsymbol{\Pi}^k \leftarrow \mathrm{diag}(u_k)\boldsymbol{K}\mathrm{diag}(v_k)$
    $k \leftarrow k+1$
until $\|(\boldsymbol{\Pi}^k)^\intercal \mathbb{1}_{N_0} - b\|_2 < \gamma$ or $k > k_{\max}$
return $\boldsymbol{\Pi}^k$

for any $x, y \in \mathbb{R}_+^d$ and, setting $x = e^w, y = e^z$ componentwise,

$$\begin{aligned}
\mathsf{d}_H(x,y) &= \log \max_{1\le i,j\le d} e^{w_i + z_j - w_j - z_i}, \\
&= \max_{1\le i,j\le d} w_i + z_j - w_j - z_i, \\
&= \max_{1\le i\le d} (\log x_i - \log y_i) - \min_{1\le i\le d} (\log x_i - \log y_i), \\
&= \max_{1\le i\le d} \log\left(\frac{x_i}{y_i}\right) - \min_{1\le i\le d} \log\left(\frac{x_i}{y_i}\right).
\end{aligned} \tag{26}$$

It was proved in Birkhoff (1957); Samelson (1957) that multiplication with a positive matrix is a strict contraction w.r.t. d_H. Precisely,

$$\mathsf{d}_H(\boldsymbol{A}x, \boldsymbol{A}y) \le \lambda(\boldsymbol{A})\mathsf{d}_H(x,y), \tag{27}$$

for any $\boldsymbol{A} \in \mathbb{R}_+^{d' \times d}$ and $x, y \in \mathbb{R}_+^d$, where

$$\lambda(\boldsymbol{A}) = \frac{\sqrt{\eta(\boldsymbol{A})} - 1}{\sqrt{\eta(\boldsymbol{A})} + 1} < 1, \quad \eta(\boldsymbol{A}) = \max_{\substack{1\le i,j\le d' \\ 1\le k,l\le d}} \frac{\boldsymbol{A}_{ik}\boldsymbol{A}_{jl}}{\boldsymbol{A}_{jk}\boldsymbol{A}_{il}}.$$

Of note is that $\lambda(\boldsymbol{A}) = \lambda(\boldsymbol{A}^\intercal)$. Let

$$E = \{\boldsymbol{A} \in \mathbb{R}_+^{N_0 \times N_1} : \boldsymbol{A} = \mathrm{diag}(x)\boldsymbol{K}\mathrm{diag}(y) \text{ for some } x \in \mathbb{R}_+^{N_0}, y \in \mathbb{R}_+^{N_1}\},$$

and observe that if $\boldsymbol{A}, \boldsymbol{B} \in E$, there exists $x_{\boldsymbol{A},\boldsymbol{B}} \in \mathbb{R}_+^{N_0}, y_{\boldsymbol{A},\boldsymbol{B}} \in \mathbb{R}_+^{N_1}$ for which $\boldsymbol{A} = \mathrm{diag}(x_{\boldsymbol{A},\boldsymbol{B}})\boldsymbol{B}\mathrm{diag}(y_{\boldsymbol{A},\boldsymbol{B}})$. In this setting, let $\mathsf{d} : E \times E \mapsto [0,\infty)$ be such that

$$\mathsf{d}(\boldsymbol{A}, \boldsymbol{B}) = \mathsf{d}_H(x_{\boldsymbol{A},\boldsymbol{B}}, \mathbb{1}_{N_0}) + \mathsf{d}_H(y_{\boldsymbol{A},\boldsymbol{B}}, \mathbb{1}_{N_1}),$$

then d is a metric on E. As the EOT coupling $\boldsymbol{\Pi}^\star$ satisfies

$$\frac{\boldsymbol{\Pi}^\star_{ij}}{a_i b_j} = e^{\frac{\varphi\left(x^{(i)}\right) + \psi\left(y^{(j)}\right) - c\left(x^{(i)}, y^{(j)}\right)}{\varepsilon}},$$

where (φ, ψ) is any pair of EOT potentials, $\mathbf{\Pi}^\star = \mathrm{diag}(u^\star)\boldsymbol{K}\mathrm{diag}(v^\star) \in E$ for

$$u_i^\star = a_i e^{\frac{\varphi\left(x^{(i)}\right)}{\varepsilon}}, \quad v_j^\star = b_j e^{\frac{\psi\left(y^{(j)}\right)}{\varepsilon}}.$$

Note that $u^\star = a/(\boldsymbol{K}v^\star)$ and $v^\star = b/(\boldsymbol{K}^\intercal u^\star)$.

In the sequel, we analyze the convergence of $\mathbf{\Pi}^k$ to $\mathbf{\Pi}^\star$ under d. The following result translates bounds on $\mathsf{d}(\mathbf{\Pi}^k, \mathbf{\Pi}^\star)$ to bounds on $\|\mathbf{\Pi}^k - \mathbf{\Pi}^\star\|_\infty$.

Lemma 26. *Fix $\delta > 0$. If $\mathsf{d}(\mathbf{\Pi}^k, \mathbf{\Pi}^\star) \leq \delta$, it follows that $\|\mathbf{\Pi}^k - \mathbf{\Pi}^\star\|_\infty \leq e^\delta - 1$.*

Proof By Lemma 3 in Franklin and Lorenz (1989), whenever $\mathsf{d}(\mathbf{\Pi}^k, \mathbf{\Pi}^\star) \leq \delta$ it holds that

$$e^{-\delta} - 1 \leq \frac{\mathbf{\Pi}^\star_{ij}}{\mathbf{\Pi}^k_{ij}} - 1 \leq e^\delta - 1,$$

for every $1 \leq i \leq N_0, 1 \leq j \leq N_1$. As such,

$$|\mathbf{\Pi}^\star_{ij} - \mathbf{\Pi}^k_{ij}| \leq \mathbf{\Pi}^k_{ij}\left((1 - e^{-\delta}) \vee (e^\delta - 1)\right) \leq (1 - e^{-\delta}) \vee (e^\delta - 1) = e^\delta - 1,$$

yielding $\|\mathbf{\Pi}^\star - \mathbf{\Pi}^k\|_\infty \leq e^\delta - 1$. ■

The number of iterations required to achieve $\mathsf{d}(\mathbf{\Pi}^k, \mathbf{\Pi}^\star) \leq \delta$ can be bounded as follows.

Proposition 27. *Let $\mathbf{\Pi}^k$ be given by Algorithm 3 and fix $\delta > 0$. Then, if*

$$k \geq 1 + \frac{1}{2\log\left(\lambda(\boldsymbol{K})\right)}\log\left(\frac{\delta(1 - \lambda(\boldsymbol{K}))}{\mathsf{d}_H((\mathbf{\Pi}^1)^\intercal \mathbb{1}_{N_0}, b)}\right),$$

it follows that $\mathsf{d}(\mathbf{\Pi}^k, \mathbf{\Pi}^\star) \leq \delta$.

The proof of Proposition 27 follows immediately from Lemma 29 ahead. We first prove an auxiliary lemma.

Lemma 28. *For $k \geq 1$, the iterates u_k, v_k of Algorithm 3 satisfy*

$$\mathsf{d}_H(u_{k+1}, u^\star) \leq \lambda(\boldsymbol{K})^2 \mathsf{d}_H(u_k, u^\star), \quad \mathsf{d}_H(v_{k+1}, v^\star) \leq \lambda(\boldsymbol{K})^2 \mathsf{d}_H(v_k, v^\star),$$
$$\mathsf{d}_H(u_k, u^\star) \leq \frac{\mathsf{d}_H(u_k, u_{k+1})}{1 - \lambda(\boldsymbol{K})^2}, \quad \mathsf{d}_H(v_k, v^\star) \leq \frac{\mathsf{d}_H(v_k, v_{k+1})}{1 - \lambda(\boldsymbol{K})^2}.$$

Proof To prove the first claim, we have by (27) that

$$\mathsf{d}_H(u_{k+1}, u^\star) = \mathsf{d}_H\left(a/(\boldsymbol{K}v_{k+1}), a/(\boldsymbol{K}v^\star)\right) = \mathsf{d}_H(\boldsymbol{K}v^\star, \boldsymbol{K}v_{k+1}) \leq \lambda(\boldsymbol{K})\mathsf{d}_H(v_{k+1}, v^\star).$$

It follows similarly that $\mathsf{d}_H(v_{k+1}, v^\star) \leq \lambda(\boldsymbol{K})\mathsf{d}_H(u_k, u^\star)$. Combining these bounds,

$$\mathsf{d}_H(u_{k+1}, u^\star) \leq \lambda(\boldsymbol{K})^2 \mathsf{d}_H(u_k, u^\star), \quad \mathsf{d}_H(v_{k+1}, v^\star) \leq \lambda(\boldsymbol{K})^2 \mathsf{d}_H(v_k, v^\star),$$

which proves the first claim. Applying the triangle inequality yields

$$\mathsf{d}_H(u_k, u^\star) \leq \mathsf{d}_H(u_{k+1}, u^\star) + \mathsf{d}_H(u_k, u_{k+1}) \leq \lambda(\boldsymbol{K})^2 \mathsf{d}_H(u_k, u^\star) + \mathsf{d}_H(u_k, u_{k+1}),$$

such that $(1-\lambda(\boldsymbol{K})^2)\mathsf{d}_H(u_k,u^\star) \le \mathsf{d}_H(u_k,u_{k+1})$ which proves the second claim; the same argument holds for the iterates v_k. ■

Lemma 29 translates the bound from Lemma 28 to a bound on $\mathsf{d}(\Pi^k,\Pi^\star)$. We introduce the notation $\odot$ to denote the componentwise product of vectors.

Lemma 29. *For $k \ge 2$, $\mathsf{d}(\boldsymbol{\Pi}^k,\boldsymbol{\Pi}^\star) \le \frac{\lambda(\boldsymbol{K})^{2(k-1)}}{1-\lambda(\boldsymbol{K})}\mathsf{d}_H((\boldsymbol{\Pi}^1)^\intercal \mathbb{1}_{N_0}, b)$.*

Proof As $\boldsymbol{\Pi}^k = \mathrm{diag}(u_k)\boldsymbol{K}\mathrm{diag}(v_k)$ and $\boldsymbol{\Pi}^\star = \mathrm{diag}(u^\star)\boldsymbol{K}\mathrm{diag}(v^\star)$,

$$\begin{aligned}\mathsf{d}(\boldsymbol{\Pi}^k,\boldsymbol{\Pi}^\star) &= \mathsf{d}_H(u_k,u^\star) + \mathsf{d}_H(v_k,v^\star)\\ &\le \lambda(\boldsymbol{K})^{2(k-1)}(1+\lambda(K))d_H(v_1,v^\star)\\ &\le \frac{\lambda(\boldsymbol{K})^{2(k-1)}}{1-\lambda(\boldsymbol{K})}d_H(v_1,v_2),\end{aligned}$$

where both inequalities follow from Lemma 28 and its proof. Finally,

$$\mathsf{d}_H(v_1,v_2) = \mathsf{d}_H\left(v_1, \frac{b}{\boldsymbol{K}^\intercal u_1}\right) = \mathsf{d}_H\left(v_1 \odot \boldsymbol{K}^\intercal u_1, b\right) = \mathsf{d}_H((\boldsymbol{\Pi}^1)^\intercal \mathbb{1}_{N_0}, b).$$

■

Now, we demonstrate why the termination condition based on the 2-norm endows us with a δ'-oracle approximation and provide a bound on the number of iterations required to achieve it. Theorem 1 in Dvurechensky et al. (2018) proves that there exists $\bar{k} \le 1 + \frac{R}{\gamma}$ satisfying

$$\|u_{\bar{k}} \odot \boldsymbol{K} v_{\bar{k}+1} - a\|_1 + \|(\boldsymbol{\Pi}^k)^\intercal \mathbb{1}_{N_0} - b\|_1 \le \gamma,$$

for $R = -2\log\left(e^{-\|\boldsymbol{C}\|_\infty/\varepsilon}\min_{\substack{1\le i\le N_0\\ 1\le j\le N_1}} a_i \wedge b_j\right)$. This gives a bound on the maximal number of iterations to achieve the 2-norm termination condition via the standard inequality $\|\cdot\|_2 \le \|\cdot\|_1$. We clarify that the analysis in Dvurechensky et al. (2018) is for a slightly different implementation of Sinkhorn's algorithm. First, running Algorithm 3 is tantamount to running their algorithm with reversed marginals. Next, one iteration of Algorithm 3 corresponds to two iterations in their analysis. Finally, their approach uses $\mathbb{1}_{N_0}$ rather than $\mathbb{1}_{N_0}/N_0$ for the initialization. This difference is innocuous for achieving the termination condition, as the iterates are identical up to multiplying v_k by N_0 and dividing u_k by N_0; $\boldsymbol{\Pi}^k$ is invariant this operation.

We now bound d_H in terms of the Euclidean distance with the aim of controlling $\mathsf{d}(\boldsymbol{\Pi}^k,\boldsymbol{\Pi}^\star)$ by $\|(\boldsymbol{\Pi}^k)^\intercal \mathbb{1}_{N_0} - b\|_2$.

Lemma 30. *Let $r,s \in \mathbb{R}^d_+$ be arbitrary, then*

$$\mathsf{d}_H(s,r) \le (r_{i^\star}^{-1} + s_{i_\star}^{-1})\|r-s\|_2,$$

where $i^* \in \mathrm{argmax}_{1\le i\le d}\frac{s_i}{r_i}$ and $i_* \in \mathrm{argmin}_{1\le i\le d}\frac{s_i}{r_i}$.

Proof We have by (26) that

$$\mathsf{d}_H(s,r) = \max_{1\le i\le N_0} \log\left(\frac{s_i}{r_i}\right) - \min_{1\le i\le N_0} \log\left(\frac{s_i}{r_i}\right).$$

Observe that $1 - \frac{r_i}{s_i} \le \log\left(\frac{s_i}{r_i}\right) \le \frac{s_i}{r_i} - 1$, as follows the inequalities $\frac{x}{1+x} \le \log(1+x) \le x$ for $x > -1$. Whence,

$$\begin{aligned}\mathsf{d}_H(s,r) &\le r_{i^\star}^{-1}\left(s_{i^\star} - r_{i^\star}\right) - s_{i_\star}^{-1}\left(s_{i_\star} - r_{i_\star}\right)\\ &\le \left(r_{i^\star}^{-1} + s_{i_\star}^{-1}\right)\|s-r\|_2.\end{aligned}$$

■

Note that the bound in Lemma 30 is symmetric in the sense that interchanging s and r does not modify the constant. This can be seen by noting that $\operatorname{argmax}_{1\le i\le d}\frac{s_i}{r_i} = \operatorname{argmin}_{1\le i\le d}\frac{r_i}{s_i}$ and $\operatorname{argmin}_{1\le i\le d}\frac{s_i}{r_i} = \operatorname{argmax}_{1\le i\le d}\frac{r_i}{s_i}$. By combining Lemmas 29 and 30 we arrive at the desired result.

Proposition 31. *Let $\underline{b} = \min_{1\le i\le N_1} b_i$ and set $0 < \gamma < \underline{b}$. Then, for $k \ge 1$,*

$$\mathsf{d}(\boldsymbol{\Pi}^k, \boldsymbol{\Pi}^\star) \le \frac{\left((\boldsymbol{\Pi}^k)^\intercal \mathbb{1}_{N_0}\right)_{i_\star}^{-1} + b_{i^\star}^{-1}}{1-\lambda(\boldsymbol{K})}\|(\boldsymbol{\Pi}^k)^\intercal\mathbb{1}_{N_0} - b\|$$

where $i^* \in \operatorname{argmax}_{1\le i\le N_1}\frac{\left((\boldsymbol{\Pi}^k)^\intercal\mathbb{1}_{N_0}\right)_i}{b_i}$ and $i_* \in \operatorname{argmin}_{1\le i\le N_1}\frac{\left((\boldsymbol{\Pi}^1)^\intercal\mathbb{1}_{N_0}\right)_i}{b_i}$. Further, there exists $\bar{k} \le 1 + \frac{R}{\gamma}$ for which $\|\boldsymbol{\Pi}^{\bar{k}\intercal}\mathbb{1}_{N_0} - b\| \le \gamma$ and $\mathsf{d}(\boldsymbol{\Pi}^{\bar{k}}, \boldsymbol{\Pi}^\star) \le \frac{(\underline{b}-\gamma)^{-1}+\underline{b}^{-1}}{1-\lambda(\boldsymbol{K})}\gamma$. In particular, setting

$$\gamma = \bar{\alpha}\underline{b} \text{ for } \bar{\alpha} = \frac{\delta(1-\lambda(\boldsymbol{K})) + 2 - \sqrt{\delta^2(1-\lambda(\boldsymbol{K}))^2 + 4}}{2} \in (0,1),$$

it holds that $\mathsf{d}(\boldsymbol{\Pi}^k, \boldsymbol{\Pi}^\star) = \delta$.

Proof From the proof of Lemma 29,

$$\mathsf{d}(\boldsymbol{\Pi}^k, \boldsymbol{\Pi}^\star) \le (1+\lambda(\boldsymbol{K}))\mathsf{d}_H(v_k, v^\star) \le \frac{\mathsf{d}_H(v_k, v_{k+1})}{1-\lambda(\boldsymbol{K})} = \frac{\mathsf{d}_H((\boldsymbol{\Pi}^k)^\intercal\mathbb{1}_{N_0}, b)}{1-\lambda(\boldsymbol{K})},$$

where the final inequality and equality stem from Lemma 28 and its proof. Applying Lemma 30 proves the first claim.

As for the second claim, it is clear from the discussion preceding Lemma 30 that $\|(\boldsymbol{\Pi}^{\bar{k}})^\intercal\mathbb{1}_{N_0} - b\| \le \gamma$ for some $\bar{k} \le 1 + \frac{R}{\gamma}$. Now, let $w = (\boldsymbol{\Pi}^{\bar{k}})^\intercal\mathbb{1}_{N_0}$ and observe that $\|w-b\|_\infty \le \|w-b\| \le \gamma$ such that $b_i - \gamma \le w_i \le b_i + \gamma$ for $i = 1,\dots,N_1$. Hence $w_i^{-1} \le (b_i-\gamma)^{-1} \le (\underline{b}-\gamma)^{-1}$ as $\gamma < \underline{b}$. Applying this bound to the previous inequality proves the claim.

For the final claim, observe that $\bar{\alpha}$ solves the equation $\frac{2\alpha-\alpha^2}{1-\alpha} = \delta(1-\lambda(\boldsymbol{K}))$ for $\alpha \in (0,1)$ (indeed, $\delta(1-\lambda(\boldsymbol{K})) < \sqrt{\delta^2(1-\lambda(\boldsymbol{K}))^2+4} < \delta(1-\lambda(\boldsymbol{K}))+2$). Setting $\gamma = \bar{\alpha}\underline{b}$ in the previously derived bound on $\mathsf{d}(\boldsymbol{\Pi}^k, \boldsymbol{\Pi}^\star)$ gives

$$\mathsf{d}(\boldsymbol{\Pi}^k, \boldsymbol{\Pi}^\star) \le \frac{(1-\bar{\alpha})^{-1}+1}{1-\lambda(\boldsymbol{K})}\bar{\alpha} = \frac{2\bar{\alpha}-\bar{\alpha}^2}{(1-\bar{\alpha})(1-\lambda(\boldsymbol{K}))} = \delta.$$

■

The proof of Proposition 8 follows by combining Propositions 27 and 31; the maximal number of iterations for Algorithm 3 to output a matrix $\tilde{\mathbf{\Pi}}$ satisfying $\mathsf{d}(\tilde{\mathbf{\Pi}}, \mathbf{\Pi}^\star) \leq \delta$ is

$$\tilde{k} = \min \Bigg\{ 1 + \frac{1}{2\log\left(\lambda(\boldsymbol{K})\right)} \log\left(\frac{\delta(1-\lambda(\boldsymbol{K}))}{\mathsf{d}_H((\mathbf{\Pi}^1)^\intercal \mathbb{1}_{N_0}, b)}\right), \\ 1 - \frac{4\underline{b}^{-1}}{\delta(1-\lambda(\boldsymbol{K})) + 2 - \sqrt{\delta^2(1-\lambda(\boldsymbol{K}))^2 + 4}} \log\left(e^{-\|\boldsymbol{C}\|_\infty/\varepsilon} \min_{\substack{1\leq i\leq N_0 \\ 1\leq j\leq N_1}} a_i \wedge b_j\right) \Bigg\}, \tag{28}$$

which corresponds to an $(e^\delta - 1)$-oracle for the entropic transport plan in light of Lemma 26.

Appendix C. Convergence of Algorithm 2

In what follows, we slightly adapt the proof of Theorem 2 in Ghadimi and Lan (2016) to conform to the inexact setting. We first clarify that they treat the composite problem

$$\inf_{x\in\mathbb{R}^d} f(x) + g(x) + \mathcal{Q}(x),$$

where f is L'-smooth and non-convex, g is L''-smooth and convex, and $\mathcal{Q}$ is non-smooth and convex with a bounded domain. Hence $f+g$ is $L = L' + L''$ smooth and possibly non-convex.

Our problem conforms to this setting (up to vectorization) with $f = \mathsf{OT}_{(\cdot),\varepsilon}(\mu_0, \mu_1)$, $g = 32\|\cdot\|_F^2$, and $\mathcal{Q} = \mathcal{I}_{\mathcal{D}_M}$, the indicator function of the set $\mathcal{D}_M$, defined by

$$\mathcal{I}_{\mathcal{D}_M}(\boldsymbol{A}) = \begin{cases} 0, & \text{if } \boldsymbol{A} \in \mathcal{D}_M, \\ +\infty, & \text{otherwise.} \end{cases}$$

When Φ is convex, we set $f = 0$ and $g = \Phi$ hence $L' = 0$, $L = L''$.

As Φ is L-smooth, by Lemma 5 in Ghadimi and Lan (2016),

$$\Phi(\boldsymbol{B}_k) \leq \Phi(\boldsymbol{A}_k) + \operatorname{tr}\left(D\Phi_{[\boldsymbol{A}_k]}^\intercal(\boldsymbol{B}_k - \boldsymbol{A}_k)\right) + \frac{L}{2}\|\boldsymbol{B}_k - \boldsymbol{A}_k\|_F^2, \tag{29}$$

and for any $\boldsymbol{H} \in \mathbb{R}^{d_0\times d_1}$, letting L' denote the Lipschitz constant of $\mathsf{OT}_{(\cdot),\varepsilon}(\mu_0,\mu_1)$, the same result yields

$$\begin{aligned} &\Phi(\boldsymbol{A}_k) - ((1-\tau_k)\Phi(\boldsymbol{B}_{k-1}) + \tau_k\Phi(\boldsymbol{H})) \\ &= \tau_k\left(\Phi(\boldsymbol{A}_k) - \Phi(\boldsymbol{H})\right) + (1-\tau_k)\left(\Phi(\boldsymbol{A}_k) - \Phi(\boldsymbol{B}_{k-1})\right) \\ &\leq \tau_k\left(\operatorname{tr}\left(D\Phi_{[\boldsymbol{A}_k]}^\intercal(\boldsymbol{A}_k - \boldsymbol{H})\right) + \frac{L'}{2}\|\boldsymbol{H} - \boldsymbol{A}_k\|_F^2\right) \\ &+ (1-\tau_k)\left(\operatorname{tr}\left(D\Phi_{[\boldsymbol{A}_k]}^\intercal(\boldsymbol{A}_k - \boldsymbol{B}_{k-1})\right) + \frac{L'}{2}\|\boldsymbol{B}_{k-1} - \boldsymbol{A}_k\|_F^2\right) \\ &= \operatorname{tr}\left(D\Phi_{[\boldsymbol{A}_k]}^\intercal(\boldsymbol{A}_k - \tau_k\boldsymbol{H} - (1-\tau_k)\boldsymbol{B}_{k-1})\right) \\ &+ \frac{L'\tau_k}{2}\|\boldsymbol{H} - \boldsymbol{A}_k\|_F^2 + \frac{L'(1-\tau_k)}{2}\underbrace{\|\boldsymbol{B}_k - \boldsymbol{A}_k\|_F^2}_{\tau_k^2\|\boldsymbol{B}_{k-1} - \boldsymbol{C}_{k-1}\|_F^2}, \end{aligned} \tag{30}$$

recalling the update $\boldsymbol{A}_k = \tau_k \boldsymbol{C}_{k-1} + (1-\tau_k)\boldsymbol{B}_{k-1}$.

Denote the subdifferential of $\mathcal{I}_{\mathcal{D}_M}$ at $\boldsymbol{A} \in \mathbb{R}^{d_0 \times d_1}$ by

$$\partial \mathcal{I}_{\mathcal{D}_M}(\boldsymbol{A}) \coloneqq \left\{ \boldsymbol{P} \in \mathbb{R}^{d_0 \times d_1} : \mathcal{I}_{\mathcal{D}_M}(\boldsymbol{X}) - \mathcal{I}_{\mathcal{D}_M}(\boldsymbol{A}) \geq \operatorname{tr}\left(\boldsymbol{P}^\intercal(\boldsymbol{X}-\boldsymbol{A})\right), \text{ for every } \boldsymbol{X} \in \mathbb{R}^{d_0 \times d_1} \right\}.$$

As $\boldsymbol{C}_k$ is optimal for the problem $\operatorname{argmin}_{\boldsymbol{V} \in \mathbb{R}^{d_0 \times d_1}} \left\{ \frac{1}{2\gamma_k} \| \boldsymbol{V} - (\boldsymbol{C}_{k-1} - \gamma_k \boldsymbol{G}_k) \|_F^2 + \mathcal{I}_{\mathcal{D}_M}(\boldsymbol{V}) \right\}$, there exists $\boldsymbol{P} \in \partial \mathcal{I}_{\mathcal{D}_M}(\boldsymbol{C}_k)$ for which $\boldsymbol{G}_k + \boldsymbol{P} + \frac{1}{\gamma_k}(\boldsymbol{C}_k - \boldsymbol{C}_{k+1}) = 0$ (see Theorem 23.8, Theorem 25.1, and p. 264 in Rockafellar 1997). Thus, for any $\boldsymbol{U} \in \mathbb{R}^{d_0 \times d_1}$,

$$\begin{aligned} \operatorname{tr}\left((\boldsymbol{G}_k + \boldsymbol{P})^\intercal(\boldsymbol{C}_k - \boldsymbol{U})\right) &= \frac{1}{\gamma_k} \operatorname{tr}\left((\boldsymbol{C}_k - \boldsymbol{C}_{k-1})^\intercal(\boldsymbol{U} - \boldsymbol{C}_k)\right) \\ &= \frac{1}{2\gamma_k}\left(\|\boldsymbol{C}_{k-1} - \boldsymbol{U}\|_F^2 - \|\boldsymbol{C}_k - \boldsymbol{U}\|_F^2 - \|\boldsymbol{C}_k - \boldsymbol{C}_{k-1}\|_F^2 \right), \end{aligned}$$

where the final line follows from some simple algebra. As $\boldsymbol{P} \in \partial \mathcal{I}_{\mathcal{D}_M}(\boldsymbol{C}_k)$, $\operatorname{tr}(\boldsymbol{P}^\intercal(\boldsymbol{C}_k - \boldsymbol{U})) \geq \mathcal{I}_{\mathcal{D}_M}(\boldsymbol{C}_k) - \mathcal{I}_{\mathcal{D}_M}(\boldsymbol{U}) = -\mathcal{I}_{\mathcal{D}_M}(\boldsymbol{U})$, whence

$$\operatorname{tr}\left(\boldsymbol{G}_k^\intercal(\boldsymbol{C}_k - \boldsymbol{U})\right) \leq \mathcal{I}_{\mathcal{D}_M}(\boldsymbol{U}) + \frac{1}{2\gamma_k}\left(\|\boldsymbol{C}_{k-1} - \boldsymbol{U}\|_F^2 - \|\boldsymbol{C}_k - \boldsymbol{U}\|_F^2 - \|\boldsymbol{C}_k - \boldsymbol{C}_{k-1}\|_F^2 \right). \quad (31)$$

By the same steps applied to the other subproblem with $\boldsymbol{B}_k$ and $\boldsymbol{A}_k$ taking the place of $\boldsymbol{C}_k$ and $\boldsymbol{C}_{k-1}$ respectively,

$$\operatorname{tr}\left(\boldsymbol{G}_k^\intercal(\boldsymbol{B}_k - \boldsymbol{U})\right) \leq \mathcal{I}_{\mathcal{D}_M}(\boldsymbol{U}) + \frac{1}{2\beta_k}\left(\|\boldsymbol{A}_k - \boldsymbol{U}\|_F^2 - \|\boldsymbol{B}_k - \boldsymbol{U}\|_F^2 - \|\boldsymbol{B}_k - \boldsymbol{A}_k\|_F^2 \right).$$

Setting $\boldsymbol{U} = \tau_k \boldsymbol{C}_k + (1-\tau_k)\boldsymbol{B}_{k-1} \in \mathcal{D}_M$ (by convexity) in the previous display, bounding $-\|\boldsymbol{B}_k - \boldsymbol{U}\|_F^2$ above by 0, and recalling that $\boldsymbol{A}_k = \tau_k \boldsymbol{C}_{k-1} + (1-\tau_k)\boldsymbol{B}_{k-1}$ such that $\boldsymbol{A}_k - \boldsymbol{U} = \tau_k(\boldsymbol{C}_{k-1} - \boldsymbol{C}_k)$,

$$\operatorname{tr}\left(\boldsymbol{G}_k^\intercal(\boldsymbol{B}_k - \tau_k \boldsymbol{C}_k + (1-\tau_k)\boldsymbol{B}_{k-1})\right) \leq \frac{1}{2\beta_k}\left(\tau_k^2 \|\boldsymbol{C}_k - \boldsymbol{C}_{k-1}\|_F^2 - \|\boldsymbol{B}_k - \boldsymbol{A}_k\|_F^2 \right).$$

Combining with (31) upon scaling by τ_k,

$$\begin{aligned} \operatorname{tr}\left(\boldsymbol{G}_k^\intercal(\boldsymbol{B}_k - \tau_k \boldsymbol{U} + (1-\tau_k)\boldsymbol{B}_{k-1})\right) \leq \tau_k \mathcal{I}_{\mathcal{D}_M}(\boldsymbol{U}) &+ \frac{1}{2\beta_k}\left(\tau_k^2 \|\boldsymbol{C}_k - \boldsymbol{C}_{k-1}\|_F^2 - \|\boldsymbol{B}_k - \boldsymbol{A}_k\|_F^2 \right), \\ &+ \frac{\tau_k}{2\gamma_k}\left(\|\boldsymbol{C}_{k-1} - \boldsymbol{U}\|_F^2 - \|\boldsymbol{C}_k - \boldsymbol{U}\|_F^2 - \|\boldsymbol{C}_k - \boldsymbol{C}_{k-1}\|_F^2 \right), \end{aligned}$$

by the choice of τ_k, β_k, γ_k, we have that $\frac{\tau_k^2}{\beta_k} - \frac{\tau_k}{\gamma_k} \leq 0$ such that

$$\begin{aligned} \operatorname{tr}\left(\boldsymbol{G}_k^\intercal(\boldsymbol{B}_k - \tau_k \boldsymbol{U} + (1-\tau_k)\boldsymbol{B}_{k-1})\right) \leq \tau_k \mathcal{I}_{\mathcal{D}_M}(\boldsymbol{U}) &+ \frac{\tau_k}{2\gamma_k}\left(\|\boldsymbol{C}_{k-1} - \boldsymbol{U}\|_F^2 - \|\boldsymbol{C}_k - \boldsymbol{U}\|_F^2 \right) \\ &- \frac{1}{2\beta_k}\|\boldsymbol{B}_k - \boldsymbol{A}_k\|_F^2. \end{aligned}$$

Combining the equation above with (29) and (30) and setting $\boldsymbol{H} = \boldsymbol{U} \in \mathcal{D}_M$ (otherwise the bound is vacuous),

$$\begin{aligned}\Phi(\boldsymbol{B}_k) - \Phi(\boldsymbol{H}) &\leq (1-\tau_k)\left(\Phi(\boldsymbol{B}_{k-1}) - \Phi(\boldsymbol{H})\right) + \operatorname{tr}\left(D\Phi_{[\boldsymbol{A}_k]}^{\intercal}(\boldsymbol{B}_k - \tau_k\boldsymbol{H} - (1-\tau_k)\boldsymbol{B}_{k-1})\right) \\ &\quad + \frac{L'\tau_k}{2}\left\|\boldsymbol{H} - \boldsymbol{A}_k\right\|_F^2 + \frac{L'(1-\tau_k)}{2}\tau_k^2\left\|\boldsymbol{B}_{k-1} - \boldsymbol{C}_{k-1}\right\|_F^2 + \frac{L}{2}\left\|\boldsymbol{B}_k - \boldsymbol{A}_k\right\|_F^2 \\ &\leq (1-\tau_k)\left(\Phi(\boldsymbol{B}_{k-1}) - \Phi(\boldsymbol{H})\right) + \delta' + \frac{\tau_k}{2\gamma_k}\left(\left\|\boldsymbol{C}_{k-1} - \boldsymbol{H}\right\|_F^2 - \left\|\boldsymbol{C}_k - \boldsymbol{H}\right\|_F^2\right) \\ &\quad + \frac{L'\tau_k}{2}\left\|\boldsymbol{H} - \boldsymbol{A}_k\right\|_F^2 + \frac{L'(1-\tau_k)}{2}\tau_k^2\left\|\boldsymbol{B}_{k-1} - \boldsymbol{C}_{k-1}\right\|_F^2 + \left(\frac{L}{2} - \frac{1}{2\beta_k}\right)\left\|\boldsymbol{B}_k - \boldsymbol{A}_k\right\|_F^2,\end{aligned}$$

where the inequality follows from the δ-oracle which implies the bound (cf. (16))

$$\sup_{\boldsymbol{Y},\boldsymbol{Z}\in\mathcal{D}_M}\left\{\left|\operatorname{tr}\left(\boldsymbol{G}_k - D\Phi_{[\boldsymbol{A}_k]}\right)^{\intercal}(\boldsymbol{Y} - \boldsymbol{Z})\right)\right|\right\} \leq \delta',$$

observing that $\boldsymbol{B}_k, \tau_k\boldsymbol{H} + (1-\tau_k)\boldsymbol{B}_{k-1} \in \mathcal{D}_M$ by convexity ($\tau_k \in (0,1]$).

Applying Lemma 1 in Ghadimi and Lan (2016) yields, for $A_i = \frac{2}{i(i+1)}$,

$$\begin{aligned}\frac{\Phi(\boldsymbol{B}_k) - \Phi(\boldsymbol{H})}{A_k} &\leq \sum_{i=1}^{k} A_i^{-1}\left(\delta' + \frac{\tau_i}{2\gamma_i}\left(\left\|\boldsymbol{C}_{i-1} - \boldsymbol{H}\right\|_F^2 - \left\|\boldsymbol{C}_i - \boldsymbol{H}\right\|_F^2\right) + \frac{L'\tau_i}{2}\left\|\boldsymbol{H} - \boldsymbol{A}_i\right\|^2\right. \\ &\quad \left. + \frac{L'(1-\tau_i)}{2}\tau_i^2\left\|\boldsymbol{B}_{i-1} - \boldsymbol{C}_{i-1}\right\|_F^2 + \left(\frac{L}{2} - \frac{1}{2\beta_i}\right)\left\|\boldsymbol{B}_i - \boldsymbol{A}_i\right\|^2\right) \\ &\leq \frac{\left\|\boldsymbol{C}_0 - \boldsymbol{H}\right\|_F^2}{2\gamma_1} + \sum_{i=1}^{k} A_i^{-1}\left(\delta' + \frac{L'\tau_i}{2}\left\|\boldsymbol{H} - \boldsymbol{A}_i\right\|^2\right. \\ &\quad \left. + \frac{L'(1-\tau_i)}{2}\tau_i^2\left\|\boldsymbol{B}_{i-1} - \boldsymbol{C}_{i-1}\right\|_F^2 + \left(\frac{L}{2} - \frac{1}{2\beta_i}\right)\left\|\boldsymbol{B}_i - \boldsymbol{A}_i\right\|^2\right).\end{aligned}$$

By convexity of $\|\cdot\|_F^2$,

$$\begin{aligned}&\left\|\boldsymbol{H} - \boldsymbol{A}_i\right\|_F^2 + \tau_i(1-\tau_i)\left\|\boldsymbol{B}_{i-1} - \boldsymbol{C}_{i-1}\right\|_F^2 \\ &\leq 2\left(\left\|\boldsymbol{H}\right\|_F^2 + \left\|\boldsymbol{A}_i\right\|_F^2 + \tau_i(1-\tau_i)\left(\left\|\boldsymbol{B}_{i-1}\right\|_F^2 + \left\|\boldsymbol{C}_{i-1}\right\|_F^2\right)\right) \\ &\leq 2\left(\left\|\boldsymbol{H}\right\|_F^2 + (1-\tau_i)\left\|\boldsymbol{B}_{i-1}\right\|_F^2 + \tau_i\left\|\boldsymbol{C}_{i-1}\right\|_F^2 + \tau_i(1-\tau_i)\left(\left\|\boldsymbol{B}_{i-1}\right\|_F^2 + \left\|\boldsymbol{C}_{i-1}\right\|_F^2\right)\right) \\ &\leq 2\left(\left\|\boldsymbol{H}\right\|_F^2 + (1 + \tau_i(1-\tau_i))\max_{\mathcal{D}_M}\|\cdot\|_F^2\right) \\ &\leq 2\left(\left\|\boldsymbol{H}\right\|_F^2 + \frac{5}{16}M^2\right),\end{aligned}$$

observing that $\tau_i \in (0,1]$ hence $\tau_i(1-\tau_i) \leq \frac{1}{4}$. Thus, for $\boldsymbol{H} = \boldsymbol{B}^\star$, a global minimizer of Φ,

$$\begin{aligned}\frac{\Phi(\boldsymbol{B}_k) - \Phi(\boldsymbol{B}^\star)}{A_k} + \sum_{i=1}^{k}\frac{1 - L\beta_i}{2A_i\beta_i}\left\|\boldsymbol{B}_i - \boldsymbol{A}_i\right\|_F^2 &\leq \frac{\left\|\boldsymbol{C}_0 - \boldsymbol{B}^\star\right\|_F^2}{2\gamma_1} \\ &\quad + \sum_{i=1}^{k} A_i^{-1}\left(\delta' + L'\tau_i\left(\left\|\boldsymbol{B}^\star\right\|_F^2 + \frac{5}{16}M^2\right)\right).\end{aligned}$$

By construction, $\sum_{i=1}^k A_i^{-1} L' \tau_i = \frac{L'}{A_k}$, and $\Phi(\boldsymbol{B}_k) - \Phi(\boldsymbol{B}^\star) \geq 0$. It follows that

$$\begin{aligned}
&\min_{i=1}^k \left\| \beta_i^{-1} (\boldsymbol{B}_i - \boldsymbol{A}_i) \right\|_F^2 \\
&\leq 2 \left(\sum_{i=1}^k \frac{\beta_i (1 - L\beta_i)}{A_i} \right)^{-1} \left(\frac{\|\boldsymbol{C}_0 - \boldsymbol{B}^\star\|_F^2}{2\gamma_1} + \sum_{i=1}^k A_i^{-1} \delta' + \frac{L'}{A_k} \left(\|\boldsymbol{B}^\star\|_F^2 + \frac{5}{16} M^2 d_0^2 d_1^2 \right) \right).
\end{aligned}$$

As $\beta_i = \frac{L}{2}$, $\gamma_1 = \frac{1}{4L}$, and $A_i = \frac{2}{i(i+1)}$, $\sum_{i=1}^k \frac{\beta_i(1-L\beta_i)}{A_i} = \frac{1}{4L} \sum_{i=1}^k A_i^{-1} = \frac{k(k+1)(k+2)}{24L}$, so

$$\min_{i=1}^k \left\| \beta_i^{-1} (\boldsymbol{B}_i - \boldsymbol{A}_i) \right\|_F^2 \leq \frac{96L^2}{k(k+1)(k+2)} \|\boldsymbol{C}_0 - \boldsymbol{B}^\star\|_F^2 + 8L\delta' + \frac{24LL'}{N} \left(\|\boldsymbol{B}^\star\|_F^2 + \frac{5M^2}{16} \right).$$

This proves the claimed result in the non-convex setting.

In the convex regime, recall from the prior discussion that we may set $L' = 0$ in the previous display, proving the claim.

Appendix D. Additional Results

D.1 Proof of Lemma 17

The proof of Lemma 17 follows from the following lemma coupled with the chain rule for Fréchet differentiable maps.

Lemma 32. *Let $\mu_i \in \mathcal{P}(\mathbb{R}^{d_i})$, for $i = 0, 1$, be compactly supported with $\mathrm{spt}(\mu_i) = S_i$. Then, the map $f \in \mathcal{C}(S_0 \times S_1) \mapsto \left(\int e^{f(\cdot,y)} d\mu_1(y), \int e^{f(x,\cdot)} d\mu_0(x) \right) \in \mathcal{C}(S_0) \times \mathcal{C}(S_1)$ is smooth with first derivative at $f \in \mathcal{C}(S_0 \times S_1)$ given by*

$$h \in \mathcal{C}(S_0 \times S_1) \mapsto \left(\int h(\cdot, y) e^{f(\cdot,y)} d\mu_1(y), \int h(x, \cdot) e^{f(x,\cdot)} d\mu_0(x) \right) \in \mathcal{C}(S_0) \times \mathcal{C}(S_1).$$

Proof First, we show that the map $f \in \mathcal{C}(S_0 \times S_1) \mapsto e^f \in \mathcal{C}(S_0 \times S_1)$ is Fréchet differentiable with $D(e^{(\cdot)})_{[f]}(h) = he^f$. Fix $f \in \mathcal{C}(S_0 \times S_1)$ and consider

$$\lim_{\substack{h \in \mathcal{C}(S_0 \times S_1) \\ \|h\|_{\infty, S_0 \times S_1} \to 0}} \frac{\left\| e^{f+h} - e^f - he^f \right\|_{\infty, S_0 \times S_1}}{\|h\|_{\infty, S_0 \times S_1}} \leq \|e^f\|_{\infty, S_0 \times S_1} \lim_{\substack{h \in \mathcal{C}(S_0 \times S_1) \\ \|h\|_{\infty, S_0 \times S_1} \to 0}} \frac{\left\| e^h - 1 - h \right\|_{\infty, S_0 \times S_1}}{\|h\|_{\infty, S_0 \times S_1}}.$$

Fix arbitrary $(x, y) \in S_0 \times S_1$. By a Taylor expansion,

$$e^{h(x,y)} = 1 + h(x, y) + \frac{1}{2} e^{\xi(x,y)} h^2(x, y),$$

where $|\xi(x, y)| \in [0, |h(x, y)|]$ i.e. $\|\xi\|_{\infty, S_0 \times S_1} \leq \|h\|_{\infty, S_0 \times S_1}$. That is,

$$\lim_{\substack{h \in \mathcal{C}(S_0 \times S_1) \\ \|h\|_{\infty, S_0 \times S_1} \to 0}} \frac{\left\| e^h - 1 - h \right\|_{\infty, S_0 \times S_1}}{\|h\|_{\infty, S_0 \times S_1}} \leq \lim_{\substack{h \in \mathcal{C}(S_0 \times S_1) \\ \|h\|_{\infty, S_0 \times S_1} \to 0}} \frac{1}{2} e^{\|\xi\|_{\infty, S_0 \times S_1}} \|h\|_{\infty, S_0 \times S_1} = 0.$$

On the other hand, the derivative of $f \in \mathcal{C}(S_0 \times S_1) \mapsto \int f(x,y)d\mu_1(y) \in \mathcal{C}(S_0)$ at any point is given by $h \in \mathcal{C}(S_0 \times S_1) \mapsto \int h(x,y)d\mu_1(y) \in \mathcal{C}(S_0)$. The claimed expression for the first derivative then follows by the chain rule. The derivatives of this map can be computed to arbitrary order inductively by the prior argument. ■

Proof of Lemma 17 Observe that the map $(\boldsymbol{A}, \varphi_0, \varphi_1) \in \mathbb{R}^{d_0 \times d_1} \times \mathfrak{E} \mapsto \varphi_0 \oplus \varphi_1 - c_{\boldsymbol{A}} \in \mathcal{C}(S_0 \times S_1)$ is smooth with first derivative at $(\boldsymbol{A}, \varphi_0, \varphi_1) \in \mathbb{R}^{d_0 \times d_1} \times \mathfrak{E}$ given by

$$(\boldsymbol{B}, h_0, h_1) \in \mathbb{R}^{d_0 \times d_1} \times \mathfrak{E} \mapsto h_0 \oplus h_1 + 32x^\intercal \boldsymbol{B} y \in \mathcal{C}(S_0 \times S_1).$$

The result then follows from Lemma 32 by applying the chain rule. ■

D.2 Compactness of $\mathcal{L}$

Lemma 33 (Example 2 in Yosida 1995)**.** *Let $\varepsilon > 0$, $\mu_0 \in \mathcal{P}(\mathbb{R}^{d_0})$, $\mu_1 \in \mathcal{P}(\mathbb{R}^{d_1})$, and $\boldsymbol{A} \in \mathbb{R}^{d_0 \times d_1}$ be arbitrary and let $(\varphi_0^{\boldsymbol{A}}, \varphi_1^{\boldsymbol{A}})$ be EOT potentials for $\mathsf{OT}_{\boldsymbol{A},\varepsilon}(\mu_0, \mu_1)$. Then, the map $\mathcal{L} : L^2(\mu_0) \times L^2(\mu_1) \mapsto L^2(\mu_0) \times L^2(\mu_1)$ defined by*

$$\mathcal{L}(f_0, f_1) = \left(\int f_1(y) e^{\frac{\varphi_0^{\boldsymbol{A}}(x) + \varphi_1^{\boldsymbol{A}}(y) - c_{\boldsymbol{A}}(x,y)}{\varepsilon}} d\mu_1(y), \int f_0(x) e^{\frac{\varphi_0^{\boldsymbol{A}}(x) + \varphi_1^{\boldsymbol{A}}(y) - c_{\boldsymbol{A}}(x,y)}{\varepsilon}} d\mu_0(x) \right),$$

is compact.

Proof For simplicity, we prove only that

$$\mathcal{L}_2 : f \in L^2(\mu_0) \mapsto \int f(x)\xi(x,\cdot)d\mu_0(x) \in L^2(\mu_1),$$

is a compact operator for $\xi : (x,y) \in \mathbb{R}^{d_0} \times \mathbb{R}^{d_1} \mapsto e^{\frac{\varphi_0^{\boldsymbol{A}}(x) + \varphi_1^{\boldsymbol{A}}(y) - c_{\boldsymbol{A}}(x,y)}{\varepsilon}}$. For any $y \in \mathbb{R}^{d_1}$ and $f \in L^2(\mu_0)$, $|\mathcal{L}_2(f)(y)|^2 \leq \|f\|^2_{L^2(\mu_0)} \int |\xi(\cdot, y)|^2 d\mu_0$, as $\xi(\cdot, y)$ is bounded on $\mathrm{spt}(\mu_0)$ this operator is well-defined.

Let f_n be a bounded sequence in $L^2(\mu_0)$. By the Eberlein-Šmulian theorem (Yosida, 1995, p. 141), up to passing to a subsequence, f_n converges weakly to $f \in L^2(\mu_0)$. For fixed $y \in \mathbb{R}^{d_1}$, $\xi(\cdot, y) \in L^2(\mu_0)$, hence $\mathcal{L}_2(f_n)(y) \to \mathcal{L}_2(f)(y)$ and it follows from the dominated convergence theorem that, for any $g \in L^2(\mu_1)$, $\int \mathcal{L}_2(f_n) g d\mu_1 \to \int \mathcal{L}_2(f) g d\mu_1$ such that $\mathcal{L}_2(f_n) \to \mathcal{L}_2(f)$ weakly in $L^2(\mu_1)$. Also, by dominated convergence,

$$\|\mathcal{L}_2(f_n)\|^2_{L^2(\mu_1)} = \int \mathcal{L}_2(f_n)^2 d\mu_1 \to \int \mathcal{L}_2(f)^2 d\mu_1 = \|\mathcal{L}_2(f)\|^2_{L^2(\mu_1)},$$

such that $\mathcal{L}_2(f_n) \to \mathcal{L}_2(f)$ strongly in $L^2(\mu_1)$. As f_n was an arbitrary bounded sequence in $L^2(\mu_0)$ and $\mathcal{L}_2(f_n) \to \mathcal{L}_2(f)$ strongly in $L^2(\mu_1)$ up to a subsequence, $\mathcal{L}_2$ is a compact operator. ■