Entropic Gromov-Wasserstein Distances: Stability and Algorithms

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

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

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

Abstract

The Gromov-Wasserstein (GW) distance quantifies discrepancy between metric measure spaces, but suffers from computational hardness. The entropic Gromov-Wasserstein (EGW) distance serves as a computationally efficient proxy for the GW distance. Recently, it was shown that the quadratic GW and EGW distances admit variational forms that tie them to the well-understood optimal transport (OT) and entropic OT (EOT) problems. By leveraging this connection, we establish convexity and smoothness properties of the objective in this variational problem. This results in the first efficient algorithms for solving the EGW problem that are subject to formal guarantees in both the convex and non-convex regimes.

1 Introduction

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

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

where $\Pi(\mu_0, \mu_1)$ is the set of couplings between μ_0 and μ_1. This approach, proposed in [25], 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 quotient space of all mm spaces modulo obtained by identifying isomorphic mm spaces (i.e. the underlying measures μ_0, μ_1) are such that $\mu_0 \circ T^{-1} = \mu_1$ for some isometry $T : \mathcal{X}_0 \to \mathcal{X}_1$). From an applied standpoint, a solution to the GW problem between two heterogeneous datasets yields both a quantification of discrepancy, and an optimal matching $\pi^\star$ between them. As such, the GW distance has seen many applications, encompassing single-cell genomics [5, 15], alignment of language models [1], shape matching [23, 24], graph matching [39, 40], heterogeneous domain adaptation [41], and generative modeling [8].

Exact computation of the GW distance is a quadratic assignment problem, which is known to be NP-complete [11]. The computational intractability of the GW problem in (1) has inspired several reformulations that aim to alleviate this issue. Recent approaches include slicing [38], relaxing

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

the strict marginal constraints using f-divergence penalties [33], and optimizing over bi-directional maps [44]. While these methods offer certain advantages, it is the approach based on entropic regularization [29, 36] that is most frequently used in application. In [29], it is proposed to solve the entropic Gromow-Wasserstein problem (EGW) via a mirror descent algorithm with a complexity of $O(N^3)$ for marginals supported on N distinct points (see, e.g., Remark 1 in [29]). The follow-up work [32] proposes a low-rank variant of the EGW problem which can be solved in linear time, wherein only couplings admitting a certain low-rank structure are considered. As an intermediate step of their analysis, they show that the complexity of mirror descent 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 quadratic and N dominates the ambient dimensions. Although mirror descent seems to solve the EGW problem well in practice, formal guarantees concerning convergence rates or local optimality are lacking.

The goal of this work is to address the computational gap described above, targeting algorithms with non-asymptotic guarantees and establishing convexity regimes of the EGW problem—all of which are consequences of a new stability analysis of the EGW variational representation from [43].

2 Notation and preliminaries

For a topological space S, we let $\mathcal{P}(S)$ be the collection of all Borel probability distributions on S. The Frobenius inner product on $\mathbb{R}^{d_0\times d_1}$ is defined by $\langle \boldsymbol{A},\boldsymbol{B}\rangle_F=\mathrm{tr}\,(\boldsymbol{A}^\intercal\boldsymbol{B})$; the associated norm is denoted by $\|\cdot\|_F$. A function $f:\mathbb{R}^d\to\mathbb{R}$ is ρ-weakly convex if $f+\frac{\rho}{2}\|\cdot\|^2$ is convex, it is L-smooth if its gradient is L-Lipschitz. For a Fréchet differentiable map $F:U\to V$ between normed vector spaces U and V, we denote the derivative of F at the point $u\in U$ evaluated at $v\in V$ by $DF_{[u]}(v)$. We adopt the shorthands $a\wedge b=\min\{a,b\}$ and $a\vee b=\max\{a,b\}$.

2.1 Entropic optimal transport

Entropic regularization transforms the linear OT problem into a strongly convex one. Given distributions $\mu_i\in\mathcal{P}(\mathbb{R}^{d_i})$, $i=0,1$, and a cost function $c:\mathbb{R}^{d_0}\times\mathbb{R}^{d_1}\to\mathbb{R}$, 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 $\mathsf{D}_{\mathsf{KL}}(\mu\|\nu)=\int\log\left(\frac{d\mu}{d\nu}\right)d\mu$, if $\mu\ll\nu$, and ∞, otherwise. 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,

$$\mathsf{OT}_\varepsilon(\mu_0,\mu_1)=\sup_{(\varphi_0,\varphi_1)\in L^1(\mu_0)\times L^1(\mu_1)}\int\varphi_0d\mu_0+\int\varphi_1d\mu_1-\varepsilon\int e^{\frac{\varphi_0\oplus\varphi_1-c}{\varepsilon}}d\mu_0\otimes\mu_1+\varepsilon,$$

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

$$\int e^{\frac{\varphi_0(x)+\varphi_1-c(x,\cdot)}{\varepsilon}}d\mu_1=1,\quad \mu_0\text{-a.e. }x\in\mathbb{R}^{d_0},\quad \int e^{\frac{\varphi_0+\varphi_1(y)-c(\cdot,y)}{\varepsilon}}d\mu_0=1,\quad \mu_1\text{-a.e. }y\in\mathbb{R}^{d_1},\tag{2}$$

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 (2) 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 $a\in\mathbb{R}$. The unique EOT coupling π_ε is characterized by $\frac{d\pi_\varepsilon}{d\mu_0\otimes\mu_1}(x,y)=e^{\frac{\varphi_0(x)+\varphi_1(y)-c(x,y)}{\varepsilon}}$, and, under some additional conditions on the cost and marginals, (2) admits a pair of continuous solutions which is unique up to additive constants and satisfies the system everywhere, i.e., 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 [28] 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 and the inner product cost. By analogy to OT, EGW serves as a proxy of the standard (p,q)-GW distance. 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 proceed to define the EGW distance for the quadratic cost.

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

$$\mathsf{S}_\varepsilon(\mu_0,\mu_1) = \inf_{\pi\in\Pi(\mu_0,\mu_1)} \int \left|\|x-x'\|^2 - \|y-y'\|^2\right|^2 d\pi\otimes\pi(x,y,x',y') + \varepsilon \mathsf{D}_{\mathsf{KL}}(\pi\|\mu_0\otimes\mu_1). \quad (3)$$

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. When μ_0, μ_1 are centered, which we may assume without loss of generality, the EGW distance decomposes as

$$\begin{aligned} \mathsf{S}_\varepsilon(\mu_0,\mu_1) &= \mathsf{S}_1(\mu_0,\mu_1) + \mathsf{S}_{2,\varepsilon}(\mu_0,\mu_1), \\ \mathsf{S}_1(\mu_0,\mu_1) &= \int \|x-x'\|^4 d\mu_0\otimes\mu_0(x,x') + \int \|y-y'\|^4 d\mu_1\otimes\mu_1(y,y') - 4M_2(\mu_0)M_2(\mu_1), \\ \mathsf{S}_{2,\varepsilon}(\mu_0,\mu_1) &= \inf_{\boldsymbol{A}\in\mathbb{R}^{d_0\times d_1}} 32\|\boldsymbol{A}\|_F^2 + \mathsf{OT}_{\boldsymbol{A},\varepsilon}(\mu_0,\mu_1), \end{aligned} \quad (4)$$

where $\mathsf{OT}_{\boldsymbol{A},\varepsilon}(\mu_0,\mu_1)$ is the EOT problem with the cost function $c_{\boldsymbol{A}} : (x,y) \in \mathbb{R}^{d_0}\times\mathbb{R}^{d_1} \mapsto -4\|x\|^2\|y\|^2 - 32x^\intercal \boldsymbol{A} y$ and regularization parameter ε. Moreover, the infimum is achieved at some $\boldsymbol{A}^\star \in \mathcal{D}_M := [-M/2, M/2]^{d_0\times d_1}$ for any $M \geq \sqrt{M_2(\mu_0)M_2(\mu_1)} =: M_{\mu_0,\mu_1}$. The proof of Theorem 1 in [43] demonstrates that if μ_0 and μ_1 are centered and $\pi_\star$ is optimal for the original EGW formulation, then $\boldsymbol{A}^\star = \frac{1}{2}\int xy^\intercal\, d\pi_\star(x,y)$ is optimal for $\mathsf{S}_{2,\varepsilon}$ and $\pi_\star = \pi_{\boldsymbol{A}^\star}$, where $\pi_{\boldsymbol{A}^\star}$ is the unique EOT coupling for $\mathsf{OT}_{\boldsymbol{A}^\star,\varepsilon}(\mu_0,\mu_1)$. Corollary 1 ahead expands on this connection by establishing a one-to-one correspondence between solutions of S_ε and $\mathsf{S}_{2,\varepsilon}$.

Although (4) illustrates a connection between the EGW and EOT problems, the outer minimization over $\mathcal{D}_M$ necessitates studying EOT with an *a priori* unknown cost function $c_{\boldsymbol{A}}$.

A similar decomposition holds for the inner product GW problem, where the difference of squared Euclidean norms is replaced by a difference of inner products. In that case, $\mathsf{F}_\varepsilon(\mu_0,\mu_1) = \mathsf{F}_1(\mu_0,\mu_1) + \mathsf{F}_{2,\varepsilon}(\mu_0,\mu_1)$, for $\mathsf{F}_1(\mu_0,\mu_1) = \int |\langle x,x'\rangle|^2 d\mu_0\otimes\mu_0(x,x') + \int |\langle y,y'\rangle|^2 d\mu_0\otimes\mu_0(y,y')$, and $\mathsf{F}_{2,\varepsilon}(\mu_0,\mu_1) = \inf_{\boldsymbol{A}\in\mathbb{R}^{d_0\times d_1}} 8\|\boldsymbol{A}\|_F^2 + \mathsf{IOT}_{\boldsymbol{A},\varepsilon}(\mu_0,\mu_1)$, with the distinction that no centering is needed and $\mathsf{IOT}_{\boldsymbol{A},\varepsilon}(\mu_0,\mu_1)$ is the EOT problem with the cost function $c_{\boldsymbol{A}}(x,y) = -8x^\intercal \boldsymbol{A} y$. We restrict our attention to the quadratic EGW problem, similar results hold in the inner product case.

3 Stability of entropic Gromov-Wasserstein distances

We now analyze the stability of the EGW problem with respect to the matrix $\boldsymbol{A}$ appearing in its variational form (4). Specifically, we characterize the first and second derivatives of the objective function whose optimization defines $\mathsf{S}_{2,\varepsilon}$ which elucidates its convexity properties and enables us to 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 2).

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 $\mathsf{S}_{2,\varepsilon}(\mu_0,\mu_1)$. We first characterize the derivatives of Φ and then prove that this map is weakly convex and L-smooth.

Proposition 1 (First and second derivatives)**.** $\Phi : \boldsymbol{A} \in \mathbb{R}^{d_0\times d_1} \mapsto 32\|\boldsymbol{A}\|_F^2 + \mathsf{OT}_{\boldsymbol{A},\varepsilon}(\mu_0,\mu_1)$ *is smooth, coercive, and has first and second-order Fréchet derivatives at* $\boldsymbol{A} \in \mathbb{R}^{d_0\times d_1}$ *given by*

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

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

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*

$$\begin{aligned} \int \left(h_0^{\boldsymbol{A},\boldsymbol{C}}(x) + h_1^{\boldsymbol{A},\boldsymbol{C}}(y) - 32x^\intercal \boldsymbol{C} y\right) e^{\frac{\varphi_0^{\boldsymbol{A}}(x)+\varphi_1^{\boldsymbol{A}}(y)-c_{\boldsymbol{A}}(x,y)}{\varepsilon}} d\mu_1(y) = 0, \quad \forall\, x\in\mathrm{spt}(\mu_0), \\ \int \left(h_0^{\boldsymbol{A},\boldsymbol{C}}(x) + h_1^{\boldsymbol{A},\boldsymbol{C}}(y) - 32x^\intercal \boldsymbol{C} y\right) e^{\frac{\varphi_0^{\boldsymbol{A}}(x)+\varphi_1^{\boldsymbol{A}}(y)-c_{\boldsymbol{A}}(x,y)}{\varepsilon}} d\mu_0(x) = 0, \quad \forall\, y\in\mathrm{spt}(\mu_1). \end{aligned} \quad (5)$$

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 1 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 (2). The derivative of $\mathsf{OT}_{(\cdot),\varepsilon}(\mu_0, \mu_1)$, which is a simple function of the EOT potentials, is then readily obtained. By differentiating the Frobenius norm, this yields the derivative of Φ. See Appendix A.1.

As $D\Phi_{[\boldsymbol{A}]}(\boldsymbol{B}) = \left\langle 64\boldsymbol{A} - 32\int xy^\intercal \, d\pi_{\boldsymbol{A}}(x,y), \boldsymbol{B}\right\rangle_F$, we can interpret $64\boldsymbol{A} - 32\int xy^\intercal \, d\pi_{\boldsymbol{A}}(x,y)$ as the gradient of Φ which we denote $D\Phi_{[\boldsymbol{A}]}$. This perspective is utilized in Section 4 when studying computational guarantees for the EGW distance, as it is simpler to view iterates as matrices.

As a direct corollary to Proposition 1, we provide an (implicit) characterization of the stationary points of Φ and connect its minimizers to solutions of S_ε. Details are provided in Appendix A.2.

Corollary 1 (Stationary points and correspondence between S_ε and $\mathsf{S}_{2,\varepsilon}$)**.**

(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, $\lambda_{\max}\left(D^2\Phi_{[\boldsymbol{A}]}\right)$ and $\lambda_{\min}\left(D^2\Phi_{[\boldsymbol{A}]}\right)$, can be controlled which enables us to characterize convexity and smoothness of Φ.

Theorem 1 (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 $L \leq 64 \vee \left(32^2\varepsilon^{-1}\sqrt{M_4(\mu_0)M_4(\mu_1)} - 64\right)$.*

Theorem 1 follows from Proposition 1 by considering the variational form of the maximal and minimal eigenvalues; see Appendix A.3 for details. In general, optimal EGW couplings may not be unique. Theorem 1 provides sufficient conditions for uniqueness of solutions to both $\mathsf{S}_{2,\varepsilon}$ and the EGW problem by the connection discussed in Corollary 1 when the marginals are centered.

4 Computational guarantees

Building on this stability theory, we now study computation of the EGW problem. The goal is to compute the distance between two discrete distributions $\mu_0 \in \mathcal{P}(\mathbb{R}^{d_0})$ and $\mu_1 \in \mathcal{P}(\mathbb{R}^{d_1})$ supported on N_0 and N_1 atoms $(x^{(i)})_{i=1}^{N_0}$ and $(y^{(j)})_{j=1}^{N_1}$, respectively. In light of the decomposition (4), we focus on $\mathsf{S}_{2,\varepsilon}$, which is given by a smooth optimization problem whose convexity depends on the value of ε. Throughout, we 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 Eq. (5)), we focus on first-order methods. Given the regularity of the $\mathsf{S}_{2,\varepsilon}$ 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 requires computing the corresponding EOT plan, which entails an approximation. We model this under the scope of gradient methods with inexact gradient oracles [13, 16, 17].

For a fixed $\varepsilon > 0$ and μ_0, μ_1 as above, we seek 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 1). 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 $\mathbf{\Pi}^{\boldsymbol{A}} \in \mathbb{R}^{N_0 \times N_1}$, where $\mathbf{\Pi}^{\boldsymbol{A}}_{ij} = \pi^{\boldsymbol{A}}(x^{(i)}, y^{(j)})$. The inexact oracle paradigm assumes that, for any $\boldsymbol{A} \in \mathcal{D}_M$, we have access to a δ-oracle $\widetilde{\mathbf{\Pi}}^{\boldsymbol{A}}$ for $\mathbf{\Pi}^{\boldsymbol{A}}$ with $\|\widetilde{\mathbf{\Pi}}^{\boldsymbol{A}} - \mathbf{\Pi}^{\boldsymbol{A}}\|_\infty < \delta$. Such oracles can be obtained, for instance, by Sinkhorn's algorithm [12, 35].

Proposition 2 (Inexact oracle via Sinkhorn iterations)**.** *Fix $\delta > 0$. Then, Sinkhorn's algorithm (Algorithm 3) returns a δ-oracle approximation $\widetilde{\mathbf{\Pi}}^{\boldsymbol{A}}$ of $\mathbf{\Pi}^{\boldsymbol{A}}$ 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* (17)*.*

The proof of Proposition 2 follows by combining a number of known results, see Appendix C. 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 1. As convexity implies that the minimal eigenvalue of $D^2\Phi_{[\boldsymbol{A}]}$ is positive for any $\boldsymbol{A} \in \mathcal{D}_M$, Theorem 1 further yields that Φ is 64-smooth. With that, we can the apply inexact oracle first-order method from [13]. To describe the approach, assume that we are given a δ-oracle $\widetilde{\mathbf{\Pi}}^{\boldsymbol{A}}$ for the EOT plan $\mathbf{\Pi}^{\boldsymbol{A}}$ for $\mathsf{OT}_{\boldsymbol{A},\varepsilon}(\mu_0, \mu_1)$, and define the corresponding gradient approximation

$$\widetilde{D}\Phi_{[\boldsymbol{A}]} = 64\boldsymbol{A} - 32 \sum_{\substack{1 \le i \le N_0 \\ 1 \le j \le N_1}} x^{(i)} (y^{(j)})^\intercal \widetilde{\mathbf{\Pi}}^{\boldsymbol{A}}_{ij}. \tag{6}$$

Algorithm 1 Fast gradient method with inexact oracle

Fix $L = 64$ and let $\alpha_k = \frac{k+1}{2}$, and $\tau_k = \frac{2}{k+3}$

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

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

The sign, min, and multiplication operations in Algorithm 1 are applied entrywise. 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 1.

We now provide formal convergence guarantees for Algorithm 1.

Theorem 2 (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* (6) *satisfy $\Phi(\boldsymbol{B}_k) - \Phi(\boldsymbol{B}^\star) \le \frac{2L\|\boldsymbol{B}^\star\|_F^2}{(k+1)(k+2)} + 3\delta'$, where $\boldsymbol{B}^\star$ is a global minimizer of Φ and $\delta' = 32M\delta \sum_{\substack{1 \le i \le N_0 \\ 1 \le j \le N_1}} \left\| x^{(i)}(y^{(j)})^\intercal \right\|_1$ where $\|\cdot\|_1$ denotes the entrywise 1-norm. Moreover, for any $\eta > 3\delta'$, Algorithm 1 requires at most $k = \left\lceil -\frac{3}{2} + \frac{1}{2}\sqrt{1 + \frac{8L\|\boldsymbol{B}^\star\|_F^2}{\eta - 3\delta'}} \right\rceil \le \left\lceil -\frac{3}{2} + \frac{1}{2}\sqrt{1 + \frac{128M^2 d_0^2 d_1^2}{\eta - 3\delta'}} \right\rceil$ iterations to achieve an η-approximate solution.*

The proof of Theorem 2, given in Appendix A.4, follows from Theorem 2.2 in [13] after casting our problem as an instance of their setting. Some implications of Theorem 2 are discussed next.

Remark 1 (Optimal rates and stopping conditions)**.** *First, consider the convergence rate of the function values. 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., [27]). 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, observe that all terms involved in the upper bound for the number of iterations are explicit as soon as a desired precision η is chosen since the oracle error δ can be fixed according to Proposition 2. Consequently, it can be used as an explicit stopping condition for Algorithm 1.

4.3 General case

Algorithm 2 Adaptive gradient method with inexact oracle

Given $\boldsymbol{C}_0 \in \mathcal{D}_M$, fix the 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**
  $\boldsymbol{D}_k \leftarrow \boldsymbol{A}_k - \beta_k \boldsymbol{G}_k$
  $\boldsymbol{B}_k \leftarrow \frac{M}{2}\operatorname{sign}(\boldsymbol{D}_k)\min\left(\frac{2}{M}|\boldsymbol{D}_k|, 1\right)$
  $\boldsymbol{E}_k \leftarrow \boldsymbol{C}_{k-1} - \gamma_k \boldsymbol{G}_k$
  $\boldsymbol{C}_k \leftarrow \frac{M}{2}\operatorname{sign}(\boldsymbol{E}_k)\min\left(\frac{2}{M}|\boldsymbol{E}_k|, 1\right)$
  $\boldsymbol{A}_{k+1} \leftarrow \tau_k \boldsymbol{C}_k + (1-\tau_k)\boldsymbol{B}_k$
  $\boldsymbol{G}_{k+1} \leftarrow \widetilde{D}\Phi_{[\boldsymbol{A}_{k+1}]}$
  $k \leftarrow k+1$
return $\boldsymbol{B}_k$

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 1, 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 [21] to account for our inexact oracle. Notably, their method adapts to the convexity of Φ as described in Theorem 3.

Unlike Algorithm 1, which is 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 $\boldsymbol{A}_0$ is set to a stationary point of Φ, then $D\Phi_{[\boldsymbol{A}_0]} = \boldsymbol{0}$ and, consequently $\widetilde{D}\Phi_{[\boldsymbol{A}_0]} \approx \boldsymbol{0}$ (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 [21]. For completeness, we provide a self-contained argument in Appendix D along with a discussion of how this problem fits in the framework of [21].

Theorem 3 (Adaptive convergence rate)**.** *Assume that Φ is L-smooth on $\mathcal{D}_M$ and that $\widetilde{\boldsymbol{\Pi}}^{\boldsymbol{A}}$ is a δ-oracle for $\boldsymbol{\Pi}^{\boldsymbol{A}}$. Then, the iterates $\boldsymbol{A}_k, \boldsymbol{B}_k$ in Algorithm 2 with $\widetilde{D}\Phi_{[\boldsymbol{A}_k]}$ given by* (6) *satisfy*

1. *If Φ is non-convex and $\mathsf{OT}_{(\cdot),\varepsilon}(\mu_0,\mu_1)$ is L'-smooth, then $\min_{1\le i\le k}\left\|\beta_i^{-1}(\boldsymbol{B}_i - \boldsymbol{A}_i)\right\|_F^2 \le \frac{96L^2}{k(k+1)(k+2)}\|\boldsymbol{C}_0 - \boldsymbol{B}^\star\|_F^2 + \frac{24LL'}{k}\left(\|\boldsymbol{B}^\star\|_F^2 + \frac{5M^2d_0^2d_1^2}{16}\right) + 8L\delta'$, where $\boldsymbol{B}^\star$ is a global minimizer of Φ, and $\delta' = 32M\delta\sum_{\substack{1\le i\le N_0\\1\le j\le N_1}}\left\|x^{(i)}(y^{(j)})^\intercal\right\|_1$.*

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

We first show that when $\left\|\beta_k^{-1}(\boldsymbol{B}_k - \boldsymbol{A}_k)\right\|_F$ is small, $D\Phi_{[\boldsymbol{A}_k]}$ is approximately stationary.

Corollary 2 (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 2 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 Appendix A.5 for 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. By analogy with Remark 1, when all iterates are interior points Algorithm 2 yields a bound on the number of iterations required to achieve an approximate stationary point.

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

Remark 2 (Adaptivity of Algorithm 2)**.** *As in Theorem 2, the convergence rates 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 [21]. 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 [34] and Theorem 3.1 in [10]). This adaptivity is beneficial, as Φ may be convex beyond the conditions derived in Theorem 1.*

An empirical comparison of Algorithms 1 and 2 in the convex setting is included in Section 4.4. 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 Sinkhorn iterations.

Remark 3 (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. [32]), the gradient approximation* (6) *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 Numerical Experiments

We conclude this section with some experiments that empirically validate the rates obtained in Theorems 2 and 3 and the computational complexity discussed in Remark 3. All experiments were performed on a desktop computer with 16 GB of RAM and an Intel i5-10600k CPU using the Python programming language. Blown-up figures are included in Appendix F

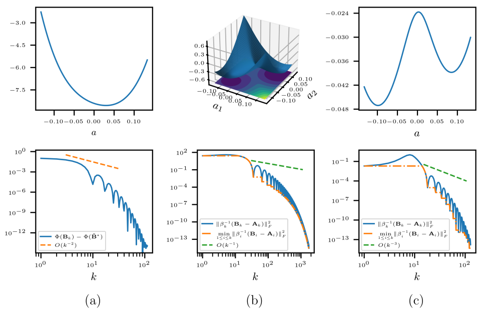


Figure 1: The top row compiles plots of Φ for the different examples. The bottom row consists of plots tracking the progress of the iterates. In (b) and (c), Algorithm 2 is initialized at $\boldsymbol{C}_0 = (1,1)\times 10^{-5}$ and $\boldsymbol{C}_0 = 1\times 10^{-5}$, respectively.

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 2 on the optimality gap $\Phi(\boldsymbol{B}_k) - \Phi(\boldsymbol{B}^\star)$ is seen to hold.[1] Figure 1 (b) illustrates the progress of Algorithm 2 applied to a non-convex Φ, for $\mu_0 = \frac{1}{3}\left(\delta_{0.3} + \delta_{-0.8} + \delta_{-0.5}\right)$ 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 $\min_{1\le i\le k}\left\|\beta_i^{-1}(\boldsymbol{B}_i - \boldsymbol{A}_i)\right\|_F^2$ in the non-convex case from Theorem 3 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 $\mu_0 = \frac{1}{5}\left(\delta_{-0.1} + \delta_{-0.2} + \delta_{0.2} + \delta_{-0.3} + \delta_{0.3}\right)$ and $\mu_1 = \frac{1}{5}\left(\delta_{0.2} + \delta_{-0.3} + \delta_{0.3} + \delta_{-0.4} + \delta_{0.4}\right)$ 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 (6) is computed using the implementation of Sinkhorn's algorithm from [19].

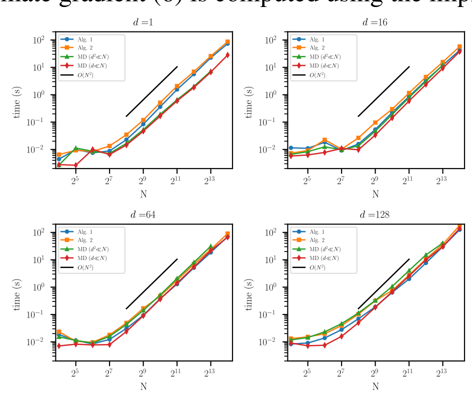


Figure 2: The various plots compile the average runtime of Algorithms 1 and 2, and two versions of the mirror descent algorithm in the convex regime for different combinations of d and N.

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 $\{(\mu_{0,i}, \mu_{1,i})\}_{i=1}^{500}$. In the sequel, a *single experiment* refers to the process of timing the computation of $\mathsf{S}_\varepsilon(\mu_{0,i}, \mu_{1,i})$ for some fixed d, N and all $i = 1, \dots, 500$. For practical reasons, we choose to abort an experiment before all 500 EGW distances have been computed if the total runtime for this experiment exceeds 20 minutes. The average runtime is then computed among all completed calculations in a single experiment.

<u>The convex case:</u> 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 1. 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

[1]The plot shows the approximate gap $\Phi(\boldsymbol{B}_k) - \Phi(\bar{\boldsymbol{B}}^\star)$, where $\bar{\boldsymbol{B}}^\star$ is the approximate minimizer.

methods with the two implementations of the $O(N^2)$ mirror descent algorithm provided in [32]. 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 [32] with some small modifications (e.g., EOT couplings are computed using Sinkhorn's algorithm from the Python Optimal Transport package [19] and some extraneous logging features are removed). We note that the first version of the mirror descent algorithm encounters "out of memory" errors for $N = 16384$.

The plots show that the four algorithms perform similarly on the considered examples, and empirically validate the $O(N^2)$ computational complexity from Remark 3. To verify that the algorithms all converge to solutions with similar objective values, we evaluate the relative error[2] between all pairs of algorithms for each d, N. The largest relative error we observe is 6.6×10^{-6} for $d = 1$ and, for the other choices of d, is at most 4.2×10^{-12}. The values obtained are thus in good agreement.

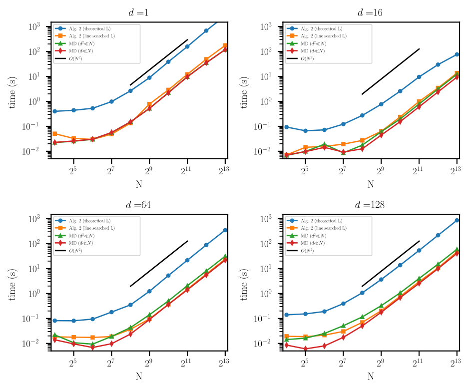


Figure 3: The various plots compile the average runtime of Algorithm 2 with the two methods for choosing L, and two versions of the mirror descent algorithm in the non-convex regime for different combinations of d and N.

The non-convex case: To evaluate the performance of Algorithm 2 when convexity is unknown, we set ε to violate the condition of Theorem 1, 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 convergence. The first is to set L equals to the theoretical upper bound from Theorem 1, 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. 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 $10\times$ 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 [3, 26, 37], 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, but at the cost of increasing the per iteration complexity.

5 Conclusion

In this work, we have addressed stability for the EGW problem over Euclidean spaces with quadratic cost. The analysis leveraged variational characterizations of these EGW distances that tie them to EOT with a certain parametrized cost function. The stability analysis was used to study convexity

[2] 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.

and smoothness properties of this variational problem, which led to two new efficient algorithms for computing the EGW distance. The complexity of these algorithms agrees with the best known complexity of $O(N^2)$ for computing the quadratic EGW distance directly, but unlike previous approaches, we provide, for the first time, non-asymptotic convergence rate guarantees in both the convex and non-convex regimes. This stability analysis also lays the groundwork for solving the EGW problem via smooth optimization methods.

Acknowledgments and Disclosure of Funding

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

References

[1] David Alvarez-Melis and Tommi Jaakkola, *Gromov-wasserstein alignment of word embedding spaces*, Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2018, pp. 1881–1890.

[2] Tom M. Apostol, *Mathematical analysis*, 5 ed., Addison-Wesley, 1974.

[3] Stephen R Becker, Emmanuel J Candès, and Michael C Grant, *Templates for convex cone problems with applications to sparse signal recovery*, Mathematical programming computation **3** (2011), 165–218.

[4] Garrett Birkhoff, *Extensions of Jentzsch's theorem*, Transactions of the American Mathematical Society **85** (1957), no. 1, 219–227.

[5] Andrew J Blumberg, Mathieu Carriere, Michael A Mandell, Raul Rabadan, and Soledad Villar, *MREC: a fast and versatile framework for aligning and matching point clouds with applications to single cell molecular data*, arXiv preprint arXiv:2001.01666 (2020).

[6] J Frédéric Bonnans and Alexander Shapiro, *Perturbation analysis of optimization problems*, Springer Science & Business Media, 2013.

[7] Haim Brézis, *Functional analysis, sobolev spaces and partial differential equations*, vol. 2, Springer, 2011.

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

[9] Guillaume Carlier and Maxime Laborde, *A differential approach to the multi-marginal Schrödinger system*, SIAM Journal on Mathematical Analysis **52** (2020), no. 1, 709–717.

[10] Shuo Chen, Bin Shi, and Ya-xiang Yuan, *Gradient norm minimization of Nesterov acceleration: $o(1/k^3)$*, arXiv preprint arXiv:2209.08862 (2022).

[11] Clayton W. Commander, *A survey of the quadratic assignment problem, with applications*, Morehead Electronic Journal of Applicable Mathematics **4** (2005), MATH–2005–01.

[12] Marco Cuturi, *Sinkhorn distances: lightspeed computation of optimal transport*, Proceedings of the 26th International Conference on Neural Information Processing Systems, 2013, pp. 2292–2300.

[13] Alexandre d'Aspremont, *Smooth optimization with approximate gradient*, SIAM Journal on Optimization **19** (2008), no. 3, 1171–1183.

[14] Alexandre d'Aspremont, Laurent El Ghaoui, Michael I. Jordan, and Gert R. G. Lanckriet, *A direct formulation for sparse PCA using semidefinite programming*, SIAM Review **49** (2007), no. 3, 434–448.

[15] Pinar Demetci, Rebecca Santorella, Björn Sandstede, William Stafford Noble, and Ritambhara Singh, *SCOT: single-cell multi-omics alignment with optimal transport*, Journal of Computational Biology **29** (2022), no. 1, 3–18.

[16] Olivier Devolder, François Glineur, and Yurii Nesterov, *First-order methods of smooth convex optimization with inexact oracle*, Mathematical Programming **146** (2014), 37–75.

[17] Pavel Dvurechensky, *Gradient method with inexact oracle for composite non-convex optimization*, arXiv preprint arXiv:1703.09180 (2017).

[18] Pavel Dvurechensky, Alexander Gasnikov, and Alexey Kroshnin, *Computational optimal transport: complexity by accelerated gradient descent is better than by Sinkhorn's algorithm*, International conference on machine learning, PMLR, 2018, pp. 1367–1376.

[19] Rémi Flamary, Nicolas Courty, Alexandre Gramfort, Mokhtar Z. Alaya, Aurélie Boisbunon, Stanislas Chambon, Laetitia Chapel, Adrien Corenflos, Kilian Fatras, Nemo Fournier, Léo Gautheron, Nathalie T.H. Gayraud, Hicham Janati, Alain Rakotomamonjy, Ievgen Redko, Antoine Rolet, Antony Schutz, Vivien Seguy, Danica J. Sutherland, Romain Tavenard, Alexander Tong, and Titouan Vayer, *POT: Python optimal transport*, Journal of Machine Learning Research **22** (2021), no. 78, 1–8.

[20] Joel Franklin and Jens Lorenz, *On the scaling of multidimensional matrices*, Linear Algebra and its applications **114** (1989), 717–735.

[21] Saeed Ghadimi and Guanghui Lan, *Accelerated gradient methods for nonconvex nonlinear and stochastic programming*, Mathematical Programming **156** (2016), no. 1-2, 59–99.

[22] Jean-Baptiste Hiriart-Urruty and Claude Lemaréchal, *Fundamentals of convex analysis*, Springer Science & Business Media, 2004.

[23] Patrice Koehl, Marc Delarue, and Henri Orland, *Computing the Gromov-Wasserstein distance between two surface meshes using optimal transport*, Algorithms **16** (2023), no. 3, 131.

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

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

[26] Yu Nesterov, *Gradient methods for minimizing composite functions*, Mathematical programming **140** (2013), no. 1, 125–161.

[27] Yurii Nesterov, *Introductory lectures on convex optimization: A basic course*, vol. 87, Springer Science & Business Media, 2003.

[28] Marcel Nutz, *Introduction to entropic optimal transport*, Lecture notes, Columbia University (2021).

[29] Gabriel Peyré, Marco Cuturi, and Justin Solomon, *Gromov-Wasserstein averaging of kernel and distance matrices*, International Conference on Machine Learning, PMLR, 2016, pp. 2664–2672.

[30] R Tyrrell Rockafellar, *Convex analysis*, vol. 11, Princeton university press, 1997.

[31] Hans Samelson, *On the Perron-Frobenius theorem.*, Michigan Mathematical Journal **4** (1957), no. 1, 57 – 59.

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

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

[34] Bin Shi, Simon S Du, Michael I Jordan, and Weijie J Su, *Understanding the acceleration phenomenon via high-resolution differential equations*, Mathematical Programming (2021), 1–70.

[35] Richard Sinkhorn, *Diagonal equivalence to matrices with prescribed row and column sums*, The American Mathematical Monthly **74** (1967), no. 4, 402–405.

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

[37] Paul Tseng, *On accelerated proximal gradient methods for convex-concave optimization*, https://www.mit.edu/~dimitrib/PTseng/papers/apgm.pdf, 2008.

[38] Titouan Vayer, Rémi Flamary, Romain Tavenard, Laetitia Chapel, and Nicolas Courty, *Sliced Gromov-Wasserstein*, arXiv preprint arXiv:1905.10124 (2020).

[39] Hongteng Xu, Dixin Luo, and Lawrence Carin, *Scalable Gromov-Wasserstein learning for graph partitioning and matching*, Advances in neural information processing systems **32** (2019).

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

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

[42] Kösaku Yosida, *Functional analysis*, Springer Science & Business Media, 1995.

[43] Zhengxin Zhang, Ziv Goldfeld, Youssef Mroueh, and Bharath K. Sriperumbudur, *Gromov-Wasserstein distances: entropic regularization, duality, and sample complexity*, 2022.

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

A Proofs

A.1 Proof of Proposition 1

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 1. *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,*

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

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

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

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 [9].

Lemma 2. *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 11 in Appendix E.2 demonstrates that $\mathcal{L}$ is a compact linear self-map of $L^2(\mu_0) \times L^2(\mu_1)$.

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

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

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 $F := \{(f_0, f_1) \in L^2(\mu_0) \times L^2(\mu_1) : \int f_0 d\mu_0 = \int f_1 d\mu_1\}$. As in the lead-up to this lemma, $\xi_{\boldsymbol{A}}(E) \subset F$. By the Fredholm alternative (cf. Theorem 6.6 in [7]), $(\mathrm{Id} + \mathcal{L})(L^2(\mu_0) \times L^2(\mu_1))$ 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 [7]). $\square$

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

Lemma 3. *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

$$\begin{aligned}
&\int \left(h_0^{\boldsymbol{A},\boldsymbol{B}}(x) + h_1^{\boldsymbol{A},\boldsymbol{B}}(y) - 32x^\intercal \boldsymbol{B} y\right) e^{\frac{\varphi_0^{\boldsymbol{A}}(x)+\varphi_1^{\boldsymbol{A}}(y)-c_{\boldsymbol{A}}(x,y)}{\varepsilon}} d\mu_1(y) = 0, \quad \forall\, x \in \operatorname{spt}(\mu_0),\\
&\int \left(h_0^{\boldsymbol{A},\boldsymbol{B}}(x) + h_1^{\boldsymbol{A},\boldsymbol{B}}(y) - 32x^\intercal \boldsymbol{B} y\right) e^{\frac{\varphi_0^{\boldsymbol{A}}(x)+\varphi_1^{\boldsymbol{A}}(y)-c_{\boldsymbol{A}}(x,y)}{\varepsilon}} d\mu_0(x) = 0, \quad \forall\, y \in \operatorname{spt}(\mu_1),
\end{aligned} \tag{7}$$

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 1). By Lemma 2, $D_2\Upsilon_{\boldsymbol{A}}$ is an isomorphism and we may invoke the implicit mapping theorem (cf. Theorem 5.14 in [6]). 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

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

i.e., $-Dg_{[\boldsymbol{A}]}(\boldsymbol{B})$ solves (7). By construction, $g(\boldsymbol{B}) = (\varphi_0^{\boldsymbol{B}}, \varphi_1^{\boldsymbol{B}})$ and by repeating this process at any $\boldsymbol{A} \in \mathbb{R}^{d_0 \times d_1}$, we extend the differentiability of the potentials to the entire space $\mathbb{R}^{d_0 \times d_1}$. $\square$

Given the dual form of the EOT cost, Lemma 3 suffices to prove Proposition 1.

Proof of Proposition 1. 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 3 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 (7) w.r.t. μ_0 while using $\frac{d\pi_{\boldsymbol{A}}}{\mu_0 \otimes \mu_1}(x,y) = e^{\frac{\varphi_0^{\boldsymbol{A}}(x)+\varphi_1^{\boldsymbol{A}}(y)-c_{\boldsymbol{A}}(x,y)}{\varepsilon}}$, yields

$$\int \left(h_0^{\boldsymbol{A},\boldsymbol{B}}(x) + h_1^{\boldsymbol{A},\boldsymbol{B}}(y)\right) d\pi_{\boldsymbol{A}}(x,y) = \int h_0^{\boldsymbol{A},\boldsymbol{B}} d\mu_0 + \int h_1^{\boldsymbol{A},\boldsymbol{B}} d\mu_1 = 32 \int x^\intercal \boldsymbol{B} y \, d\pi_{\boldsymbol{A}}(x,y), \tag{8}$$

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 Section 2.1 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 1, 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

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

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,

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

Coercivity of Φ follows from the fact that

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

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

A.2 Proof of Corollary 1

Item (i). The expression for the stationary points follows immediately from Proposition 1. 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

$$|\boldsymbol{A}_{ij}| = \frac{1}{2}\left|\int x_i y_j \, d\pi_{\boldsymbol{A}}(x,y)\right| \leq \frac{1}{2}\int |x_i y_j| d\pi_{\boldsymbol{A}}(x,y) \leq \frac{1}{2}\sqrt{M_2(\mu_0)M_2(\mu_1)}.$$

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

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

By (4),

$$\begin{aligned}
\mathsf{S}_\varepsilon(\mu_0,\mu_1) &= \mathsf{S}_\varepsilon(\mu_0,\mu_1) + \mathsf{S}_{2,\varepsilon}(\mu_0,\mu_1) \\
&= \int \left|\|x-x'\|^2 - \|y-y'\|^2\right|^2 + 2\|x-x'\|^2\|y-y'\|^2 d\pi_{\boldsymbol{A}}\otimes\pi_{\boldsymbol{A}}(x,y,x',y') \\
&\quad - 4\int \|x\|^2\|y\|^2 d\mu_0\otimes\mu_1(x,y) - 4\int \|x\|^2\|y\|^2 d\pi_{\boldsymbol{A}}(x,y) \\
&\quad - 8\left\|\int xy^\intercal \, d\pi_{\boldsymbol{A}}(x,y)\right\|_F^2 + \varepsilon \mathsf{D}_{\mathsf{KL}}(\pi_{\boldsymbol{A}}||\mu_0\otimes\mu_1).
\end{aligned} \tag{9}$$

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 (9) 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 $\pi = \pi_{\boldsymbol{A}^\star}$. □

A.3 Proof of Theorem 1

The proof of Theorem 1 depends on the following lemma. The variance bound is seen to be sharp up to constants in Appendix B.

Lemma 4 (Hessian eigenvalue bounds). *The following hold for any $\boldsymbol{A}\in\mathbb{R}^{d_0\times d_1}$:*

(i) The minimal eigenvalue of $D^2\Phi_{[\boldsymbol{A}]}$, $\lambda_{\min}\left(D^2\Phi_{[\boldsymbol{A}]}\right)$, admits the lower bound $64 - 32^2\varepsilon^{-1}\sup_{\|\boldsymbol{C}\|_F=1}\mathrm{Var}_{\pi_{\boldsymbol{A}}}(X^\intercal \boldsymbol{C} Y)$, where $\sup_{\|\boldsymbol{C}\|_F=1}\mathrm{Var}_{\pi_{\boldsymbol{A}}}(X^\intercal \boldsymbol{C} Y) \leq \sqrt{M_4(\mu_0)M_4(\mu_1)}$.

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

Proof. 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 1. Recall that $(h_0^{\boldsymbol{A},\boldsymbol{C}}, h_1^{\boldsymbol{A},\boldsymbol{C}})$ satisfy

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

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

$$\begin{aligned}
&\int\left[\left(h_0^{\boldsymbol{A},\boldsymbol{C}}(x)\right)^2 + h_1^{\boldsymbol{A},\boldsymbol{C}}(y)h_0^{\boldsymbol{A},\boldsymbol{C}}(x) - 32x^\intercal\boldsymbol{C}yh_0^{\boldsymbol{A},\boldsymbol{C}}(x)\right]d\pi_{\boldsymbol{A}}(x,y) = 0, \\
&\int\left[h_0^{\boldsymbol{A},\boldsymbol{C}}(x)h_1^{\boldsymbol{A},\boldsymbol{C}}(y) + \left(h_1^{\boldsymbol{A},\boldsymbol{C}}(y)\right)^2 - 32x^\intercal\boldsymbol{C}yh_1^{\boldsymbol{A},\boldsymbol{C}}(y)\right]d\pi_{\boldsymbol{A}}(x,y) = 0.
\end{aligned}$$

Summing these equations gives

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

such that

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

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

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

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 (8), proving the desired inequality.

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

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

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 [22] 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 $\lambda_{\max}\left(D^2\Phi_{[\boldsymbol{A}]}\right) \leq 64$. $\square$

Proof of Theorem 1. We first discuss the convexity properties of Φ. By Lemma 4, $\lambda_{\min}\left(D^2\Phi_{[\boldsymbol{A}]} + \frac{\rho}{2}\|\boldsymbol{A}\|_F^2\right) \geq 64 - 32^2\varepsilon^{-1}\sqrt{M_4(\mu_0)M_4(\mu_1)} + \rho$ for any $\boldsymbol{A}\in\mathbb{R}^{d_0\times d_1}$ and $\rho\geq 0$. When this lower bound is non-negative, Φ is ρ-weakly convex on $\mathbb{R}^{d_0\times d_1}$ by definition. 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 [2, p.356])

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

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. $\square$

A.4 Proof of Theorem 2

In this section, we show that Theorem 2.2 in [13] 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 [13] as the proximal operators

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

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

$$\operatorname*{argmin}_{\boldsymbol{V}\in\mathcal{D}_M}\left\{\|\boldsymbol{V}-\boldsymbol{U}\|_F^2\right\}, \tag{10}$$

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 (10) is given by $\boldsymbol{V}^\star$ defined entrywise by (cf. Section 5.2.2 in [14])

$$\boldsymbol{V}^\star = \frac{M}{2}\mathrm{sign}(\boldsymbol{U})\min\left(\frac{2}{M}|\boldsymbol{U}|,1\right).$$

Next, we show that our notion of δ-oracle yields a δ'-approximate gradient in the sense of Equation (2.3) in [13]. Precisely, we prove that

$$\left|\mathrm{tr}\left(\left(\widetilde{D}\Phi_{[\boldsymbol{A}]} - D\Phi_{[\boldsymbol{A}]}\right)^\intercal(\boldsymbol{B}-\boldsymbol{C})\right)\right| \leq \delta', \tag{11}$$

for any $\boldsymbol{A},\boldsymbol{B},\boldsymbol{C}\in\mathcal{D}_M$. By Hölder's inequality,

$$\left|\mathrm{tr}\left(\left(\widetilde{D}\Phi_{[\boldsymbol{A}]} - D\Phi_{[\boldsymbol{A}]}\right)^\intercal(\boldsymbol{B}-\boldsymbol{C})\right)\right| \leq M\left\|\widetilde{D}\Phi_{[\boldsymbol{A}]} - D\Phi_{[\boldsymbol{A}]}\right\|_1,$$

and the choice $\boldsymbol{B} = -\boldsymbol{C} = \frac{M}{2}\mathrm{sign}\left(\widetilde{D}\Phi_{[\boldsymbol{A}]} - D\Phi_{[\boldsymbol{A}]}\right)$ saturates the above bound. Recall that

$$\widetilde{D}\Phi_{[\boldsymbol{A}]} - D\Phi_{[\boldsymbol{A}]} = 32 \sum_{\substack{1\le i\le N_0 \\ 1\le j\le N_1}} x^{(i)} \left(y^{(j)}\right)^\intercal \left(\widetilde{\boldsymbol{\Pi}}^{\boldsymbol{A}}_{ij} - \boldsymbol{\Pi}^{\boldsymbol{A}}_{ij}\right), \tag{12}$$

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 such that

$$\left\|\widetilde{D}\Phi_{[\boldsymbol{A}]} - D\Phi_{[\boldsymbol{A}]}\right\|_1 \le 32 \left\|\widetilde{\boldsymbol{\Pi}}^{\boldsymbol{A}} - \boldsymbol{\Pi}^{\boldsymbol{A}}\right\|_\infty \sum_{\substack{1\le i\le N_0 \\ 1\le j\le N_1}} \left\|x^{(i)}\left(y^{(j)}\right)^\intercal\right\|_1 < 32\delta \sum_{\substack{1\le i\le N_0 \\ 1\le j\le N_1}} \left\|x^{(i)}\left(y^{(j)}\right)^\intercal\right\|_1,$$

where $|\cdot|$ is applied componentwise in the above display. Combining the displayed equations yields

$$\left|\mathrm{tr}\left(\left(\widetilde{D}\Phi_{[\boldsymbol{A}]} - D\Phi_{[\boldsymbol{A}]}\right)^\intercal (\boldsymbol{B} - \boldsymbol{C})\right)\right| \le 32M\delta \sum_{\substack{1\le i\le N_0 \\ 1\le j\le N_1}} \left\|x^{(i)}\left(y^{(j)}\right)^\intercal\right\|_1 = \delta',$$

proving Eq. (11).

With these preparations Theorem 2 follows from Theorem 2.2 in [13] and the discussion following its proof, noting that $\sum_{i=0}^k \frac{i+1}{2} = \frac{(k+1)(k+2)}{4}$. □

A.5 Proof of Corollary 2

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

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

It remains to bound $\|D\Phi_{[\boldsymbol{A}_k]} - \widetilde{D}\Phi_{[\boldsymbol{A}_k]}\|_F$ using the δ-oracle. By (12),

$$\begin{aligned}
\|D\Phi_{[\boldsymbol{A}]} - \widetilde{D}\Phi_{[\boldsymbol{A}]}\|_F &= 32 \left\| \sum_{\substack{1\le i\le N_0 \\ 1\le j\le N_1}} x^{(i)}\left(y^{(j)}\right)^\intercal \left(\widetilde{\boldsymbol{\Pi}}^{\boldsymbol{A}}_{ij} - \boldsymbol{\Pi}^{\boldsymbol{A}}_{ij}\right)\right\|_F \\
&\le 32 \sum_{\substack{1\le i\le N_0 \\ 1\le j\le N_1}} |\widetilde{\boldsymbol{\Pi}}^{\boldsymbol{A}}_{ij} - \boldsymbol{\Pi}^{\boldsymbol{A}}_{ij}| \left\|x^{(i)}\left(y^{(j)}\right)^\intercal\right\|_F \\
&\le 32 \|\widetilde{\boldsymbol{\Pi}}^{\boldsymbol{A}} - \boldsymbol{\Pi}^{\boldsymbol{A}}\|_\infty \sum_{\substack{1\le i\le N_0 \\ 1\le j\le N_1}} \|x^{(i)}\|\|y^{(j)}\| \\
&< 32\delta \sum_{\substack{1\le i\le N_0 \\ 1\le j\le N_1}} \|x^{(i)}\|\|y^{(j)}\|
\end{aligned}$$

□

B Sharpness of variance bound from Lemma 4

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 $\boldsymbol{C} \in \mathbb{R}^{d_0\times d_1}$,

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

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

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

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

C 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 $\mu_1 = \sum_{j=1}^{N_1} b_j\delta_{x^{(j)}} \in \mathcal{P}(\mathbb{R}^{d_1})$, 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. [12, 19]).

Algorithm 3 Sinkhorn Algorithm

Fix a threshold γ and a maximum iteration number $k_{\max}$.
$v_0 \leftarrow \mathbb{1}_{N_1}/N_1$
$k\leftarrow 1$
repeat
  $u_k \leftarrow a/(\boldsymbol{K}v_{k-1})$
  $v_k \leftarrow b/(\boldsymbol{K}^\intercal u_k)$
  $\boldsymbol{\Pi}^k \leftarrow \mathrm{diag}(u_k)\boldsymbol{K}\mathrm{diag}(v_k)$
  $k\leftarrow k+1$
until $\|\boldsymbol{\Pi}^k\mathbb{1}_{N_1} - a\|_2 < \gamma$ or $k > k_{\max}$
return $\boldsymbol{\Pi}^k$

In Algorithm 3, the division of two vectors is understood componentwise. The stopping condition is based only on one of the marginal constraints as $\mathbb{1}_{N_0}^\intercal\boldsymbol{\Pi}^k = b^\intercal$ by construction.

The following definitions enable describing the convergence properties of Algorithm 3; we follow the approach of [20] with only minor modifications. Let $\mathbb{R}^d_+$ denote the set of vectors with positive entries and, for $x,y\in\mathbb{R}^d_+$ let

$$\mathsf{d}_H(x,y) = \log\max_{1\le i,j\le d}\frac{x_iy_j}{y_ix_j},$$

denote Hilbert's projective metric[3] on $\mathbb{R}^d_+$. By definition,

$$\mathsf{d}_H(x,y) = \mathsf{d}_H(x/y,\mathbb{1}_d), \tag{13}$$

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

$$\begin{aligned} d_H(x,y) &= \log\max_{1\le i,j\le d} e^{w_i+z_j-w_j-z_i},\\ &= \max_{1\le i,j\le d} w_i+z_j-w_j-z_i,\\ &= \max_{1\le i\le d}(\log x_i - \log y_i) - \min_{1\le i\le d}(\log x_i-\log y_i),\\ &= \max_{1\le i\le d}\log\left(\frac{x_i}{y_i}\right) - \min_{1\le i\le d}\log\left(\frac{x_i}{y_i}\right). \end{aligned} \tag{14}$$

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

It was proved in [4, 31] that multiplication with a positive matrix is a strict contraction w.r.t. d_H. Precisely,

$$d_H(\boldsymbol{A}x, \boldsymbol{A}y) \le \lambda(\boldsymbol{A}) d_H(x, y), \tag{15}$$

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

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

Let

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

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

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

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

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

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

$$u^\star_i = a_i e^{\frac{\varphi\left(x^{(i)}\right)}{\varepsilon}}, \quad v^\star_j = b_j e^{\frac{\psi\left(y^{(j)}\right)}{\varepsilon}}.$$

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

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

Lemma 5. *Fix $\delta > 0$. If $\mathsf{d}(\boldsymbol{\Pi}^k, \boldsymbol{\Pi}^\star) \le \delta$, it follows that $\|\boldsymbol{\Pi}^k - \boldsymbol{\Pi}^\star\|_\infty \le e^\delta - 1$.*

Proof. By Lemma 3 in [20], whenever $\mathsf{d}(\boldsymbol{\Pi}^k, \boldsymbol{\Pi}^\star) \le \delta$ it holds that

$$e^{-\delta} - 1 \le \frac{\boldsymbol{\Pi}^\star_{ij}}{\boldsymbol{\Pi}^k_{ij}} - 1 \le e^\delta - 1,$$

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

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

yielding $\|\boldsymbol{\Pi}^\star - \boldsymbol{\Pi}^k\|_\infty \le e^\delta - 1$. □

Towards bounding $\mathsf{d}(\boldsymbol{\Pi}^k, \boldsymbol{\Pi}^\star)$, we first show that the iterates u_k, v_k defined in Algorithm 3 converge to $u^\star, v^\star$ under d_H.

Lemma 6. *Let u_k, v_k be iterates generated by Algorithm 3. Then,*

$$\mathsf{d}_H(u_k, u^\star) \le \lambda(\boldsymbol{K})^{2(k-1)}\mathsf{d}_H(u_1, u^\star)$$
$$\mathsf{d}_H(v_k, v^\star) \le \lambda(\boldsymbol{K})^{2k}\mathsf{d}_H(v_0, v^\star).$$

In particular, $\mathsf{d}_H(u_k, u^\star) \to 0$, $\mathsf{d}_H(v_k, v^\star) \to 0$ as $k \to \infty$.

Proof. The second claim follows from the first and the fact that $\lambda(\boldsymbol{K}) < 1$. To prove the first claim, we have, by definition,

$$\begin{aligned} \mathsf{d}_H(u_{k+1}, u^\star) &= \mathsf{d}_H\left(\frac{a}{\boldsymbol{K}v_k}, \frac{a}{\boldsymbol{K}v^\star}\right) = \mathsf{d}_H\left(\boldsymbol{K}v_k, \boldsymbol{K}v^\star\right) \le \lambda(\boldsymbol{K})\mathsf{d}_H\left(v_k, v^\star\right), \\ \mathsf{d}_H(v_k, v^\star) &= \mathsf{d}_H\left(\frac{b}{\boldsymbol{K}^\intercal u_k}, \frac{b}{\boldsymbol{K}^\intercal u^\star}\right) = \mathsf{d}_H\left(\boldsymbol{K}^\intercal u_k, \boldsymbol{K}^\intercal u^\star\right) \le \lambda(\boldsymbol{K})\mathsf{d}_H\left(u_k, u^\star\right), \end{aligned} \tag{16}$$

as $\lambda(\boldsymbol{K}) = \lambda(\boldsymbol{K}^\intercal)$. Thus, $\mathsf{d}_H(u_{k+1}, u^\star) \le \lambda(\boldsymbol{K})^2\mathsf{d}_H(u_k, u^\star)$ and $\mathsf{d}_H(v_{k+1}, v^\star) \le \lambda(\boldsymbol{K})^2\mathsf{d}_H(v_k, v^\star)$. The conclusion follows by applying these bounds repeatedly. □

Next, we bound the progress of the iterates $\mathbf{\Pi}^k$ to $\mathbf{\Pi}^\star$ under d in terms of $\mathsf{d}_H(u_k, u^\star), \mathsf{d}_H(v_k, v^\star)$.

Lemma 7. *In the setting of Lemma 6, we have that*

$$\mathsf{d}(\mathbf{\Pi}^k, \mathbf{\Pi}^\star) = \mathsf{d}_H(u_k, u^\star) + \mathsf{d}_H(v_k, v^\star) \leq \lambda(\boldsymbol{K})^{2(k-1)}\mathsf{d}_H(u_1, u^\star) + \lambda(\boldsymbol{K})^{2k}\mathsf{d}_H(v_0, v^\star).$$

Proof. The second inequality follows from the first by Lemma 6. By construction,

$$\mathbf{\Pi}^{k+j} = \mathrm{diag}\left(\frac{u_j}{u_k}\right)\mathbf{\Pi}^k \mathrm{diag}\left(\frac{v_j}{v_k}\right),$$

such that

$$\begin{aligned}\mathsf{d}(\mathbf{\Pi}^k, \mathbf{\Pi}^{k+j}) &= \mathsf{d}_H\left(\frac{u_j}{u_k}, \mathbb{1}_{N_0}\right) + \mathsf{d}_H\left(\frac{v_j}{v_k}, \mathbb{1}_{N_1}\right), \\ &= \mathsf{d}_H\left(u_j, u_k\right) + \mathsf{d}_H\left(v_j, v_k\right),\end{aligned}$$

taking the limit as $j \to \infty$ and applying Lemma 6 yields

$$\mathsf{d}(\mathbf{\Pi}^k, \mathbf{\Pi}^\star) = \mathsf{d}_H\left(u_k, u^\star\right) + \mathsf{d}_H\left(v_k, v^\star\right),$$

where we have also used the fact that $\lim_{j\to\infty}\mathsf{d}(\mathbf{\Pi}^k, \mathbf{\Pi}^{k+j}) = \mathsf{d}(\mathbf{\Pi}^k, \mathbf{\Pi}^\star)$ as follows from [20, p. 731] and [35]. □

As $u^\star$ and $v^\star$ are *a priori* unknown, we now bound $\mathsf{d}(u^k, u^\star)$ in terms of $\mathsf{d}(a, u_k \odot \boldsymbol{K}v_k)$, which is a measure of how much $\mathbf{\Pi}^k$ violates the marginal constraint,[4] and another analogous term.

Lemma 8. *In the setting of Lemma 6, we have that*

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

Proof. By construction, we have that

$$\begin{aligned}\mathsf{d}_H(u_k, u^\star) &\leq \mathsf{d}_H(u_{k+1}, u_k) + \mathsf{d}_H(u_{k+1}, u^\star), \\ &\leq \mathsf{d}_H\left(\frac{a}{\boldsymbol{K}v_k}, u_k\right) + \lambda(\boldsymbol{K})^2\mathsf{d}_H(u_k, u^\star), \\ &= \mathsf{d}_H\left(a, u_k \odot \boldsymbol{K}v_k\right) + \lambda(\boldsymbol{K})^2\mathsf{d}_H(u_k, u^\star),\end{aligned}$$

where we have applied the triangle inequality and (16). The claimed result for v_k follows from the same argument. □

Combined, Lemmas 7 and 8 provide an explicit bound on the total number of iterations required to ensure that $\mathsf{d}(\mathbf{\Pi}^k, \mathbf{\Pi}^\star)$ achieves a given precision.

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

$$k \geq 1 + \frac{1}{2\log\left(\lambda(\boldsymbol{K})\right)}\log\left(\frac{\delta(1-\lambda(\boldsymbol{K})^2)}{\mathsf{d}_H(a, u_1 \odot \boldsymbol{K}v_1) + \lambda(\boldsymbol{K})^2\mathsf{d}_H(b, v_0 \odot \boldsymbol{K}^\intercal u_1)}\right),$$

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

Proof. It follows from Lemma 7 and Lemma 8 that

$$\mathsf{d}(\mathbf{\Pi}^k, \mathbf{\Pi}^\star) \leq \frac{\lambda(\boldsymbol{K})^{2(k-1)}\mathsf{d}_H(a, u_1 \odot \boldsymbol{K}v_1) + \lambda(\boldsymbol{K})^{2k}\mathsf{d}_H(b, v_0 \odot \boldsymbol{K}^\intercal u_1)}{1-\lambda(\boldsymbol{K})^2}.$$

The upper bound on the number of iterations required to achieve $\mathsf{d}(\mathbf{\Pi}^k, \mathbf{\Pi}^\star) \leq \delta$ then follows from basic algebra. □

[4] $\mathbf{\Pi}^k\mathbb{1}_{N_1} = u_k \odot \boldsymbol{K}v_k$, where $\odot$ denotes elementwise multiplication.

Now, we demonstrate why the termination condition based on the 2-norm (or an equivalent condition based on the 1-norm) endows us with a δ-oracle approximation and provide a bound on the number of iterations required to achieve it. Theorem 1 in [18] proves that there exists $\bar{k} \le 1 + \frac{R}{\delta}$ satisfying

$$\|u_{\bar{k}} \odot \boldsymbol{K} v_{\bar{k}} - a\|_1 + \|v_{\bar{k}} \odot \boldsymbol{K}^\intercal u_{\bar{k}+1} - b\|_1 \le \delta,$$

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 now bound d_H in terms of the Euclidean distance as to control $\mathsf{d}(\boldsymbol{\Pi}^k, \boldsymbol{\Pi}^\star)$ by $\|u_k \odot \boldsymbol{K} v_k - a\|_2$.

Lemma 9. *Let $r, s \in \mathbb{R}^d_+$ be arbitrary, then*

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

where $i^* \in \operatorname{argmax}_{1\le i\le d} r_i^{-1}(s_i - r_i)$ and $i_* \in \operatorname{argmin}_{1\le i\le d} s_i^{-1}(s_i - r_i)$.

Proof. We have by (14) 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

$$1 - \frac{r_i}{s_i} \le \log\left(\frac{s_i}{r_i}\right) \le \frac{s_i}{r_i} - 1,$$

using the inequalities $\frac{x}{1+x} \le \log(1+x) \le x$ for $x > -1$. Whence,

$$\begin{aligned}
\mathsf{d}_H(s, r) &\le \max_{1\le i\le N_0} r_i^{-1}(s_i - r_i) - \min_{1\le i\le N_0} s_i^{-1}(s_i - r_i) \\
&= r_{i^\star}^{-1}(s_{i^\star} - r_{i^\star}) - s_{i_\star}^{-1}(s_{i_\star} - r_{i_\star}) \\
&\le \left(r_{i^\star}^{-1} + s_{i_\star}^{-1}\right)\|s - r\|_2.
\end{aligned}$$

□

By combining Lemmas 7 and 9 we arrive at the desired result.

Proposition 4. *Let $\underline{a} = \min_{1\le i\le N_0} a_i$. In the setting of Lemma 9, the iterates $\boldsymbol{\Pi}^k$ generated by Algorithm 3 with the threshold $\underline{a} \ge \gamma > 0$ satisfy*

$$\mathsf{d}(\boldsymbol{\Pi}^k, \boldsymbol{\Pi}^\star) \le \frac{E_k\|a - u_k \odot \boldsymbol{K} v_k\|_2}{1 - \lambda(\boldsymbol{K})},$$

where E_k denotes the constant from Lemma 9. Hence, the 2-norm termination criterion in Algorithm 3 is satisfied in $\bar{k}$ iterations for some $\bar{k} \le 1 + \frac{R}{\gamma}$ and

$$\mathsf{d}(\boldsymbol{\Pi}^{\bar{k}}, \boldsymbol{\Pi}^\star) \le \frac{E_{\bar{k}}\gamma}{1 - \lambda(\boldsymbol{K})} \le \frac{\gamma}{1 - \lambda(\boldsymbol{K})}\left(\underline{a}^{-1} + (\underline{a} - \gamma)^{-1}\right).$$

Proof. The first claim follows directly from Lemmas 7 and 9 together with the fact that $\mathsf{d}_H(v_k, v^\star) \le \lambda(\boldsymbol{K})\mathsf{d}_H(u_k, u^\star)$ (see (16)).

It is clear from the discussion preceding Lemma 9 that $\|u_{\bar{k}} \odot \boldsymbol{K} v_{\bar{k}} - a\|_2 \le \gamma$ for some $\bar{k} \le 1 + \frac{R}{\gamma}$, which corresponds to the 2-norm termination condition for Algorithm 3. To see that $E_{\bar{k}} \le \underline{a}^{-1} + (\underline{a} - \gamma)^{-1}$, let $w^k = u_k \odot \boldsymbol{K} v_k$ and observe that $\|a - w^k\|_\infty \le \|a - w^k\|_2 \le \gamma$. Hence, for any index $1 \le i \le N_0$, $a_i - \gamma \le w_i^k$ and, as such, $(w_i^k)^{-1} \le \frac{1}{a_i - \gamma} \le \frac{1}{\underline{a} - \gamma}$. □

The proof of Proposition 2 then follows by combining Propositions 3 and Proposition 4. The maximal number of iterations for Algorithm 3 to output a δ-oracle approximation of the EOT coupling

$\mathbf{\Pi}^{\boldsymbol{A}}$ is thus

$$\tilde{k} = \min\left\{1 + \frac{1}{2\log\left(\lambda(\boldsymbol{K})\right)}\log\left(\frac{\delta(1-\lambda(\boldsymbol{K})^2)}{\mathsf{d}_H(a, u_1 \odot \boldsymbol{K}v_1) + \lambda(\boldsymbol{K})^2\mathsf{d}_H(b, v_0 \odot \boldsymbol{K}^\intercal u_1)}\right),\right.$$
$$\left. 1 - 2\delta^{-1}\log\left(e^{-\|\boldsymbol{C}\|_\infty/\varepsilon}\min_{\substack{1\le i\le N_0\\1\le j\le N_1}} a_i \wedge b_j\right)\right\}. \tag{17}$$

D Convergence of Algorithm 2

In what follows, we slightly adapt the proof of Theorem 2 in [21] 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 [21],

$$\Phi(\boldsymbol{B}_k) \le \Phi(\boldsymbol{A}_k) + \mathrm{tr}\left(D\Phi^\intercal_{[\boldsymbol{A}_k]}(\boldsymbol{B}_k - \boldsymbol{A}_k)\right) + \frac{L}{2}\|\boldsymbol{B}_k - \boldsymbol{A}_k\|_F^2, \tag{18}$$

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 gives

$$\begin{aligned}
&\Phi(\boldsymbol{A}_k) - ((1-\tau_k)\Phi(\boldsymbol{B}_{k-1}) + \tau_k\Phi(\boldsymbol{H}))\\
&= \tau_k\left(\Phi(\boldsymbol{A}_k) - \Phi(\boldsymbol{H})\right) + (1-\tau_k)\left(\Phi(\boldsymbol{A}_k) - \Phi(\boldsymbol{B}_{k-1})\right)\\
&\le \tau_k\left(\mathrm{tr}\left(D\Phi^\intercal_{[\boldsymbol{A}_k]}(\boldsymbol{A}_k - \boldsymbol{H})\right) + \frac{L'}{2}\|\boldsymbol{H} - \boldsymbol{A}_k\|_F^2\right)\\
&+ (1-\tau_k)\left(\mathrm{tr}\left(D\Phi^\intercal_{[\boldsymbol{A}_k]}(\boldsymbol{A}_k - \boldsymbol{B}_{k-1})\right) + \frac{L'}{2}\|\boldsymbol{B}_{k-1} - \boldsymbol{A}_k\|_F^2\right)\\
&= \mathrm{tr}\left(D\Phi^\intercal_{[\boldsymbol{A}_k]}(\boldsymbol{A}_k - \tau_k\boldsymbol{H} - (1-\tau_k)\boldsymbol{B}_{k-1})\right)\\
&+ \frac{L'\tau_k}{2}\|\boldsymbol{H} - \boldsymbol{A}_k\|_F^2 + \frac{L'(1-\tau_k)}{2}\underbrace{\|\boldsymbol{B}_k - \boldsymbol{A}_k\|_F^2}_{\tau_k^2\|\boldsymbol{B}_{k-1} - \boldsymbol{C}_{k-1}\|_F^2},
\end{aligned} \tag{19}$$

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

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

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

As $\boldsymbol{C}_k$ is optimal for the problem $\mathrm{argmin}_{\boldsymbol{V}\in\mathbb{R}^{d_0\times d_1}}\left\{\frac{1}{2\gamma_k}\|\boldsymbol{V} - (\boldsymbol{C}_{k-1} - \gamma_k\boldsymbol{G}_k)\|_F^2 + \mathcal{I}_{\mathcal{D}_M}(\boldsymbol{V})\right\}$, 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 [30]). Thus, for any $\boldsymbol{U}\in\mathbb{R}^{d_0\times d_1}$,

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

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

$$\operatorname{tr}\left(\boldsymbol{G}_k^\intercal(\boldsymbol{C}_k - \boldsymbol{U})\right) \leq \mathcal{I}_{\mathcal{D}_M}(\boldsymbol{U}) + \frac{1}{2\gamma_k}\left(\|\boldsymbol{C}_{k-1} - \boldsymbol{U}\|_F^2 - \|\boldsymbol{C}_k - \boldsymbol{U}\|_F^2 - \|\boldsymbol{C}_k - \boldsymbol{C}_{k-1}\|_F^2\right). \quad (20)$$

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 (20) upon scaling by τ_k,

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

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

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

Combining the equation above with Eq. (18) and Eq. (19) and setting $\boldsymbol{H} = \boldsymbol{U} \in \mathcal{D}_M$ (otherwise the bound is vacuous),

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

where the inequality follows from the δ-oracle which implies the bound (cf. (11))

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

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 [21] yields, for $A_i = \frac{2}{i(i+1)}$,

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

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

$$\begin{aligned}
&\|\boldsymbol{H}-\boldsymbol{A}_i\|_F^2+\tau_i(1-\tau_i)\|\boldsymbol{B}_{i-1}-\boldsymbol{C}_{i-1}\|_F^2 \\
&\le 2\left(\|\boldsymbol{H}\|_F^2+\|\boldsymbol{A}_i\|_F^2+\tau_i(1-\tau_i)\left(\|\boldsymbol{B}_{i-1}\|_F^2+\|\boldsymbol{C}_{i-1}\|_F^2\right)\right) \\
&\le 2\left(\|\boldsymbol{H}\|_F^2+(1-\tau_i)\|\boldsymbol{B}_{i-1}\|_F^2+\tau_i\|\boldsymbol{C}_{i-1}\|_F^2+\tau_i(1-\tau_i)\left(\|\boldsymbol{B}_{i-1}\|_F^2+\|\boldsymbol{C}_{i-1}\|_F^2\right)\right) \\
&\le 2\left(\|\boldsymbol{H}\|_F^2+(1+\tau_i(1-\tau_i))\max_{\mathcal{D}_M}\|\cdot\|_F^2\right) \\
&\le 2\left(\|\boldsymbol{H}\|_F^2+\frac{5}{16}M^2d_0^2d_1^2\right),
\end{aligned}$$

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

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

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

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

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

$$\min_{i=1}^k\left\|\beta_i^{-1}\left(\boldsymbol{B}_i-\boldsymbol{A}_i\right)\right\|_F^2\le\frac{96L^2}{k(k+1)(k+2)}\|\boldsymbol{C}_0-\boldsymbol{B}^\star\|_F^2+8L\delta'+\frac{24LL'}{N}\left(\|\boldsymbol{B}^\star\|_F^2+\frac{5M^2d_0^2d_1^2}{16}\right).$$

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.

E Additional Results

E.1 Proof of Lemma 1

The proof of Lemma 1 follows from the following lemma coupled with the chain rule for Fréchet differentiable maps.

Lemma 10. *Let $\mu_i\in\mathcal{P}(\mathbb{R}^{d_i})$, for $i=0,1$, be compactly supported with $\operatorname{spt}(\mu_i)=S_i$. 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

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

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. $\|\xi\|_{\infty,S_0\times S_1} \le \|h\|_{\infty,S_0\times S_1}$. That is,

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

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 1. 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 10 by applying the chain rule. □

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

Lemma 11 (Example 2 in [42])**.** *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

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

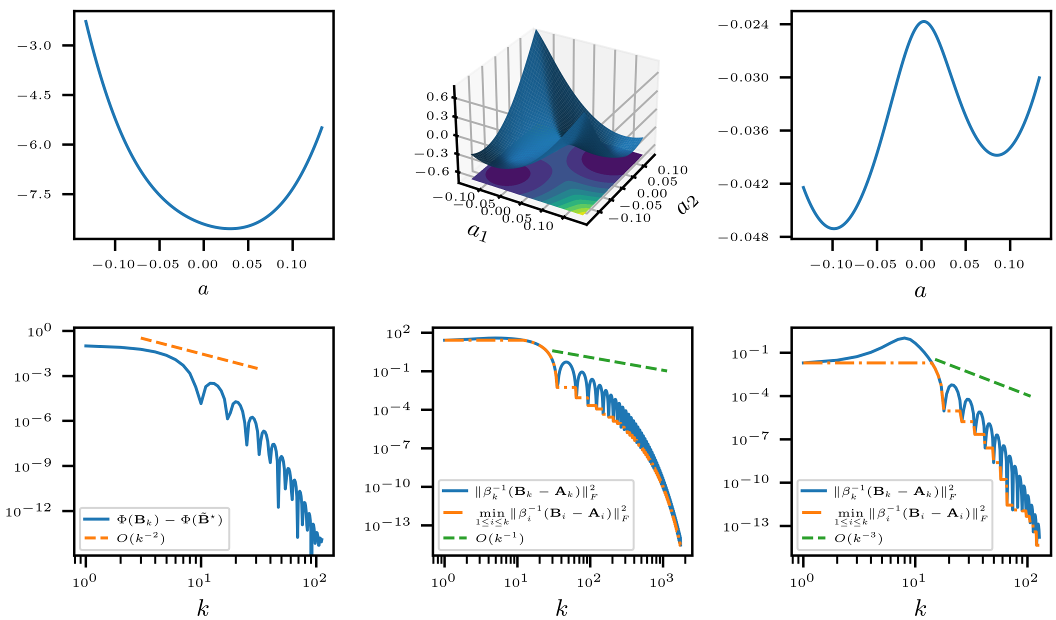
is a compact operator for $\xi : (x,y) \in \mathbb{R}^{d_0} \times \mathbb{R}^{d_1} \mapsto e^{\frac{\varphi_0^{\boldsymbol{A}}(x)+\varphi_1^{\boldsymbol{A}}(y)-c_{\boldsymbol{A}}(x,y)}{\varepsilon}}$. For any $y \in \mathbb{R}^{d_1}$ and $f \in L^2(\mu_0)$, $|\mathcal{L}_2(f)(y)|^2 \le \|f\|^2_{L^2(\mu_0)}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

Let f_n be a bounded sequence in $L^2(\mu_0)$. By the Eberlein-Šmulian theorem [42, 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. □

F Blown-up figures



(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.

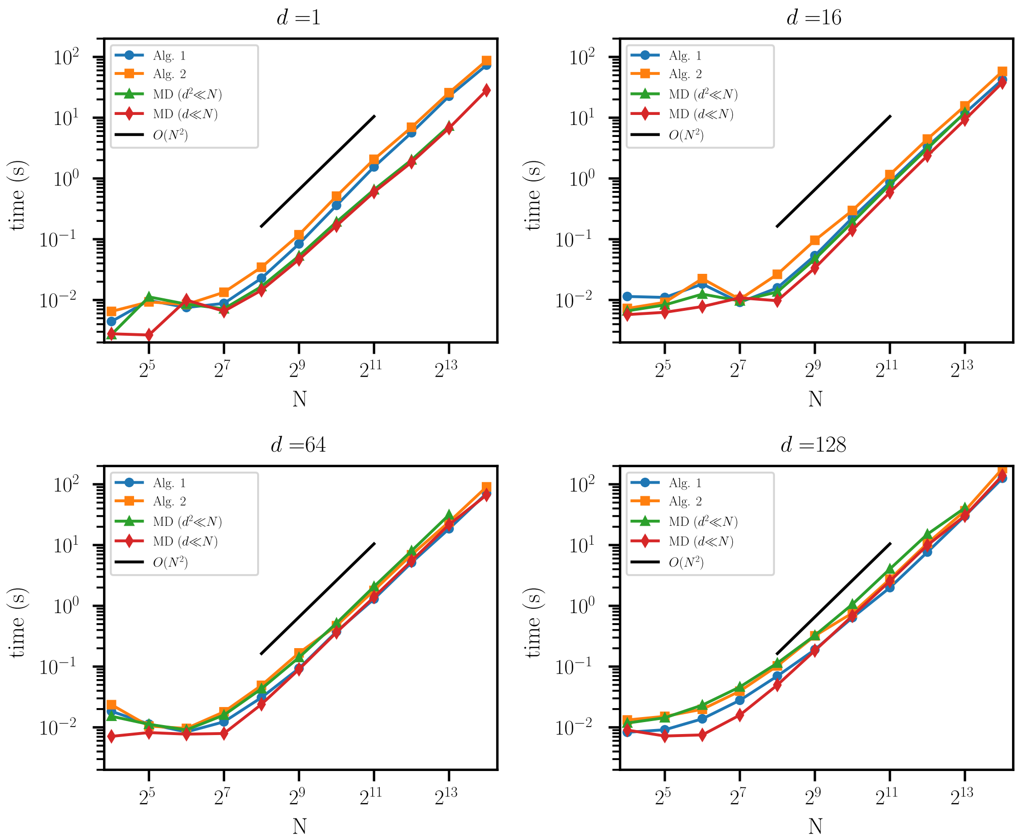


Figure 2: The various plots compile the average runtime of Algorithms 1 and 2, and two versions of the mirror descent algorithm in the convex regime for different combinations of d and N.

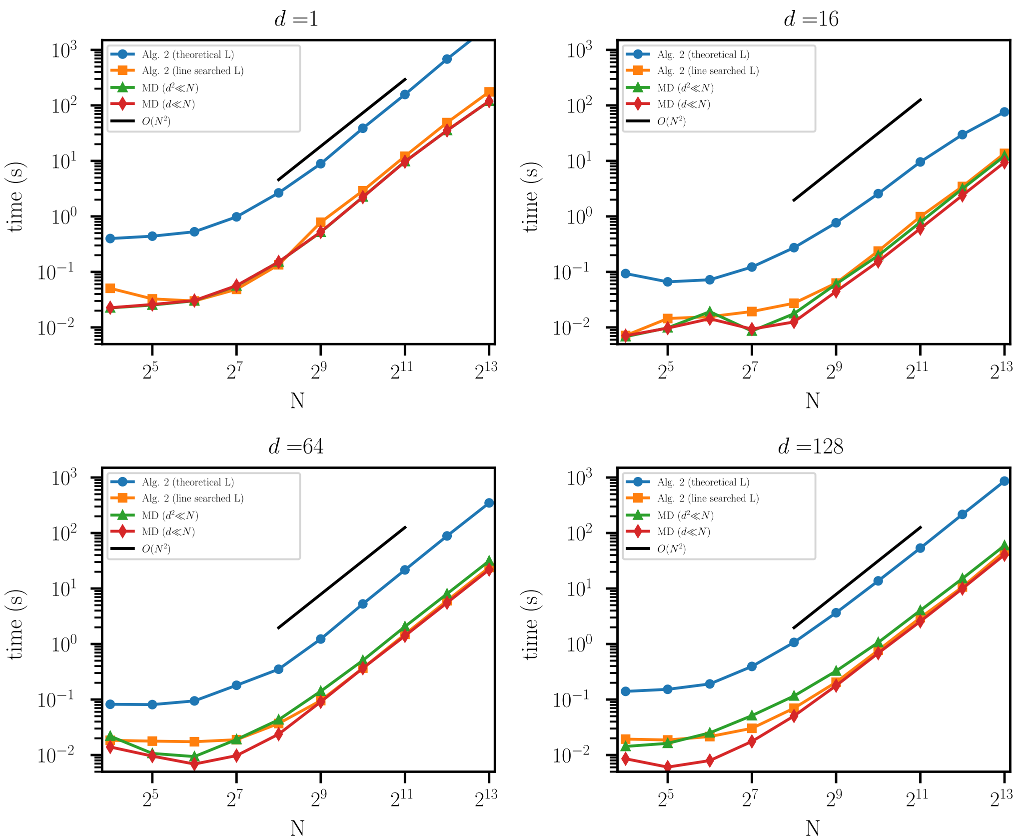


Figure 3: The various plots compile the average runtime of Algorithm 2 with the two methods for choosing L, and two versions of the mirror descent algorithm in the non-convex regime for different combinations of d and N.