Designing Algorithms for Entropic Optimal Transport from an Optimisation Perspective

Vishwak Srinivasan VISHWAKS@MIT.EDU
Department of EECS, MIT

Qijia Jiang QJANG@UCDAVIS.EDU
Department of Statistics, UC Davis

Abstract

In this work, we study and develop a collection of methods for the entropy-regularised optimal transport (eOT) problem through the lens of optimisation. These methods are proposed as optimisation methods for the semi-dual of the eOT problem [8], and gives rise to two broad classes of methods for solving this problem: either based on solving a non-convex constrained problem over a structured subset of couplings, or by working directly with the semi-dual. This is inspired by recent optimisation interpretations of the Sinkhorn algorithm – the de-facto method for solving the eOT problem. By adopting this new viewpoint, we obtain non-asymptotic rates of convergence for the methods studied in this work (old and new) and particularly under minimal assumptions on the problem structure.

1. Introduction

Given two probability distributions μ and ν over $\mathcal{X}, \mathcal{Y} \subseteq \mathbb{R}^d$ respectively, the optimal transport (OT) problem concerns finding an *optimal* map that transforms samples from one to another. Recent advances in computing resources has renewed interest in the OT problem, both in the design of approximate methods for this problem suited for large-scale settings [23], and in the development of a theoretical understanding of its properties [26, 29]. The "optimality" in the OT problem is defined in terms of a cost function $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, and the optimal value of the problem results in a notion of discrepancy between μ and ν that complements information-theoretic discrepancy measures like the total variation distance or the Kullback-Leibler (KL) divergence. Formally, let $\Pi(\mu, \nu)$ be the set of all joint distributions whose marginals are μ and ν. The Kantorovich formulation of the OT problem is given by the following program:

$$\inf_{\pi \in \Pi(\mu,\nu)} \iint c(x, y) \mathrm{d}\pi(x, y) =: \mathrm{OT}(\mu, \nu; c) \ . \tag{1}$$

In modern machine learning and statistics however, methods based on solving the exact OT problem have not seen widespread use owing to both computational and statistical reasons resulting primarily from the high-dimensional nature of the domains involved i.e., of $\mathcal{X}$ and $\mathcal{Y}$. These bottlenecks can surprisingly be alleviated by adding an entropy regularisation to the OT problem, referred to as the entropic optimal transport (eOT) problem. More precisely, for a regularisation parameter $\varepsilon > 0$,

$$\inf_{\pi \in \Pi(\mu,\nu)} \iint c(x,y) \mathrm{d}\pi(x,y) + \varepsilon \cdot \mathrm{d}_{\mathrm{KL}}(\pi \| \mu \otimes \nu) =: \mathrm{OT}_{\varepsilon}(\mu,\nu;c) \ . \tag{2}$$

In Eq. (2), $\mathrm{d}_{\mathrm{KL}}(\pi \| \mu \otimes \nu)$ is the KL divergence between π and the product distribution $\mu \otimes \nu$. A popular algorithm for solving this problem is the *Sinkhorn* algorithm [27], and this was recently popularised by Cuturi [7] by demonstrating it as a viable solution for solving the eOT problem over large datasets, and by extension for solving the OT problem approximately.

Of relevance to this paper is a more recent interpretation of the Sinkhorn algorithm as an *infinite-dimensional optimisation procedure*. This has proven instrumental in obtaining assumption-free guarantees for the Sinkhorn algorithm and has led to a growing body of literature since its inception [2, 9, 14, 18, 24]. In this work, we draw inspiration from this refreshing viewpoint, and propose a collection of new methods for the eOT problem that have provable guarantees. We do so by considering the *semi-dual formulation* of the eOT problem, which is a *concave unconstrained program* (see Section 2 for more details). We find that a gradient ascent procedure for maximising the semi-dual [12] and the Sinkhorn algorithm are special cases of an abstract class of methods that can be seen as iteratively minimising the discrepancy between ν and the $\mathcal{Y}$-marginal of a joint distribution of a certain form (Φ-match). We extend prior *primal* interpretations of the Sinkhorn algorithm [2, 24] for these class of methods, which along with relative smoothness properties contribute to clean, non-asymptotic guarantees that scale as $1/N$ to obtain the eOT coupling for a subclass of Φ-match methods that are related to the maximum mean discrepancy [13].

We complement this *primal interpretation* by adopting principles from finite-dimensional optimisation to propose new methods based on growth conditions of the semi-dual objective, thereby extending the Φ-match class of methods for the eOT problem. These lead to methods that provably optimise the semi-dual (and consequently solving the eOT problem) and converge at a rate that scales $\frac{1}{N}$ under minimal assumptions. We find that one of these methods can be accelerated and converge at a faster rate that scales as $\frac{1}{N^2}$, analogous to FISTA for ISTA [3].

2. Background

The eOT problem defined in Eq. (2) is a (strictly) convex minimisation problem, and under certain regularity conditions, this problem admits a unique solution $\pi^\star \in \Pi(\mu,\nu)$ [15, 22] of the form

$$\mathrm{d}\pi^\star(x,y) = \exp\left(\phi^\star(y) - \psi^\star(x) - \frac{c(x,y)}{\varepsilon}\right) \mathrm{d}\mu(x)\mathrm{d}\nu(y) \tag{3}$$

where $\psi^\star$ and $\phi^\star$ are called *Schrödinger potentials*. The eOT problem has a dual formulation with zero duality gap, and is *unconstrained*:

$$\begin{aligned}\mathrm{OT}_{\varepsilon}(\mu,\nu;c) = \varepsilon \cdot \sup \Bigg\{ &\int \phi(y)\mathrm{d}\nu(y) - \int \psi(x)\mathrm{d}\mu(x) \\ &- \log \iint \exp\left(\phi(y) - \psi(x) - \frac{c(x,y)}{\varepsilon}\right)\mathrm{d}\mu(x)\mathrm{d}\nu(y)\Bigg\} \ . \end{aligned}\tag{4}$$

Any solution of the dual problem above corresponds to a pair of Schrödinger potentials and vice versa [21, Thm. 3.2]. Note that $\mathrm{OT}_{\varepsilon}(\mu,\nu;c) = \varepsilon \cdot \mathrm{OT}_1(\mu,\nu;c/\varepsilon)$, and hence without loss of generality, we focus on $\mathrm{OT}_1(\mu, \nu; {}^{c}\!/\!_{\varepsilon})$ in the rest of this work. We denote the objective in the dual form of $\mathrm{OT}_1(\mu, \nu; {}^{c}\!/\!_{\varepsilon})$ by $D(\psi, \phi)$ and use $\pi^\star$ to denote the (primal) solution of $\mathrm{OT}_1(\mu, \nu; {}^{c}\!/\!_{\varepsilon})$.

The semi-dual problem Originally discussed in [12], this problem is motivated for the use of stochastic algorithms to solve the eOT problem in the setting where $\mathcal{X}$ and $\mathcal{Y}$ are discrete spaces, and later work [7] examines this in more detail while predominantly focusing on the discrete setting.

In the notation of Léger [14], define the operation $\phi \mapsto \phi^+$ for $\phi \in L^1(\nu)$ as $\phi^+(x) := \log \mathbb{E}_{y\sim\nu}\left[\exp\left(\phi(y) - \frac{c(x,y)}{\varepsilon}\right)\right]$. By partially maximising over ψ in Eq. (4) (without loss of generality), we obtain the *semi-dual problem* for eOT, whose objective J satisfies for any $\phi \in L^1(\nu)$ that $J(\phi) := D(\phi^+, \phi) = \mathbb{E}_\nu[\phi] - \mathbb{E}_\mu[\phi^+]$. To translate a dual potential $\phi \in L^1(\nu)$ to a joint distribution over $\mathcal{X} \times \mathcal{Y}$, define the joint distribution $\pi(\phi, \phi^+)$ with density w.r.t. $\mu \otimes \nu$ as

$$\mathrm{d}\pi(\phi, \phi^+)(x, y) = \exp\left(\phi(y) - \phi^+(x) - \frac{c(x,y)}{\varepsilon}\right) \mathrm{d}\nu(y)\mathrm{d}\mu(x)\ . \tag{5}$$

This is a valid probability density function over $\mathcal{X} \times \mathcal{Y}$ as evidenced by the fact that its $\mathcal{X}$-marginal is always μ. Recall that this joint distribution with $\phi \leftarrow \phi^\star$ in Eq. (5) corresponds to the unique solution of the eOT problem in Eq. (3) by virtue of $(\phi^\star)^+ = \psi^\star$. The collection of all such π in Eq. (5) formed by $\phi \in L^1(\nu)$ is referred to as $\mathcal{Q}$.

We henceforth assume that μ and ν have densities w.r.t. the Lebesgue measure. The first variation δJ of the semi-dual J can be succinctly expressed in terms of the marginal $\pi(\phi, \phi^+)_\mathcal{Y}$ [14, Lem. 1]: for any $\phi \in L^1(\nu)$, $\delta J(\phi)(y) = \nu(y) - \pi(\phi, \phi^+)_\mathcal{Y}(y)$. Now we note two important properties of the semi-dual J: the Bregman divergence induced by J is non-positive, or in other words that J is a concave functional (which was informally stated in prior work, see [14, Pg. 5] for instance), and is bounded "quadratically" from below, yielding the regularity condition that the semi-dual J does not grow arbitrarily quickly (which is a new finding in this work, specifically without assumptions on μ and ν). These are summarised in the following lemma.

Lemma 1 *Let* $\phi, \overline{\phi} \in L^1(\nu)$. *Then,* $-\frac{\|\overline{\phi}-\phi\|^2_{L^\infty(\mathcal{Y})}}{2} \leq J(\overline{\phi}) - J(\phi) - \left\langle \delta J(\phi), \overline{\phi} - \phi \right\rangle_{L^2(\mathcal{Y})} \leq 0.$

3. From semi-dual gradient ascent to a new class of methods for eOT

Lemma 1 implies that one could ostensibly use a gradient ascent-like procedure to find a maximiser of the semi-dual J and consequently solve the eOT problem to within a desired tolerance. More precisely, from an initialisation $\phi^0 \in L^1(\nu)$, we can obtain a sequence of iterates $\{\phi^n\}_{n\geq 1}$ based on the recursion

$$\phi^{n+1} = \mathsf{M}^{\mathsf{SGA}}(\phi^n; \eta) := \phi^n + \eta \cdot \delta J(\phi^n)\ . \tag{SGA}$$

The update $\mathsf{M}^{\mathsf{SGA}}$ was previously considered by [12] for discrete spaces $\mathcal{X}$ and $\mathcal{Y}$, where ϕ can be represented as a finite-dimensional vector. In this setting, Lemma 1 implies a standard notion of smoothness for the semi-dual J [20, Chap. 2] by the monotonicity of finite-dimensional norms. This consequently results in an *assumption-free* non-asymptotic convergence guarantee for SGA with $\eta < 2$. When generalising to continuous spaces, a temporary setback towards establishing such rates of convergence for this update is that $\|\phi\|_{L^\infty(\mathcal{Y})} \leq \|\phi\|_{L^2(\mathcal{Y})}$ is not generally true, and we rectify this by instead adopting an alternate perspective on SGA as minimising the "discrepancy" between the $\mathcal{Y}$-marginal of $\pi(\phi, \phi^+)$ and ν. From the form of $\delta J(\phi)$, we have

$$\mathsf{M}^{\mathsf{SGA}}(\phi; \eta) = \phi + \eta \cdot \left(\nu - \pi(\phi, \phi^+)_{\mathcal{Y}}\right) \ . \tag{6}$$

Note that a fixed point $\phi^\star$ of this update satisfies $\pi(\phi^\star, (\phi^\star)^+)_{\mathcal{Y}} = \nu$. This also corresponds to a maximiser of J since $\pi(\phi^\star, (\phi^\star)^+)_{\mathcal{Y}} = \nu$ which is equivalent to $\delta J(\phi^\star) = 0$. More generally, we can define a class of updates described below formed by composition $\pi(\phi, \phi^+)_{\mathcal{Y}}$ and ν with a map $\Phi : L^1(\mathcal{Y}) \to L^\infty(\mathcal{Y})$ called $\mathsf{M}^{\Phi\text{-match}}$ defined as

$$\boxed{\mathsf{M}^{\Phi\text{-match}}(\phi; \eta) = \phi - \eta \cdot \left(\log \Phi(\pi(\phi, \phi^+)_{\mathcal{Y}}) - \log \Phi(\nu)\right) \ . \qquad (\Phi\text{-match})}$$

From this, we see that (1) when $\Phi(f) : f \mapsto e^f$, this recovers SGA; and (2) when $\Phi(f) : f \mapsto f$, this recovers the dual version of the Sinkhorn algorithm [2] with a general step size η (also called η-Sinkhorn in [24]). In the following subsection, we provide an interpretations of Φ-match over the space of distributions $\pi \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ as opposed to dual potentials $\phi \in L^1(\nu)$ as we do here.

Φ-match as a local greedy method For a map $\mathcal{F} : L^1(\mathcal{X} \times \mathcal{Y}) \to L^\infty(\mathcal{X} \times \mathcal{Y})$:

$$\mathsf{root}_{\mathcal{X},\mu}(\pi; \mathcal{F}, \eta) := \operatorname*{argmin}_{\bar{\pi}} \left\{\langle \mathcal{F}(\pi), \bar{\pi} - \pi \rangle_{L^2(\mathcal{Y})} + \eta^{-1} \cdot \mathrm{d}_{\mathrm{KL}}(\bar{\pi} \| \pi) : \bar{\pi}_{\mathcal{X}} = \mu\right\} \ . \tag{7}$$

When $\mathcal{F}$ is the first variation of a functional that measures the discrepancy between the $\mathcal{Y}$-marginal of π and ν, $\mathsf{root}_{\mathcal{X},\mu}(\pi; \mathcal{F}, \eta)$ can be viewed as minimising a local first-order approximation of this functional, thereby approximately matching the $\mathcal{Y}$-marginal and enforcing the $\mathcal{X}$-marginal to be μ.

In the following lemma, we show that the sequence of joint distributions obtained by updating ϕ using Φ-match can be viewed as iteratively solving the problem Eq. (7) for $\mathcal{F} \leftarrow \mathcal{V}_\Phi$ where $\mathcal{V}_\Phi(\pi)(x, y) := \log \Phi(\pi_{\mathcal{Y}})(y) - \log \Phi(\nu)(y)$. By the structure of the joint distribution induced by the potential (Eq. (5)), these are guaranteed to lie in $\mathcal{Q}$.

Theorem 1 *Let $\phi^0 \in L^1(\nu)$ and let $\{\phi^n\}_{n \geq 1}$ be the sequence of potentials obtained as $\phi^{n+1} = \mathsf{M}^{\Phi\text{-match}}(\phi^n; \eta)$ for $\eta \in [0, 1]$. Then, the sequence of distributions $\{\pi^n\}_{n \geq 0}$ where $\pi^n = \pi(\phi^n, (\phi^n)^+)$ satisfy for every $n \geq 0$, $\boxed{\pi^{n+1} = \mathsf{root}_{\mathcal{X},\mu}(\pi^n; \mathcal{V}_\Phi, \eta)}$.*

When $\Phi : f \mapsto f$, $\mathcal{V}_\Phi$ is $\log \frac{\pi_{\mathcal{Y}}}{\nu}$ in this case, which corresponds to the first variation of the functional $\rho \mapsto \mathrm{d}_{\mathrm{KL}}(\rho_{\mathcal{Y}} \| \nu)$, and the equivalence in Theorem 1 for the Sinkhorn method was originally derived in [2] and generalised to an arbitrary step size $\eta \in (0, 1)$ in [24]. In the case where $\Phi : f \mapsto e^f$, the map $\mathcal{V}_\Phi$ here is $\pi_{\mathcal{Y}} - \nu$, which is the first variation of another notion of discrepancy between $\pi_{\mathcal{Y}}$ and ν given by a *Maximum Mean Discrepancy* (abbrev. MMD) [13]. Specifically, this coincides with the first variation of

$$\rho \mapsto \mathcal{L}_{k_{\mathrm{Id}}}(\rho_{\mathcal{Y}}; \nu) := \frac{1}{2}\mathrm{MMD}_{k_{\mathrm{Id}}}(\rho_{\mathcal{Y}}, \nu)^2$$

for the identity kernel k_{Id} defined as $k_{\mathrm{Id}}(y, y') = 1$ iff $y = y'$. More generally, the first variation of $\xi \mapsto \mathcal{L}_k(\xi; \rho) := \frac{1}{2}\mathrm{MMD}_k(\xi, \rho)^2$ is given by $\mathfrak{m}_k(\xi) - \mathfrak{m}_k(\rho)$ [19, Lem. 1] where $\mathfrak{m}_k(\rho)(y) = \int k(y, y')\mathrm{d}\rho(y')$, and notably for characteristic kernels (such as the identity and Gaussian kernels), the map is $\mathfrak{m}_k$ one-to-one. This motivates the consideration of the kernel SGA update $\mathsf{M}^{k\text{-SGA}}$ for iteratively minimising $\mathcal{L}_k(\cdot; \nu)$ for any characteristic kernel k, defined as

$$\mathsf{M}^{k\text{-SGA}}(\phi; \eta) := \phi + \eta \cdot \left\{\mathfrak{m}_k(\nu) - \mathfrak{m}_k(\pi(\phi, \phi^+)_{\mathcal{Y}})\right\} \ . \tag{k-SGA}$$

3.1. Rates of convergence for k-SGA

Due to the generality of the abstraction Φ-match, we can also view k-SGA as an instance of Φ-match with the choice $\Phi_k(f) : f \mapsto e^{\mathfrak{m}_k(f)}$ where we overload $\mathfrak{m}_k(f) := \int_{\mathcal{Y}} k(y, y') \cdot f(y')\mathrm{d}y'$. As a consequence, Theorem 1 shows that the sequence of iterates $\{\phi^n\}_{n\geq 1}$ formed by k-SGA results in a sequence of distributions $\{\pi(\phi^n, (\phi^n)^+)\}_{n\geq 1}$ formed by iteratively applying $\mathsf{root}_{\mathcal{X},\mu}$ with $\mathcal{F} \leftarrow \mathcal{V}_{\Phi_k}$. Crucially, when k is a bounded kernel, the functional $\mathcal{L}_k(.;\nu)$ is *convex* and *relatively smooth* w.r.t. the entropy functional [2], and this leads to the following non-asymptotic guarantee for k-SGA for such kernels.

Theorem 2 *Let $\phi^0 \in L^1(\nu)$, and let $\{\phi^n\}_{n\geq 1}$ be the sequence of potentials obtained as $\phi^{n+1} = \mathsf{M}^{k\text{-SGA}}\left(\phi^n; \min\left\{\frac{1}{2c_k}, 1\right\}\right)$ for a bounded, positive-definite kernel k. Define the sequence of distributions $\{\pi^n\}_{n\geq 1}$ where $\pi^n = \pi(\phi^n, (\phi^n)^+)$. Then, for any $N \geq 1$*

$$\mathcal{L}_k(\pi^N_{\mathcal{Y}}; \nu) \leq \frac{\max\{2c_k, 1\}}{N} \cdot \mathrm{d}_{\mathrm{KL}}(\pi^\star \| \pi^0)\,. \tag{8}$$

When the kernel k is characteristic, Theorem 2 shows that $\pi^n_{\mathcal{Y}}$ approaches ν and consequently establishes a rate of convergence to the optimal coupling $\pi^\star$ since $\pi^n \in \mathcal{Q}$ by construction. As a result, this also gives first known *assumption-free non-asymptotic convergence guarantee* for SGA which matches the $^1\!/_N$ rate in the discrete setting. This highlights the benefit of considering an alternate (primal) viewpoint to analyse a method that is motivated by direct (dual) concave optimisation.

4. Adapting to smoothness of the semi-dual J

As mentioned previously, the fundamental obstacle to establishing guarantees for SGA is the mismatch between the inner product in the Bregman divergence (which is in $L^2(\mathcal{Y})$) and the squared growth (which is in $L^\infty(\mathcal{Y})$) from Lemma 1. We find that the semi-dual also satisfies a different notion of smoothness, but non-uniformly depending on the "size" of the domain considered.

Lemma 2 *Let $\phi, \overline{\phi} \in L^2(\nu)$ be such that $\|\phi\|_{L^\infty(\mathcal{Y})}, \|\overline{\phi}\|_{L^\infty(\mathcal{Y})} \leq B$ for a given $B > 0$. Assume that the cost $c(\cdot, \cdot) \geq 0$. Then,*

$$J(\overline{\phi}) - J(\phi) - \langle \delta J(\phi), \overline{\phi} - \phi \rangle \geq -\frac{\lambda(B) \cdot \|\overline{\phi} - \phi\|^2_{L^2(\nu)}}{2}\;;\;\; \lambda(B) = e^{2B} \cdot \mathbb{E}_{(x,y')\sim\mu\otimes\nu}\left[\exp\left(\frac{c(x,y')}{\varepsilon}\right)\right].$$

The domain $\mathcal{S}_B = \{\phi : \|\phi\|_{L^\infty(\mathcal{Y})} \leq B\}$ considered in the lemma above is of particular interest when the cost function c is uniformly bounded; a result from [10] states that the Schrödinger potentials belong in such $\mathcal{S}_B$ where B depends on the bound on c.

Here, we leverage the smoothness properties in Lemmas 1 and 2 to propose two new methods apart from SGA (and more generally Φ-match) - sign-SGA and proj-SGA.

$$\boxed{\left.\begin{aligned} \phi^{n+1/2} &= \mathsf{M}^{\text{sign-SGA}}(\phi^n; \eta) := \phi^n + \eta \cdot \|\delta J(\phi^n)\|_{L^1(\mathcal{Y})} \cdot \mathrm{sign}(\delta J(\phi^n)) \\ \phi^{n+1} &= \phi^{n+1/2} - (\phi^{n+1/2}(y_{\mathsf{anc}}) - \phi^n(y_{\mathsf{anc}})) \cdot \mathbf{1} \end{aligned}\right\}} \tag{sign-SGA}$$

$$\boxed{\phi^{n+1} = \mathsf{M}^{\text{proj-SGA}}_{\mathcal{S}_B}(\phi; \eta) := \underset{\overline{\phi}\in\mathcal{S}_B}{\mathrm{argmin}} \left\|\overline{\phi} - \left(\phi + \eta \cdot \frac{\delta J(\phi)}{\nu}\right)\right\|^2_{L^2(\nu)}} \tag{proj-SGA}$$

Both of these methods result from maximising the "quadratic" lower bounds on the semi-dual at each iteration, akin to how gradient descent for smooth functions can be seen as minimising its quadratic upper bound. These are provably ascent methods: from a suitable ϕ (either in $L^1(\nu) \cap L^\infty(\mathcal{Y})$ for sign-SGA, or from $\mathcal{S}_B$ for proj-SGA), we have $J(\mathsf{M}(\phi; \eta)) \geq J(\phi)$ for $\mathsf{M} \in \{\mathsf{M}^{\text{sign-SGA}}, \mathsf{M}^{\text{proj-SGA}}_{\mathcal{S}_B}\}$ for sufficiently small $\eta > 0$. More precisely, we can also give the following non-asymptotic convergence guarantees, as stated in the theorems below.

Theorem 3 (Informal) *Let $\phi^0 \in L^1(\nu) \cap L^\infty(\mathcal{Y})$, $y_{\mathrm{anc}} \in \mathcal{Y}$. Consider the sequence of potentials $\{\phi^n\}_{n\geq 1}$ generated according to* sign-SGA *with $\eta = 1$. Then, there exists $C > 0$ depending on ϕ^0, y_{anc} such that for all $N \geq 1$,*

$$J(\phi^N) - J(\phi^\star) \geq -\frac{C}{N+1} \,.$$

Theorem 4 (Informal) *Suppose $c(\cdot, \cdot)$ is a non-negative cost function such that $\lambda(B) < \infty$ (*Lemma 2*) and $\phi^0 \in \mathcal{S}_B$. Consider the sequence of potentials $\{\phi^n\}_{n\geq 1}$ generated according to* proj-SGA *with $\eta = \lambda(B)^{-1}$. Then, there exists $C > 0$ depending on ϕ^0 such that for all $N \geq 1$,*

$$J(\phi^N) - \min_{\phi \in \mathcal{S}_B} J(\phi) \geq -\frac{C \cdot \lambda(B)}{N} \,.$$

Note that the growth condition in Lemma 2 that proj-SGA is conceptually based on is given in terms of $L^2(\nu)$ which is a Hilbert space. This inspires us to adopt the structure of FISTA [3] to design an accelerated version of proj-SGA, which would lead to a rate of convergence that scales as $1/N^2$. This method is based on minor adjustments to handle the non-uniform growth condition in Lemma 2, and is presented below. For initial values $\phi^1 = \overline{\phi}^0 \in \mathcal{S}_B$, and $t_1 = 1$:

$$\boxed{\left.\begin{aligned} &\overline{\phi}^n = \mathsf{M}^{\text{proj-SGA}}_{\mathcal{S}_B}(\phi^n; \lambda(3B)^{-1}) \,; \;\; t_{n+1} = \frac{1 + \sqrt{1 + 4t_n^2}}{2} \,; \\ &\phi^{n+1} = \overline{\phi}^n + \left(\frac{t_n - 1}{t_{n+1}}\right) \cdot (\overline{\phi}^n - \overline{\phi}^{n-1}) \,. \end{aligned}\right\}} \qquad \text{(proj-SGA++)}$$

Theorem 5 (Informal) *Consider the setting of* Theorem 4*, and let $\{\phi^n\}_{n\geq 2}, \{\overline{\phi}^n\}_{n\geq 1}$ be generated according to* proj-SGA++*. Then, there exists $C > 0$ depending on ϕ^0 such that for all $N \geq 1$,*

$$J(\overline{\phi}^N) - \min_{\phi \in \mathcal{S}_B} J(\phi) \geq -\frac{C \cdot \lambda(3B)}{(N+1)^2} \,.$$

When B is sufficiently large (for e.g., as suggested in [10] for bounded costs), then the minimiser of J over $\mathcal{S}_B$ would coincide with $\phi^\star$ as noted previously, which implies that $\min_{\phi \in \mathcal{S}_B} J(\phi) = J(\phi^\star)$. The more detailed versions of Theorems 3 to 5 are given in the Appendix with their proofs. We conclude this discussion by drawing a comparison to *how* the rates for SGA, sign-SGA, and proj-SGA have been established. Specifically, the techniques used to prove Theorems 3 to 5 are based use more standard optimization arguments (generalized to ∞-dimensional space), and are fundamentally different from Theorem 2. This highlights the benefits of considering alternate viewpoints for establishing provable guarantees.

References

[1] Jason Altschuler, Jonathan Niles-Weed, and Philippe Rigollet. Near-linear time approximation algorithms for optimal transport via Sinkhorn iteration. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

[2] Pierre-Cyril Aubin-Frankowski, Anna Korba, and Flavien Léger. Mirror Descent with Relative Smoothness in Measure Spaces, with application to Sinkhorn and EM. In *Advances in Neural Information Processing Systems*, volume 35, pages 17263–17275, 2022.

[3] Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

[4] Alberto Chiarini, Giovanni Conforti, Giacomo Greco, and Luca Tamanini. A semiconcavity approach to stability of entropic plans and exponential convergence of Sinkhorn's algorithm. *arXiv preprint arXiv:2412.09235*, 2024.

[5] Lénaïc Chizat, Alex Delalande, and Tomas Vaškevičius. Sharper Exponential Convergence Rates for Sinkhorn's Algorithm in Continuous Settings. *arXiv preprint arXiv:2407.01202*, 2024.

[6] Giovanni Conforti, Daniel Lacker, and Soumik Pal. Projected Langevin dynamics and a gradient flow for entropic optimal transport. *arXiv preprint arXiv:2309.08598*, 2023.

[7] Marco Cuturi. Sinkhorn distances: Lightspeed Computation of Optimal Transport. In *Advances in Neural Information Processing Systems*, volume 26, 2013.

[8] Marco Cuturi and Gabriel Peyré. Semidual regularized optimal transport. *SIAM Review*, 60 (4):941–965, 2018.

[9] Nabarun Deb, Young-Heon Kim, Soumik Pal, and Geoffrey Schiebinger. Wasserstein mirror gradient flow as the limit of the Sinkhorn algorithm. *arXiv preprint arXiv:2307.16421*, 2023.

[10] Simone Di Marino and Augusto Gerolin. An optimal transport approach for the Schrödinger bridge problem and convergence of Sinkhorn algorithm. *Journal of Scientific Computing*, 85 (2), 2020.

[11] Pavel Dvurechensky, Alexander Gasnikov, and Alexey Kroshnin. Computational Optimal Transport: Complexity by Accelerated Gradient Descent Is Better Than by Sinkhorn's Algorithm. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1367–1376, 2018.

[12] Aude Genevay, Marco Cuturi, Gabriel Peyré, and Francis Bach. Stochastic Optimization for Large-scale Optimal Transport. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

[13] Arthur Gretton, Karsten Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alex Smola. A kernel method for the two-sample-problem. In *Advances in Neural Information Processing Systems*, volume 19. MIT Press, 2006.

[14] Flavien Léger. A gradient descent perspective on Sinkhorn. *Applied Mathematics and Optimization*, 84(2):1843–1855, 2021.

[15] Christian Léonard. A survey of the Schrödinger problem and some of its connections with optimal transport. *Discrete and Continuous Dynamical Systems. Series A*, 34(4):1533–1574, 2014.

[16] Tianyi Lin, Nhat Ho, and Michael I. Jordan. On the Efficiency of Entropic Regularized Algorithms for Optimal Transport. *Journal of Machine Learning Research*, 23(137):1–42, 2022.

[17] Arthur Mensch and Gabriel Peyré. Online Sinkhorn: Optimal Transport distances from sample streams. In *Advances in Neural Information Processing Systems*, volume 33, pages 1657–1667, 2020.

[18] Konstantin Mishchenko. Sinkhorn algorithm as a special case of stochastic mirror descent. *arXiv preprint arXiv:1909.06918*, 2019.

[19] Youssef Mroueh, Tom Sercu, and Anant Raj. Sobolev Descent. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 2976–2985, 2019.

[20] Yurii Nesterov. *Lectures on Convex Optimization*, volume 137 of *Springer Optimization and Its Applications*. Springer, Cham, second edition, 2018.

[21] Marcel Nutz. Introduction to Entropic Optimal Transport. *Lecture notes, Columbia University*, 2021.

[22] Marcel Nutz and Johannes Wiesel. Entropic optimal transport: convergence of potentials. *Probability Theory and Related Fields*, 184(1-2):401–424, 2022.

[23] Gabriel Peyré and Marco Cuturi. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11:355–607, 2019.

[24] Mohammad Reza Karimi, Ya-Ping Hsieh, and Andreas Krause. Sinkhorn Flow as Mirror Flow: A Continuous-Time Framework for Generalizing the Sinkhorn Algorithm. In *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pages 4186–4194. PMLR, 2024.

[25] Walter Rudin. *Real and complex analysis, 3rd ed.* McGraw-Hill, Inc., USA, 1987. ISBN 0070542341.

[26] Filippo Santambrogio. *Optimal transport for applied mathematicians*, volume 87 of *Progress in Nonlinear Differential Equations and their Applications*. Birkhäuser/Springer, Cham, 2015. Calculus of variations, PDEs, and modeling.

[27] Richard Sinkhorn and Paul Knopp. Concerning nonnegative matrices and doubly stochastic matrices. *Pacific Journal of Mathematics*, 21(2):343–348, 1967.

[28] Ingo Steinwart and Andreas Christmann. *Support vector machines*. Information Science and Statistics. Springer, New York, 2008.

[29] Cédric Villani. *Topics in optimal transportation*, volume 58 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2003.

NOTATION

We present some notation for the convenience of the reader. For a set $\mathcal{Z}$, the set of probability measures over $\mathcal{Z}$ is denoted by $\mathcal{P}(\mathcal{Z})$. Given a distribution $\pi \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$, $\pi_{\mathcal{X}}$ and $\pi_{\mathcal{Y}}$ are the $\mathcal{X}$-marginal and $\mathcal{Y}$-marginal respectively i.e., for $A \subseteq \mathcal{X}, B \subseteq \mathcal{Y}$,

$$\pi_{\mathcal{X}}(A) = \int_{A \times \mathcal{Y}} \mathrm{d}\pi(x, y) \; ; \quad \pi_{\mathcal{Y}}(B) = \int_{\mathcal{X} \times B} \mathrm{d}\pi(x, y) \; .$$

Given two distributions $\mu \in \mathcal{P}(\mathcal{X}), \nu \in \mathcal{P}(\mathcal{Y})$, we say that $\pi \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ is a coupling of μ and ν if $\pi_{\mathcal{X}} = \mu$ and $\pi_{\mathcal{Y}} = \nu$. For $\rho \in \mathcal{P}(\mathcal{Z})$, we use $\mathrm{d}\rho$ to represent its density. For convenience, if ρ has a density w.r.t. the Lebesgue measure, we use ρ to also denote its density depending on the setting.

Let $\rho, \rho' \in \mathcal{P}(\mathcal{Z})$ be probability measures such that ρ is absolutely continuous w.r.t. ρ'. The KL divergence between ρ, ρ' is denoted by $\mathrm{d}_{\mathrm{KL}}(\rho \| \rho') := \int_{\mathcal{Z}} \mathrm{d}\rho \log \frac{\mathrm{d}\rho}{\mathrm{d}\rho'}$. For a functional $\mathcal{F} : \mathcal{P}(\mathcal{Z}) \to \mathbb{R}$, we call $\delta\mathcal{F}(\rho)$ its first variation at $\rho \in \mathcal{P}(\mathcal{Z})$, and this is the function (up to additive constants) that satisfies [26, Def. 7.12]

$$\langle \delta\mathcal{F}(\rho), \chi \rangle = \int_{\mathcal{Z}} \delta\mathcal{F}(\rho)(z) \mathrm{d}\chi(z) = \lim_{h \to 0} \frac{\mathcal{F}(\rho + h \cdot \chi) - \mathcal{F}(\rho)}{h} \qquad \forall\, \chi \text{ such that} \int \mathrm{d}\chi(z) = 0 \; .$$

For a measurable function $f : \mathcal{Z} \to \mathbb{R}$, its L^p-norm w.r.t. ρ is denoted by $\|f\|_{L^p(\rho)}$, which is defined as $\left(\int_{\mathcal{Z}} |f(z)|^p \mathrm{d}\rho(z)\right)^{1/p}$. When ρ is replaced with $\mathcal{Z}$ as $\|f\|_{L^p(\mathcal{Z})}$, then this is understood to be the L^p-norm of f w.r.t. the Lebesgue measure of $\mathcal{Z}$. For another measurable function $g : \mathcal{Z} \to \mathbb{R}$, the $L^2(\rho)$ inner product is defined as $\langle f, g \rangle_{L^2(\rho)} = \int_{\mathcal{Z}} f(z) g(z) \mathrm{d}\rho(z)$. As a special case, when the subscript in the norm and inner product are omitted as in $\|\cdot\|$ and $\langle \cdot, \cdot \rangle$, then these correspond to the $L^2(\mathcal{Z})$-norm and inner product respectively. Following the notation in Rudin [25, Chap. 3], we use L^p to also denote the space of measurable functions whose L^p norm is finite.

Appendix A. More details about the eOT problem

The eOT problem is also directly related to the *static Schrödinger bridge problem* [15, Def. 2.2] denoted by $\mathrm{SB}(\mu, \nu; \pi^{\mathsf{ref}})$ for the reference measure π^{ref} whose density is $\mathrm{d}\pi^{\mathsf{ref}} \propto \exp\left(-{}^{c}\!/_{\varepsilon}\right) \mathrm{d}(\mu \otimes \nu)$ and marginals μ and ν. Then, direct calculation gives

$$\mathrm{OT}_{\varepsilon}(\mu, \nu; c) = \varepsilon \cdot \Big\{ \underbrace{\inf_{\pi \in \Pi(\mu, \nu)} \mathrm{d}_{\mathrm{KL}}(\pi \| \pi^{\mathsf{ref}})}_{\mathrm{SB}(\mu, \nu; \pi^{\mathsf{ref}})} - \log Z_{\mathsf{ref}} \Big\} \; ; \; Z_{\mathsf{ref}} = \iint \exp\left(-\frac{c(x, y)}{\varepsilon} \right) \mathrm{d}\mu(x) \mathrm{d}\nu(y) \; . \tag{9}$$

The above problem is a (strictly) convex minimisation problem since $\Pi(\mu, \nu)$ is convex, and $\pi \mapsto \mathrm{d}_{\mathrm{KL}}(\pi \| \pi^{\mathsf{ref}})$ is strictly convex, and the optimal value of the problem is

$$\mathrm{OT}_{\varepsilon}(\mu, \nu; c) = \varepsilon \cdot \left(\int \phi^{\star}(y) \mathrm{d}\nu(y) - \int \psi^{\star}(x) \mathrm{d}\mu(x) \right) \; .$$

While the solution $\pi^{\star}$ is unique, the Schrödinger potentials are unique only up to constants, that is, if $\psi^{\star}$ and $\phi^{\star}$ are Schrödinger potentials, then $\psi^{\star} + \beta$ and $\phi^{\star} + \beta$ are also Schrödinger potentials for any constant $\beta \in \mathbb{R}$.

A.1. Related work

Traditional analyses view the Sinkhorn algorithm as either alternating projection on the two marginals μ, ν or block maximization on the two dual potentials ψ, ϕ. These render a linear convergence with a contraction rate of the form $1 - e^{-\|c\|_\infty/\varepsilon}$. An important limitation of this analysis is that the rate becomes *exponentially slower* with growing $\|c\|_\infty$ or decreasing ε. Recently, several analyses have focused on the Sinkhorn algorithm in the setting where $\mathcal{X}$ and $\mathcal{Y}$ are discrete spaces [1, 11, 16]. More relevant to us is where $\mathcal{X}$ and $\mathcal{Y}$ are continuous spaces where many probabilistic approaches have been taken for analyzing the Sinkhorn algorithm. Most recently, Chiarini et al. [4] leverage the stability of optimal plans with respect to the marginals to obtain exponential convergence with unbounded cost for all $\varepsilon > 0$, albeit under various sets of conditions on the marginals. This relaxes the assumptions made in Chizat et al. [5] for semi-concave bounded costs while still maintaining a contraction rate that only deteriorates *polynomially* in ε. We refer the reader to Chiarini et al. [4, Sec. 1.5] for a more comprehensive literature review for analyses of the Sinkhorn algorithm. In summary, the most recent analyses place assumptions on the growth of the cost, decay of the tails of μ, ν and/or log-concavity, to obtain exponential convergence guarantees.

In contrast, the advantage of taking the optimisation route i.e., viewing the Sinkhorn algorithm as performing infinite-dimensional mirror descent [2, 14, 24] is that it provides a guarantee under *minimal* assumptions. From non-asymptotic guarantee standpoint, these aforementioned works furnish a discrete-time iteration complexity that scales as $1/N\varepsilon$. If the costs are additionally assumed to be bounded, then Aubin-Frankowski et al. [2] recover a contractive rate reminiscent of the classical Hilbert analysis. In this work, we achieve the similar rates while significantly expanding the scope of algorithm design and shed more light on the eOT problem, unifying both the primal and dual perspectives. [17] gives another mirror descent interpretation of Sinkhorn but the change of variable results in a non-convex objective which is hard to prove convergence for. There has also been interest in designing alternative algorithms for the eOT problem, among them [6] that designs Wasserstein gradient flow dynamics over the submanifold of $\Pi(\mu, \nu)$ which borrow tools from SDEs and PDEs in its analysis.

Appendix B. The origins of Φ-match and its interpretations

Related to Eq. (6), the Sinkhorn algorithm corresponds to an update map $\mathsf{M}^{\mathsf{Sinkhorn}}$ [24, Lem. 2] that is defined as, which corresponds to Φ-match with $\Phi : f \to f$.

$$\mathsf{M}^{\mathsf{Sinkhorn}}(\phi) := \phi - \left(\log \frac{\pi(\phi, \phi^+)_{\mathcal{Y}}}{\nu}\right) . \tag{Sinkhorn}$$

Note that Φ-match takes a density function ($L^1(\mathcal{Y})$) and returns a function that is not necessarily a density ($L^\infty(\mathcal{Y})$).

B.1. Interpretations of Φ-match

The interpretations of Φ-match that we discuss here are motivated by recent work in understanding the Sinkhorn algorithm (Sinkhorn) [2, 24]. In essence, these prior works view the Sinkhorn algorithm as minimising $\mathrm{d}_{\mathrm{KL}}(\pi_{\mathcal{Y}}, \nu)$ over $\mathcal{Q}$ that is formally defined as

$$\mathcal{Q} := \left\{ \pi : \exists\, \phi \in L^1(\nu) \text{ such that } \pi(x, y) = \exp\left(\phi(y) - \phi^+(x) - \frac{c(x, y)}{\varepsilon}\right) \mu(x)\nu(y) \right\} \,. \quad (10)$$

While $\pi^\star$ belongs to $\mathcal{Q}$, this constrained set is *not* a convex subset of $\mathcal{P}(\mathcal{X} \times \mathcal{Y})$ owing to the factorisation structure. We specifically show that Φ-match can be interpreted in the following two ways: (1) as an alternating projection scheme and (2) as a local greedy method analogous to gradient descent / mirror descent (as discussed in Section 3). While Φ-match is derived from the semi-dual, these interpretations do not involve the semi-dual and solely operate in $\mathcal{P}(\mathcal{X} \times \mathcal{Y})$.

Φ-match as iterative projections on $\mathcal{P}(\mathcal{X} \times \mathcal{Y})$ Consider the following projection operations.

$$\mathsf{project}_{\mathcal{Y},\nu}(\pi; \Phi) := \operatorname*{argmin}_{\bar{\pi}} \left\{ \mathrm{d}_{\mathrm{KL}}(\bar{\pi} \| \pi) : \bar{\pi}_{\mathcal{Y}} \propto \pi_{\mathcal{Y}} \cdot \frac{\Phi(\nu)}{\Phi(\pi_{\mathcal{Y}})} \right\}, \quad (11a)$$

$$\mathsf{project}_{\mathcal{X},\mu}(\pi', \pi; \eta) := \operatorname*{argmin}_{\bar{\pi}} \left\{ \eta \cdot \mathrm{d}_{\mathrm{KL}}(\bar{\pi} \| \pi') + (1 - \eta) \cdot \mathrm{d}_{\mathrm{KL}}(\bar{\pi} \| \pi) : \bar{\pi}_{\mathcal{X}} = \mu \right\} \,. \quad (11b)$$

For a given $\pi \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$, $\mathsf{project}_{\mathcal{Y},\nu}$ can be seen as correcting the $\mathcal{Y}$-marginal of π towards ν, and the nature of this correction depends on Φ. On the other hand, $\mathsf{project}_{\mathcal{X},\mu}$ finds a "midpoint" (which depends on the stepsize $\eta \in [0, 1]$) while ensuring that the $\mathcal{X}$-marginal is μ. In the following lemma, we show that if $\pi \in \mathcal{Q}$ (corresponding to some $\phi \in L^1(\nu)$), then applying the projections Eqs. (11a) and (11b) successively is equivalent to updating ϕ using Φ-match.

Lemma 3 *Let $\phi^0 \in L^1(\nu)$ and let $\{\phi^n\}_{n \geq 1}$ be the sequence of potentials obtained as $\phi^{n+1} = \mathsf{M}^{\Phi\text{-match}}(\phi^n; \eta)$ for $\eta \in [0, 1]$. Then, the sequence of distributions $\{\pi^n\}_{n \geq 0}$ where $\pi^n = \pi(\phi^n, (\phi^n)^+)$ satisfy for every $n \geq 0$*

$$\pi^{n+1} = \mathsf{project}_{\mathcal{X},\mu}(\pi^{n+1/2}, \pi^n; \eta) \quad \textit{where } \pi^{n+1/2} = \mathsf{project}_{\mathcal{Y},\nu}(\pi^n; \Phi) \,.$$

This is related to Eq. (7) by the following lemma.

Lemma 4 *Let $\pi \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ be such that $\pi_{\mathcal{X}} = \mu$. Then for $\eta \in [0, 1]$,*

$$\mathsf{project}_{\mathcal{X},\mu}(\mathsf{project}_{\mathcal{Y},\nu}(\pi; \Phi), \pi; \eta) = \mathsf{root}_{\mathcal{X},\mu}(\pi; \mathcal{V}_{\Phi}, \eta) \,.$$

Theorem 1 is a direct corollary of Lemma 3 and Lemma 4. The proofs of Lemmas 3 and 4 is given in Appendix B.3 respectively.

Recall from earlier that Φ-match with $\Phi : f \mapsto f$ and $\eta = 1$ coincides with Sinkhorn. In this setting, Lemma 3 recovers the classical iterative Bregman projection interpretation of the Sinkhorn method [23, Remark 4.8].

B.2. More details about MMD and k-SGA

Let $k : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$ be a positive-definite kernel, and let $\mathcal{H}_k$ be its RKHS. We refer the reader to Steinwart and Christmann [28, Chap. 4] for a more detailed exposition about kernels and their RKHS. For $\xi \in \mathcal{P}(\mathcal{Z})$, the mean function w.r.t. kernel k is defined as

$$\mathfrak{m}_k(\xi)(y) := \int_{\mathcal{Y}} k(y, y') \cdot \mathrm{d}\xi(y') \,.$$

The MMD (for a kernel k) [13] is defined as

$$\mathrm{MMD}_k(\xi, \rho) = \sup_{\substack{f \in \mathcal{H}_k \\ \|f\|_{\mathcal{H}_k} \leq 1}} |\mathbb{E}_\xi[f] - \mathbb{E}_\rho[f]| = \|\mathfrak{m}_k(\xi) - \mathfrak{m}_k(\rho)\|_{\mathcal{H}_k} .$$

k-SGA as kernelised SGA Since the map $\mathfrak{m}_k$ is additive, i.e., $\mathfrak{m}_k(f_1 + f_2) = \mathfrak{m}_k(f_1) + \mathfrak{m}_k(f_2)$ for suitable functions f_1 and f_2, we have

$$\begin{aligned}\mathfrak{m}_k(\delta J(\phi))(y) &= \int_{\mathcal{Y}} k(y, y') \cdot \delta J(\phi)(y')\mathrm{d}y' \\ &= \int_{\mathcal{Y}} k(y, y') \cdot (\nu(y') - \pi(\phi, \phi^+)_{\mathcal{Y}}(y'))\mathrm{d}y' \\ &= \mathfrak{m}_k(\nu)(y) - \mathfrak{m}_k(\pi(\phi, \phi^+)_{\mathcal{Y}})(y) .\end{aligned}$$

Then, k-SGA can be equivalently written as

$$\mathsf{M}^{k\text{-SGA}}(\phi; \eta) = \phi + \eta \cdot \mathfrak{m}_k(\delta J(\phi))$$

and can be viewed as performing kernel smoothing δJ before using it for the update.

B.3. Proofs of Lemma 3 and Lemma 4

Proof [Proof of Lemma 3] Consider some $n \geq 0$. By the decomposition of KL divergence,

$$\mathrm{d}_{\mathrm{KL}}(\pi\|\pi^n) = \mathbb{E}_{y \sim \pi_{\mathcal{Y}}}[\mathrm{d}_{\mathrm{KL}}(\pi_{\mathcal{X}|\mathcal{Y}}(.|y)\|\pi^n_{\mathcal{X}|\mathcal{Y}}(.|y))] + \mathrm{d}_{\mathrm{KL}}(\pi_{\mathcal{Y}}\|\pi^n_{\mathcal{Y}}) .$$

For convenience, we denote $\mathsf{project}_{\mathcal{Y},\nu}(\pi^n; \Phi)$ as $\pi^{n+1/2}$. By definition of $\mathsf{project}_{\mathcal{Y},\nu}(\pi^n; \Phi)$, we have

$$\pi^{n+1/2}_{\mathcal{Y}}(y) = \frac{1}{Z} \cdot \pi^n_{\mathcal{Y}}(y) \cdot \frac{\Phi(\nu)(y)}{\Phi(\pi^n_{\mathcal{Y}})(y)}; \quad \pi^{n+1/2}_{\mathcal{X}|\mathcal{Y}}(x|y) = \pi^n_{\mathcal{X}|\mathcal{Y}}(x|y) .$$

Above, $Z = \mathbb{E}_{\mathsf{y} \sim \pi^n_{\mathcal{Y}}}\left[\frac{\Phi(\nu)(\mathsf{y})}{\Phi(\pi^n_{\mathcal{Y}})(\mathsf{y})}\right]$. Therefore,

$$\pi^{n+1/2}(x, y) = \frac{1}{Z} \cdot \pi^n(x, y) \cdot \frac{\Phi(\nu)(y)}{\Phi(\pi^n_{\mathcal{Y}})(y)} .$$

Since $\pi^n = \pi(\phi^n, (\phi^n)^+)$, this shows that $\pi^{n+1/2}$ factorises as

$$\pi^{n+1/2}(x, y) = \exp\left(-\psi^{n+1/2}(x) + \phi^{n+1/2}(y) - \frac{c(x, y)}{\varepsilon}\right)\mu(x)\nu(y)$$

where $\phi^{n+1/2}(y) = \phi^n(y) + (\log\Phi(\nu)(y) - \log\Phi(\pi^n_{\mathcal{Y}})(y))$ and $\psi^{n+1/2}(x) = \psi^n(x) + \log Z$. From [24, Corr. B.1], we have that $\mathsf{project}_{\mathcal{X},\mu}(\pi^{n+1/2}, \pi^n; \eta)$ satisfies

$$\begin{aligned}\mathsf{project}_{\mathcal{X},\mu}(\pi^{n+1/2}, \pi^n; \eta)_{\mathcal{Y}|\mathcal{X}}(y|x) &= \frac{\pi^{n+1/2}_{\mathcal{Y}|\mathcal{X}}(y|x)^\eta \cdot \pi^n_{\mathcal{Y}|\mathcal{X}}(y|x)^{1-\eta}}{C(x)} \\ C(x) &= \int_{\mathcal{Y}} \pi^{n+1/2}_{\mathcal{Y}|\mathcal{X}}(y|x)^\eta \cdot \pi^n_{\mathcal{Y}|\mathcal{X}}(y|x)^{1-\eta}\mathrm{d}y .\end{aligned}$$

The factorisations of π^n and $\pi^{n+1/2}$ results in $\mathsf{project}_{\mathcal{X},\mu}(\pi^{n+1/2}, \pi^n; \eta)$ factorising as

$$\mathsf{project}_{\mathcal{X},\mu}(\pi^{n+1/2}, \pi^n; \eta)(x, y) = \exp\left(\bar{\phi}(y) - \bar{\psi}(x) - \frac{c(x, y)}{\varepsilon}\right) \mu(x)\nu(y)$$

and specifically,

$$\begin{aligned} \bar{\phi}(y) &= \eta \cdot \phi^{n+1/2}(y) + (1 - \eta) \cdot \phi^n(y) \\ &= \phi^n(y) + \eta \cdot (\log \Phi(\nu)(y) - \log \Phi(\pi^n_{\mathcal{Y}})(y)) \; . \end{aligned} \tag{12}$$

Since $\mathsf{project}_{\mathcal{X},\mu}(\pi^{n+1/2}, \pi^n; \eta)_{\mathcal{X}} = \mu$, this implies

$$\bar{\psi}(x) = \log \int_{\mathcal{Y}} \exp\left(\bar{\phi}(y) - \frac{c(x, y)}{\varepsilon}\right) \nu(y)\mathrm{d}y = \bar{\phi}^{+}(x) \; .$$

Hence, comparing Eq. (12) with Φ-match we have $\mathsf{project}_{\mathcal{X},\mu}(\pi^{n+1/2}, \pi^n; \eta) = \pi(\phi^{n+1}; (\phi^{n+1})^{+})$ which completes the proof. ■

Proof [Proof of Lemma 4] For convenience, we use the shorthand notation $\tilde{\pi} = \mathsf{project}_{\mathcal{Y},\nu}(\pi; \Phi)$. As in the proof of Lemma 3, Eq. (11a) ensures that

$$\tilde{\pi}(x, y) = \frac{1}{Z} \cdot \pi(x, y) \cdot \frac{\Phi(\nu)(y)}{\Phi(\pi_{\mathcal{Y}})(y)} \; ; \qquad Z = \mathbb{E}_{\mathsf{y}\sim\pi_{\mathcal{Y}}}\left[\frac{\Phi(\nu)(\mathsf{y})}{\Phi(\pi_{\mathcal{Y}})(\mathsf{y})}\right]$$

$$\Rightarrow \log \Phi(\pi_{\mathcal{Y}})(y) - \log \Phi(\nu)(y) = \log \frac{\pi(x, y)}{\tilde{\pi}(x, y)} - \log Z \; .$$

The objective in Eq. (7) with $\mathcal{F} \leftarrow \mathcal{V}_\Phi$ can be simplified as

$$\begin{aligned} &\langle \mathcal{V}_\Phi(\pi), \bar{\pi} - \pi \rangle + \frac{1}{\eta} \cdot \mathrm{d}_{\mathrm{KL}}(\bar{\pi} \| \pi) \\ &\qquad \iint \mathcal{V}_\Phi(\pi)(x, y)(\bar{\pi}(x, y) - \pi(x, y))\mathrm{d}x\mathrm{d}y \\ &\qquad\quad + \frac{1}{\eta} \cdot \iint \bar{\pi}(x, y) \log\left(\frac{\bar{\pi}(x, y)}{\pi(x, y)}\right) \mathrm{d}x\mathrm{d}y \\ &\qquad = \iint \left(\log \Phi(\pi^n_{\mathcal{Y}})(y) - \log \Phi(\nu)(y)\right) \cdot \bar{\pi}(x, y)\mathrm{d}x\mathrm{d}y + \log Z \\ &\qquad\quad + \frac{1}{\eta} \cdot \iint \bar{\pi}(x, y) \log\left(\frac{\bar{\pi}(x, y)}{\pi(x, y)}\right) \mathrm{d}x\mathrm{d}y \\ &\qquad\quad - \underbrace{\left(\log Z + \iint (\log \Phi(\pi_{\mathcal{Y}})(y) - \log \Phi(\nu)(y)) \cdot \pi(x, y)\mathrm{d}x\mathrm{d}y\right)}_{c(\pi)} \\ &\qquad = \iint \bar{\pi}(x, y) \cdot \log\left(\frac{\pi(x, y)}{\tilde{\pi}(x, y)}\right) \mathrm{d}x\mathrm{d}y \\ &\qquad\quad + \frac{1}{\eta} \cdot \iint \bar{\pi}(x, y) \log\left(\frac{\bar{\pi}(x, y)}{\pi(x, y)}\right) \mathrm{d}x\mathrm{d}y + c(\pi) \\ &\qquad = \frac{1}{\eta}\left\{\iint \bar{\pi}(x, y) \cdot \log\left[\left(\frac{\bar{\pi}(x, y)}{\tilde{\pi}(x, y)}\right)^{\eta} \left(\frac{\bar{\pi}(x, y)}{\pi(x, y)}\right)^{1-\eta}\right] \mathrm{d}x\mathrm{d}y\right\} \\ &\qquad\quad + c(\pi) \; . \end{aligned}$$

The objective in $\mathsf{project}_{\mathcal{X},\mu}(\tilde{\pi}, \pi; \eta)$ can be expanded as

$$\eta \mathrm{d}_{\mathrm{KL}}(\overline{\pi}\|\tilde{\pi}) + (1-\eta)\mathrm{d}_{\mathrm{KL}}(\overline{\pi}\|\pi) = \iint \overline{\pi}(x,y)\cdot\log\left[\left(\frac{\overline{\pi}(x,y)}{\tilde{\pi}(x,y)}\right)^{\eta}\left(\frac{\overline{\pi}(x,y)}{\pi(x,y)}\right)^{1-\eta}\right]\mathrm{d}x\mathrm{d}y$$

thus establishing the equivalence in the statement as $\mathsf{project}_{\mathcal{X},\mu}$ also minimises over the set $\{\overline{\pi} : \overline{\pi}_{\mathcal{X}} = \mu\}$ and $c(\pi)$ is a constant. ■

B.4. Deriving non-asymptotic rates for k-SGA

To prove Theorem 2, we use the following key lemma from [2] which characterises the growth of $\mathcal{L}_k(\cdot;\nu)$ relative to the entropy functional $H : \xi \mapsto \int_{\mathcal{Y}} \mathrm{d}\xi \log \mathrm{d}\xi$. This, along with the convexity of $\mathcal{L}_k(\cdot;\nu)$, is instrumental in establishing a rate of convergence for k-SGA.

Proposition 1 ([2, Prop. 14]) *Let $k : \mathcal{Y}\times\mathcal{Y}\to\mathbb{R}$ be a bounded, positive definite kernel where $c_k := \sup_{y\in\mathcal{Y}} k(y,y) < \infty$. Then for any $\xi, \overline{\xi} \in \mathcal{P}(\mathcal{Y})$,*

$$0 \le \left\langle \delta\mathcal{L}_k(\overline{\xi};\nu) - \delta\mathcal{L}_k(\xi;\nu), \mathrm{d}\overline{\xi} - \mathrm{d}\xi\right\rangle \le 2c_k \cdot \left\langle \delta H(\overline{\xi}) - \delta H(\xi), \mathrm{d}\overline{\xi} - \mathrm{d}\xi\right\rangle \ .$$

Consequently,

$$0 \le \mathcal{L}_k(\overline{\xi};\nu) - \mathcal{L}_k(\xi;\nu) - \langle \delta\mathcal{L}_k(\xi;\nu), \mathrm{d}\overline{\xi} - \mathrm{d}\xi\rangle \le 2c_k \cdot \mathrm{d}_{\mathrm{KL}}(\overline{\xi}\|\xi) \ .$$

Proof [Proof of Theorem 2] The proof is be obtained in the manner of the proof of Aubin-Frankowski et al. [2, Thm. 4] while catering to the squared MMD $\mathcal{L}_k$. We give the details here for completeness. For an arbitrary $n \ge 0$, we have the following identity for any $\overline{\pi}$ such that $\overline{\pi}_{\mathcal{X}} = \mu$ that

$$\begin{aligned}\eta\cdot\langle \mathcal{V}_{\Phi_k}(\pi^n), \overline{\pi} - \pi^n\rangle &+ \mathrm{d}_{\mathrm{KL}}(\overline{\pi}\|\pi^n) \\ &\ge \eta\cdot\langle \mathcal{V}_{\Phi_k}(\pi^n), \pi^{n+1} - \pi^n\rangle + \mathrm{d}_{\mathrm{KL}}(\pi^{n+1}\|\pi^n) + \mathrm{d}_{\mathrm{KL}}(\overline{\pi}\|\pi^{n+1}) \ . \end{aligned} \tag{13}$$

This is obtained by the three-point identity [2, Lem. 3] with

$$C \leftarrow \{\pi \in \mathcal{P}(\mathcal{X}\times\mathcal{Y}) : \pi_{\mathcal{X}} = \mu\} \ , \quad \mathcal{G} \leftarrow \eta\cdot\langle \mathcal{V}_{\Phi_k}(\pi^n), \cdot - \pi^n\rangle \ , \quad D_\phi(\cdot|\cdot) \leftarrow \mathrm{d}_{\mathrm{KL}}(\cdot\|\cdot) \ .$$

By the definition of $\mathcal{V}_{\Phi_k}(\pi^n) = \mathfrak{m}_k(\pi^n_{\mathcal{Y}}) - \mathfrak{m}_k(\nu) = \delta\mathcal{L}_k(\pi^n_{\mathcal{Y}};\nu)$, we have

$$\begin{aligned}\langle \mathcal{V}_{\Phi_k}(\pi^n), \pi^{n+1} - \pi^n\rangle &= \int_{\mathcal{Y}}\int_{\mathcal{X}} \delta\mathcal{L}_k(\pi^n_{\mathcal{Y}};\nu)(y)\cdot(\pi^{n+1}(x,y) - \pi^n(x,y))\,\mathrm{d}x\mathrm{d}y \\ &= \langle \delta\mathcal{L}_k(\pi^n_{\mathcal{Y}};\nu), \pi^{n+1}_{\mathcal{Y}} - \pi^n_{\mathcal{Y}}\rangle \ .\end{aligned}$$

From Proposition 1, we know that that in this case

$$\begin{aligned}\mathcal{L}_k(\pi^{n+1}_{\mathcal{Y}};\nu) &\le \mathcal{L}_k(\pi^n_{\mathcal{Y}};\nu) + \langle \delta\mathcal{L}_k(\pi^n_{\mathcal{Y}};\nu), \pi^{n+1}_{\mathcal{Y}} - \pi^n_{\mathcal{Y}}\rangle + 2c_k\cdot\mathrm{d}_{\mathrm{KL}}(\pi^{n+1}_{\mathcal{Y}}\|\pi^n_{\mathcal{Y}}) \\ &= \mathcal{L}_k(\pi^n_{\mathcal{Y}};\nu) + \langle \mathcal{V}_{\Phi}(\pi^n), \pi^{n+1} - \pi^n\rangle + 2c_k\cdot\mathrm{d}_{\mathrm{KL}}(\pi^{n+1}_{\mathcal{Y}}\|\pi^n_{\mathcal{Y}}) \\ &\overset{(a)}{\le} \mathcal{L}_k(\pi^n_{\mathcal{Y}};\nu) + \langle \mathcal{V}_{\Phi}(\pi^n), \pi^{n+1} - \pi^n\rangle + 2c_k\cdot\mathrm{d}_{\mathrm{KL}}(\pi^{n+1}\|\pi^n) \qquad (14) \\ &\overset{(b)}{\le} \mathcal{L}_k(\pi^n_{\mathcal{Y}};\nu) + \left(2c_k - \frac{1}{\eta}\right)\cdot\mathrm{d}_{\mathrm{KL}}(\pi^{n+1}\|\pi^n) - \frac{1}{\eta}\cdot\mathrm{d}_{\mathrm{KL}}(\pi^n\|\pi^{n+1}) \\ &\overset{(c)}{\le} \mathcal{L}_k(\pi^n_{\mathcal{Y}};\nu) \ . \qquad (15)\end{aligned}$$

Step (a) is a consequence of the KL decomposition, step (b) applies Eq. (13) for $\overline{\pi} \leftarrow \pi^n$, and step (c) uses the fact that $\eta = \min\left\{\frac{1}{2c_k}, 1\right\} \leq \frac{1}{2c_k}$ and the non-negativity of the KL divergence. We also have by Proposition 1 that

$$\begin{aligned}\mathcal{L}_k(\overline{\pi}_{\mathcal{Y}}; \nu) - \mathcal{L}_k(\pi^n_{\mathcal{Y}}; \nu) &\geq \langle \delta\mathcal{L}_k(\pi^n_{\mathcal{Y}}; \nu), \overline{\pi}_{\mathcal{Y}} - \pi^n_{\mathcal{Y}}\rangle \\ &= \langle \mathcal{V}_\Phi(\pi^n), \overline{\pi} - \pi^n\rangle \ .\end{aligned}$$

Substituting the above and Eq. (14) in Eq. (13) with $\eta = \min\left\{\frac{1}{2c_k}, 1\right\}$, we obtain

$$\frac{1}{\max\{2c_k, 1\}}\mathcal{L}_k(\pi^{n+1}_{\mathcal{Y}}; \nu) - \frac{1}{\max\{2c_k, 1\}}\mathcal{L}_k(\overline{\pi}_{\mathcal{Y}}; \nu) \leq \mathrm{d}_{\mathrm{KL}}(\overline{\pi}\|\pi^n) - \mathrm{d}_{\mathrm{KL}}(\overline{\pi}\|\pi^{n+1}) \ .$$

Summing both sides from $n = 0$ to $n = N - 1$ yields

$$\frac{1}{\max\{2c_k, 1\}}\sum_{n=1}^{N}\{\mathcal{L}_k(\pi^n_{\mathcal{Y}}; \nu) - \mathcal{L}_k(\overline{\pi}_{\mathcal{Y}}; \nu)\} \leq \mathrm{d}_{\mathrm{KL}}(\overline{\pi}\|\pi^0) - \mathrm{d}_{\mathrm{KL}}(\overline{\pi}\|\pi^N) \ .$$

We know that $\mathcal{L}_k(\pi^{n+1}_{\mathcal{Y}}; \nu) \leq \mathcal{L}_k(\pi^n_{\mathcal{Y}}; \nu)$ from Eq. (15). Noting that $\pi^\star$ satisfies $\pi^\star_{\mathcal{X}} = \mu$ and $\pi^\star_{\mathcal{Y}} = \nu$, we substitute $\overline{\pi} \leftarrow \pi^\star$ above and this leads to

$$\mathcal{L}_k(\pi^N_{\mathcal{Y}}; \nu) \leq \frac{\max\{2c_k, 1\}}{N} \cdot \mathrm{d}_{\mathrm{KL}}(\pi^\star\|\pi^0) \ .$$

■

Appendix C. More details about the steepest ascent methods in Section 4

C.1. Signed semi-dual gradient ascent

Here we consider the update in sign-SGA. Note that for any $\phi \in L^1(\nu) \cap L^\infty(\mathcal{Y})$, $\mathsf{M}^{\textsf{sign-SGA}}(\phi; \eta) \in L^1(\nu) \cap L^\infty(\mathcal{Y})$ due to the form of $\delta J(\phi)$. This is because for any $\phi \in L^1(\nu)$, $\|\delta J(\phi)\|_{L^1(\mathcal{Y})} = \|\pi(\phi, \phi^+)_{\mathcal{Y}} - \nu\|_{L^1(\mathcal{Y})} \leq 2$, and the sign function being bounded pointwise. Also, for a sufficiently small $\eta > 0$, the growth property implied by Lemma 1 asserts that $J(\mathsf{M}^{\textsf{sign-SGA}}(\phi; \eta)) \geq J(\phi)$ since one can show

$$J(\mathsf{M}^{\textsf{sign-SGA}}(\phi; \eta)) \geq J(\phi) + \eta \cdot \|\delta J(\phi)\|^2_{L^1(\mathcal{Y})} - \frac{\eta^2}{2} \cdot \|\delta J(\phi)\|^2_{L^1(\mathcal{Y})} \ .$$

This ascent property in conjunction with the concavity of J enables us to give a non-asymptotic convergence rate for sign-SGA with an "anchoring" step, which does not change the function value due to the following fact.

Fact 1 (Shift-invariance) *The semi-dual is invariant to additive perturbations of its argument. Formally, for any $C \in \mathbb{R}$ and $\phi \in L^1(\nu)$, $J(\phi + C \cdot \mathbf{1}) = J(\phi)$ where $\mathbf{1} : x \mapsto 1$. This is due to the fact that $(\phi + C \cdot \mathbf{1})^+ = \phi^+ + C \cdot \mathbf{1}$.*

Let $y_{\texttt{anc}} \in \mathcal{Y}$ be an anchor point, and $\phi^0 \in L^1(\nu) \cap L^\infty(\mathcal{Y})$ be such that $\phi^0(y_{\texttt{anc}}) = 0$. Define the set

$$\mathcal{T}_{\phi^0, y_{\texttt{anc}}} := \{\phi \in L^1(\nu) \cap L^\infty(\mathcal{Y}) : \phi(y_{\texttt{anc}}) = 0, J(\phi) \geq J(\phi^0)\} \ .$$

Theorem 6 (Formal version of Theorem 3) *Let $\{\phi^n\}_{n\geq 1}$ be the sequence of potentials generated according to the following recursion for $n \geq 0$:*

$$\begin{aligned}\phi^{n+1/2} &= \mathsf{M}^{\textsf{sign-SGA}}(\phi^n; \eta)\\ \phi^{n+1} &= \phi^{n+1/2} - (\phi^{n+1/2}(y_{\texttt{anc}}) - \phi^n(y_{\texttt{anc}})) \cdot \mathbf{1}\end{aligned}$$

with $\phi^0, y_{\texttt{anc}}$ as defined above. Then, for all $N \geq 1$, $\phi^N \in \mathcal{T}_{\phi^0, y_{\texttt{anc}}}$ and for $\widetilde{\phi}^\star = \operatorname{argmax}\{J(\phi) : \phi \in \mathcal{T}_{\phi^0, y_{\texttt{anc}}}\}$ we have

$$J(\phi^N) - J(\widetilde{\phi}^\star) \geq -\frac{2 \cdot \operatorname{diam}(\mathcal{T}_{\phi^0, y_{anc}}; L^\infty(\mathcal{Y}))^2}{N+1} \ , \quad \textit{where} \ \ \operatorname{diam}(\mathcal{S}; L^\infty(\mathcal{Y})) := \sup_{\overline{\phi}, \phi \in \mathcal{S}} \|\overline{\phi} - \phi\|_{L^\infty(\mathcal{Y})} \ .$$

The first step applies sign-SGA, and the second step recenters the iterate to satisfy $\phi^{n+1}(y_{\texttt{anc}}) = \phi^n(y_{\texttt{anc}})$. The recentering does not affect the value of the semi-dual as noted previously in Fact 1, and if $\phi^\star \in L^\infty(\mathcal{Y})$ is a Schrödinger potential, then $\phi^\star - \phi^\star(y_{\texttt{anc}}) \cdot \mathbf{1}$ is also a maximiser of J, which lies in $\mathcal{T}_{\phi^0, y_{\texttt{anc}}}$. This shift-invariance also explains the use of the anchoring step: without anchoring, the superlevel set of J is unbounded. We can hence infer that the sequence $\{J(\phi^n)\}_{n\geq 1}$ converges to the maximum of the semi-dual J.

C.1.1. PROOF OF THEOREM 6

Proof From Lemma 1, we have for any $\phi, \overline{\phi} \in L^1(\nu) \cap L^\infty(\mathcal{Y})$ that

$$J(\overline{\phi}) - J(\phi) - \langle \delta J(\phi), \overline{\phi} - \phi \rangle \geq -\frac{\|\overline{\phi} - \phi\|_\infty^2}{2} \ .$$

For any $n \geq 0$, substituting $\phi \leftarrow \phi^n$ and $\overline{\phi} \leftarrow \phi^{n+1/2} = \mathsf{M}^{\textsf{sign-SGA}}(\phi^n; 1)$, we get

$$\begin{aligned} J(\phi^{n+1/2}) &\geq J(\phi^n) + \langle \delta J(\phi^n), \phi^{n+1/2} - \phi^n \rangle - \frac{\|\phi^{n+1/2} - \phi^n\|_\infty^2}{2}\\ &\overset{(a)}{=} J(\phi^n) + \|\delta J(\phi^n)\|_{L^1(\mathcal{Y})} \cdot \|\delta J(\phi^n)\|_{L^1(\mathcal{Y})} - \frac{1}{2} \cdot \|\delta J(\phi^n)\|_{L^1(\mathcal{Y})}^2\\ &= J(\phi^n) + \frac{\|\delta J(\phi^n)\|_{L^1(\mathcal{Y})}^2}{2} \ . \end{aligned} \tag{16}$$

Step (a) above is due to the fact that $\langle \operatorname{sign}(\delta J(\phi^n)), \delta J(\phi^n) \rangle = \|\delta J(\phi^n)\|_1$. By the shift invariance of the semi-dual, $J(\phi^{n+1}) = J(\phi^{n+1/2})$, which implies

$$J(\phi^{n+1}) \geq J(\phi^n) + \frac{\|\delta J(\phi^n)\|_{L^1(\mathcal{Y})}^2}{2} \ .$$

Hence $\phi^{n+1} \in \mathcal{T}_{\phi^0, y_{\texttt{anc}}}$ as $\phi^{n+1}(y_{\texttt{anc}}) = \phi^n(y_{\texttt{anc}})$. Next, by concavity of J that

$$J(\widetilde{\phi}^\star) \leq J(\phi^n) + \langle \delta J(\phi^n), \widetilde{\phi}^\star - \phi^n \rangle \ . \tag{17}$$

Define the Lyapunov function $E_n := \frac{n(n+1)}{2} \cdot (J(\phi^n) - J(\widetilde{\phi}^\star))$. We have

$$\begin{aligned} E_{n+1} - E_n &= \frac{(n+2)(n+1)}{2} \cdot (J(\phi^{n+1}) - J(\phi^n)) + (n+1) \cdot (J(\phi^n) - J(\widetilde{\phi}^\star)) \\ &\overset{(a)}{\geq} (n+1) \cdot \left\{ \frac{n+2}{4} \cdot \|\delta J(\phi^n)\|_1^2 + \langle \delta J(\phi^n), \phi^n - \widetilde{\phi}^\star \rangle \right\} \\ &\overset{(b)}{\geq} -\frac{n+1}{n+2} \cdot \|\phi^n - \widetilde{\phi}^\star\|_\infty^2 \\ &\overset{(c)}{\geq} -\mathrm{diam}(\mathcal{T}_{\phi^0, y_{\mathrm{anc}}}; L^\infty(\mathcal{Y}))^2 \,. \end{aligned}$$

Above, step (a) applies Eqs. (16) and (17), and step (b) applies the Hölder-Young inequality. Finally, step (c) uses the fact that $\phi^n, \widetilde{\phi}^\star \in \mathcal{T}_{\phi^0, y_{\mathrm{anc}}}$ shown previously. Summing the above inequality from $n = 0$ to $n = N-1$, we get

$$\begin{aligned} E_N - E_0 &\geq -N \cdot \mathrm{diam}(\mathcal{T}_{\phi^0, y_{\mathrm{anc}}}; L^\infty(\mathcal{Y}))^2 \\ \Rightarrow J(\phi^N) - J(\widetilde{\phi}^\star) &\geq -\frac{2 \cdot \mathrm{diam}(\mathcal{T}_{\phi^0, y_{\mathrm{anc}}}; L^\infty(\mathcal{Y}))^2}{N+1} \,. \end{aligned}$$

■

C.2. Projected semi-dual gradient ascent

Here we consider proj-SGA and its accelerated variant proj-SGA++. As mentioned briefly previously, when the cost function is bounded in a certain manner, it is possible to show that the semi-dual satisfies a different notion of smoothness, but non-uniformly depending on the "size" of the domain considered. While Lemma 1 is a general statement, regularity condition Lemma 2 is parameterised by a "size" parameter B, and hence it is useful to understand what a reasonable choice of B is for the purposes of solving the eOT problem. If B is too small, then it is likely that the Schrödinger potential $\phi^\star$ would not satisfy $\|\phi^\star\|_{L^\infty(\mathcal{Y})} \leq B$. Interestingly however, the Schrödinger potentials $\phi^\star$ and $\psi^\star = (\phi^\star)^+$ inherit properties from the cost function $c(\cdot, \cdot)$, which allow us to determine a reasonable choice of B based on the cost function. This is formalised in the following proposition.

Proposition 2 ([10, Lem. 2.7]) *Consider the dual eOT problem defined in Eq. (4). There exists Schrödinger potentials $\phi^\star, \psi^\star$ such that*

$$\|\phi^\star\|_{L^\infty(\mathcal{Y})} \leq \frac{3\|c\|_{L^\infty(\mathcal{X}\times\mathcal{Y})}}{2}; \quad \|\psi^\star\|_{L^\infty(\mathcal{X})} \leq \frac{3\|c\|_{L^\infty(\mathcal{X}\times\mathcal{Y})}}{2} \,.$$

The intriguing aspect of this proposition is the lack of a dependence on the regularisation parameter $\varepsilon > 0$. This proposition also suggests that solving the semi-dual problem for eOT over the space of functions $\phi \in L^2(\nu)$ such that $\|\phi\|_{L^\infty(\mathcal{Y})} \leq \frac{3\|c\|_{L^\infty(\mathcal{X}\times\mathcal{Y})}}{2}$ is sufficient to recover a Schrödinger potential.

While proj-SGA may appear fortuitous, this is actually a natural recommendation based on Lemma 2. This is because it can be obtained as the solution to a truncated local quadratic approximation of the semi-dual given below (truncated due to the restriction to $\mathcal{S}_B$), which is inspired by ISTA [3]:

$$\mathsf{M}^{\mathsf{proj-SGA}}_{\mathcal{S}_B}(\phi;\eta) = \underset{\overline{\phi}\in\mathcal{S}_B}{\operatorname{argmax}}\ J(\phi) + \langle \delta J(\phi), \overline{\phi} - \phi\rangle - \frac{1}{2\eta}\cdot\|\overline{\phi}-\phi\|^2_{L^2(\nu)}\ .$$

From Lemma 2, when the cost $c(\cdot,\cdot)$ is non-negative and the step size satisfies $\eta \leq \lambda(B)^{-1}$, $J(\mathsf{M}^{\text{proj-SGA}}_{\mathcal{S}_B}(\phi;\eta)) \geq J(\phi)$ for any $\phi \in \mathcal{S}_B$ as

$$\begin{aligned} J(\mathsf{M}^{\text{proj-SGA}}_{\mathcal{S}_B}(\phi;\eta)) &\geq J(\phi) + \left\langle \delta J(\phi), \mathsf{M}^{\text{proj-SGA}}_{\mathcal{S}_B}(\phi;\eta) - \phi\right\rangle - \frac{\lambda(B)}{2}\|\mathsf{M}^{\text{proj-SGA}}_{\mathcal{S}_B}(\phi;\eta) - \phi\|^2_{L^2(\nu)} \\ &\geq J(\phi) + \left\langle \delta J(\phi), \mathsf{M}^{\text{proj-SGA}}_{\mathcal{S}_B}(\phi;\eta) - \phi\right\rangle - \frac{1}{2\eta}\|\mathsf{M}^{\text{proj-SGA}}_{\mathcal{S}_B}(\phi;\eta) - \phi\|^2_{L^2(\nu)} \\ &\geq J(\phi)\ . \end{aligned}$$

The final step uses the optimality of $\mathsf{M}^{\text{proj-SGA}}_{\mathcal{S}_B}(\phi;\eta)$. Analogous to sign-SGA, the concavity of J results in the following non-asymptotic convergence guarantee for proj-SGA.

Theorem 7 (Formal version of Theorem 4) *Suppose $c(\cdot,\cdot)$ is a non-negative cost function such that $\lambda(B) < \infty$. Let $\{\phi^n\}_{n\geq 1}$ be the sequence of potentials generated according to the following recursion for $n \geq 0$:*

$$\phi^{n+1} = \mathsf{M}^{\text{proj-SGA}}_{\mathcal{S}_B}(\phi^n; \lambda(B)^{-1})\ ,$$

where $\lambda(B)$ is defined in Lemma 2. Then, for all $N \geq 1$, $\phi^N \in \mathcal{S}_B$ and for $\widetilde{\phi}^\star = \underset{\phi\in\mathcal{S}_B}{\operatorname{argmax}}\, J(\phi)$

$$J(\phi^N) - J(\widetilde{\phi}^\star) \geq -\frac{\lambda(B)\cdot\|\phi^0 - \widetilde{\phi}^\star\|^2_{L^2(\nu)}}{2N}\ .$$

Theorem 8 (Formal version of Theorem 5) *Suppose $c(\cdot,\cdot)$ is a non-negative cost function such that $\lambda(B) < \infty$. Consider the sequences $\{\phi^n\}_{n\geq 2}, \{\overline{\phi}^n\}_{n\geq 1}$ generated according to proj-SGA++. Then, for any $N \geq 1$,*

$$J(\overline{\phi}^N) - J(\widetilde{\phi}^\star) \geq -\frac{2\cdot\lambda(3B)\cdot\|\overline{\phi}^0 - \tilde{\phi}^\star\|^2_{L^2(\nu)}}{(N+1)^2};\quad \widetilde{\phi}^\star \in \underset{\phi\in\mathcal{S}_B}{\operatorname{argmax}}\ J(\phi)\ .$$

From Proposition 2, we know that the maximum value of the semi-dual J over $\mathcal{S}_B$ for $B = \frac{3\|c\|_{L^\infty(\mathcal{X}\times\mathcal{Y})}}{2}$ is $J(\phi^\star)$ for the Schrödinger potential $\phi^\star$. This implies that the sequences of semi-dual values generated by proj-SGA and proj-SGA++ with the appropriate step sizes converge to $J(\phi^\star)$ at a $\frac{1}{N}$ and $\frac{1}{N^2}$ rate respectively. A crude bound on $\lambda(B)$ in this setting is given by $\lambda(B) \leq \exp\left(B\cdot(\varepsilon^{-1}+2)\right)$.

We would like to mention that this is not the only accelerated method for the eOT problem. One can take advantage of the structure of $\mathcal{X}$ and $\mathcal{Y}$, particularly when they are discrete spaces to directly accelerate $\mathsf{M}^{\mathsf{SGA}}$ instead of $\mathsf{M}^{\text{proj-SGA}}_{\mathcal{S}_B}$ analogous to accelerating gradient descent to minimise a convex, smooth function in finite dimension. In that special case, the semi-dual is concave and also smooth in the canonical sense as implied by Lemma 1 due to the monotonicity of norms, and this leads to a rate that scales as $\frac{1}{N^2}$. Alternatively, one could possibly also design accelerated algorithms for minimising $\rho \mapsto \mathcal{L}_k(\rho_{\mathcal{Y}}; \nu)$ subject to the constraint $\rho \in \mathcal{Q}$ in contrast to proj-SGA++ which is based on the semi-dual. The constraint ensures that the solution has the form of the optimal coupling $\pi^\star$. However, designing an accelerated method for this problem in the flavour of accelerated MD for constrained optimisation can prove challenging, due to the non-convexity of the set $\mathcal{Q}$ and the need for linear combination of past iterates when considering momentum.

C.2.1. PROOF OF THEOREMS 7 AND 8

Before we give the proof, we lay out some preliminaries and intermediate results that will come in handy to prove Theorem 8 later.

The truncated quadratic approximation to J centered at $\phi \in L^2(\nu)$ that proj-SGA is based on:

$$\widetilde{J}_{\eta,\mathcal{S}_B}(\overline{\phi}; \phi) := J(\phi) + \left\langle \frac{\delta J(\phi)}{\nu}, \overline{\phi} - \phi \right\rangle_{L^2(\nu)} - \frac{1}{2\eta}\|\overline{\phi} - \phi\|^2_{L^2(\nu)} - \mathbb{I}_{\mathcal{S}_B}(\overline{\phi}) \,.$$

Above, $\mathbb{I}_{\mathcal{S}_B}$ is the convex indicator for $\mathcal{S}_B$ which evaluates to 0 if $\overline{\phi} \in \mathcal{S}_B$ and ∞ otherwise. Note that

$$\widetilde{J}_{\eta,\mathcal{S}_B}(\overline{\phi}; \phi) = J(\phi) + \frac{\eta}{2} \cdot \left\| \frac{\delta J(\phi)}{\nu} \right\|^2_{L^2(\nu)} - \frac{1}{2\eta} \cdot \left\| \overline{\phi} - \left(\phi + \eta \cdot \frac{\delta J(\phi)}{\nu} \right) \right\|^2_{L^2(\nu)} - \mathbb{I}_{\mathcal{S}_B}(\overline{\phi}) \,.$$

As a result, we have the alternate characterisation of $\mathsf{M}^{\text{proj-SGA}}_{\mathcal{S}_B}$ as

$$\mathsf{M}^{\text{proj-SGA}}_{\mathcal{S}_B}(\phi; \eta) = \underset{\overline{\phi} \in \mathcal{S}_B}{\operatorname{argmax}} \ \widetilde{J}_{\eta,\mathcal{S}_B}(\overline{\phi}; \phi) \,.$$

We use $\overline{J}_{\mathcal{S}_B}$ to denote the composite function $J + \mathbb{I}_{\mathcal{S}_B}$. Also recall that $\mathcal{S}_B = \{\phi \in L^2(\nu) : \|\phi\|_{L^\infty(\mathcal{Y})} \leq B\}$.

Lemma 5 *Let $\phi \in \mathcal{S}_{\bar{B}}$. Then, for $\eta \leq \lambda(\max\{B, \bar{B}\})^{-1}$,*

$$\overline{J}_{\mathcal{S}_B}(\mathsf{M}^{\text{proj-SGA}}_{\mathcal{S}_B}(\phi; \eta)) \geq \widetilde{J}_{\eta,\mathcal{S}_B}(\mathsf{M}^{\text{proj-SGA}}_{\mathcal{S}_B}(\phi; \eta); \phi) \,.$$

Lemma 6 *Let $\phi \in \mathcal{S}_{\bar{B}}$. For any $\overline{\phi} \in L^2(\nu)$ and $\eta \leq \lambda(\max\{B, \bar{B}\})^{-1}$, we have that*

$$\overline{J}_{\mathcal{S}_B}(\mathsf{M}^{\textit{proj-SGA}}_{\mathcal{S}_B}(\phi; \eta)) - \overline{J}_{\mathcal{S}_B}(\overline{\phi}) \geq \frac{1}{2\eta} \cdot \|\mathsf{M}^{\text{proj-SGA}}_{\mathcal{S}_B}(\phi; \eta) - \phi\|^2_{L^2(\nu)} + \frac{1}{\eta} \cdot \langle \mathsf{M}^{\text{proj-SGA}}_{\mathcal{S}_B}(\phi; \eta) - \phi, \phi - \overline{\phi} \rangle_{L^2(\nu)} \,.$$

Proof [Proof of Theorem 7] For $\phi^0 \in \mathcal{S}_B$, each step according to proj-SGA ensures that $\phi^n \in \mathcal{S}_B$ for all $n \geq 1$. For $\eta \leq \frac{1}{\lambda(B)}$, we have from Lemma 6 applied to $\overline{\phi} \leftarrow \widetilde{\phi}^\star$ and $\phi \leftarrow \phi^n$ for an arbitrary $n \geq 0$ that

$$\begin{aligned}
\overline{J}_{\mathcal{S}_B}(\phi^{n+1}) - \overline{J}_{\mathcal{S}_B}(\widetilde{\phi}^\star) &\geq \frac{1}{2\eta} \cdot \|\phi^{n+1} - \phi^n\|^2_{L^2(\nu)} + \frac{1}{\eta} \cdot \langle \phi^{n+1} - \phi^n, \phi^n - \widetilde{\phi}^\star \rangle_{L^2(\nu)} \\
&= \frac{1}{2\eta} \cdot \|\phi^{n+1} - \widetilde{\phi}^\star\|^2_{L^2(\nu)} - \frac{1}{2\eta} \cdot \|\phi^n - \widetilde{\phi}^\star\|^2_{L^2(\nu)} \,.
\end{aligned}$$

Summing both sides from $n = 0$ to $n = N - 1$ for $N \geq 1$ we get

$$\sum_{n=0}^{N-1} (\overline{J}_{\mathcal{S}_B}(\phi^{n+1}) - \overline{J}_{\mathcal{S}_B}(\widetilde{\phi}^\star)) \geq \frac{1}{2\eta} \cdot \|\phi^N - \widetilde{\phi}^\star\|^2_{L^2(\nu)} - \frac{1}{2\eta} \cdot \|\phi^0 - \widetilde{\phi}^\star\|^2_{L^2(\nu)} \,.$$

Additionally from Lemma 5, we have for the choice of η,

$$\overline{J}_{\mathcal{S}_B}(\phi^{n+1}) \geq \widetilde{J}_{\eta,\mathcal{S}_B}(\phi^{n+1}; \phi^n) \geq \widetilde{J}_{\eta,\mathcal{S}_B}(\phi^n; \phi^n) = \overline{J}_{\mathcal{S}_B}(\phi^n) \,.$$

Hence,

$$N \cdot (\overline{J}_{\mathcal{S}_B}(\phi^N) - \overline{J}_{\mathcal{S}_B}(\widetilde{\phi}^\star)) \geq -\frac{1}{2\eta} \cdot \|\phi^0 - \widetilde{\phi}^\star\|^2_{L^2(\nu)} \,.$$

Since $\phi^n \in \mathcal{S}_B$ for all $n \geq 0$, $\overline{J}_{\mathcal{S}_B}(\phi^n) = J(\phi^n)$. ■

C.2.2. Proof of Theorem 8

Prior to stating the proof for Theorem 8, we first make the following observations about the sequence $\{\overline{\phi^n}\}_{n \geq 0}$ and $\{t_n\}_{n \geq 1}$ generated by proj-SGA++. These are:

- for every $n \geq 0$, $\overline{\phi}^n \in \mathcal{S}_B$, and
- for every $n \geq 1$, $\frac{t_n - 1}{t_{n+1}} \in (0, 1)$.

A key step towards the proof of Theorem 8 is the following lemma, analogous to [3, Lem. 4.1].

Lemma 7 *Let $\{\overline{\phi}^n\}_{n \geq 1}$ be obtained from proj-SGA++. Define $v_n = \overline{J}(\phi^\star) - \overline{J}(\overline{\phi}^n)$ and $u_n = t_n \cdot \overline{\phi}^n - (t_n - 1) \cdot \overline{\phi}^{n-1} - \widetilde{\phi}^\star$. Then,*

$$\frac{2}{\lambda(3B)} \cdot (t_n^2 v_n - t_{n+1}^2 v_{n+1}) \geq \|u_{n+1}\|^2_{L^2(\nu)} - \|u_n\|^2_{L^2(\nu)} \,.$$

Proof [Proof of Theorem 8] Since $\lambda(3B) \geq \lambda(B)$ and $\phi^1, \overline{\phi}^1 \in \mathcal{S}_B$, Lemma 6 with $\phi \leftarrow \phi^1, \overline{\phi} \leftarrow \widetilde{\phi}^\star$ gives

$$\begin{aligned}\overline{J}(\overline{\phi}^1) - \overline{J}(\phi^\star) &\geq \frac{\lambda(3B)}{2} \cdot \|\overline{\phi}^1 - \phi^1\|^2_{L^2(\nu)} + \lambda(3B) \cdot \langle \overline{\phi}^1 - \phi^1, \phi^1 - \widetilde{\phi}^\star \rangle_{L^2(\nu)} \\ &= \frac{\lambda(3B)}{2} \cdot \left\{ \|\overline{\phi}^1 - \phi^\star\|^2_{L^2(\nu)} - \|\phi^1 - \phi^\star\|^2_{L^2(\nu)} \right\} \,.\end{aligned}$$

In the notation of Lemma 7,

$$-v_1 \geq \frac{\lambda(3B)}{2} \cdot \|u_1\|^2_{L^2(\nu)} - \frac{\lambda(3B)}{2} \cdot \|\overline{\phi}^0 - \widetilde{\phi}^\star\|^2_{L^2(\nu)} \,. \tag{18}$$

Telescoping the identity from Lemma 7 for $n = 1$ to $N - 1$ gives

$$\frac{2}{\lambda(3B)} \cdot (t_1^2 v_1 - t_N^2 v_N) \geq \|u_N\|^2_{L^2(\nu)} - \|u_1\|^2_{L^2(\nu)} \geq -\|u_1\|^2_{L^2(\nu)} \,.$$

Rearranging the terms, we have

$$v_N t_N^2 \le \frac{\lambda(3B)}{2} \cdot \|u_1\|^2_{L^2(\nu)} + t_1^2 v_1 \le \frac{\lambda(3B)}{2} \cdot \|\overline{\phi}^0 - \widetilde{\phi}^\star\|^2_{L^2(\nu)} ,$$

where the last step follows from Eq. (18). Since $t_N \ge \frac{N+1}{2}$, we have

$$v_N \le \frac{2 \cdot \lambda(3B) \cdot \|\overline{\phi}^0 - \phi^\star\|^2_{L^2(\nu)}}{(N+1)^2} .$$

■

Appendix D. Proofs of Lemmas 1 and 2

Lemmas 1 and 2 are corollaries of the following lemma; its proof is given in Appendix D.3.

Lemma 8 *Let $\phi, \overline{\phi} \in L^1(\nu)$. For $t \in [0,1]$, define $\tilde{\phi}_t := \phi + t \cdot (\overline{\phi} - \phi)$ and the conditional distribution $\rho_t(.;x)$ whose density is*

$$\rho_t(y;x) := \frac{\exp\left(\tilde{\phi}_t(y) - \frac{c(x,y)}{\varepsilon}\right)\nu(y)}{\int_{\mathcal{Y}} \exp\left(\tilde{\phi}_t(y') - \frac{c(x,y')}{\varepsilon}\right)\nu(y')\mathrm{d}y'} .$$

Then,

$$J(\overline{\phi}) - J(\phi) - \langle \delta J(\phi), \overline{\phi} - \phi \rangle = -\frac{1}{2}\int_0^1 \mathbb{E}_{x\sim\mu}\left[\mathbb{V}_{\rho_t(.;x)}[\overline{\phi} - \phi]\right] \, \mathrm{d}t .$$

D.1. Proof of Lemma 1

Proof From Lemma 8 and the definition of the variance,

$$\mathbb{V}_{\rho_t(.;x)}[\overline{\phi} - \phi] \le \int_{\mathcal{Y}} (\overline{\phi}(y) - \phi(y))^2 \, \rho_t(y;x)\mathrm{d}y \le \|\overline{\phi} - \phi\|^2_{L^\infty(\mathcal{Y})} .$$

The final inequality is by Hölder's inequality. Substituting this in the result of Lemma 8, we get

$$J(\overline{\phi}) - J(\phi) - \langle \delta J(\phi), \overline{\phi} - \phi \rangle \ge -\frac{\|\overline{\phi} - \phi\|^2_{L^\infty(\mathcal{Y})}}{2} .$$

The upper bound is due to the non-negativity of the variance. ■

D.2. Proof of Lemma 2

Proof From the convexity of $\mathcal{S}_B$, note that $\tilde{\phi}_t = \phi + t \cdot (\overline{\phi} - \phi) \in \mathcal{S}_B$. Since $\mathcal{S}_B \subset L^1(\nu)$, we have by Lemma 8 that

$$\begin{aligned} J(\overline{\phi}) - J(\phi) - \langle \delta J(\phi), \overline{\phi} - \phi \rangle &= -\frac{1}{2}\int_0^1 \mathbb{E}_{x\sim\mu}\left[\mathbb{V}_{\rho_t(.;x)}[\overline{\phi} - \phi]\right] \, \mathrm{d}t \\ &\ge -\frac{1}{2}\int_0^1 \left\{\int_{\mathcal{X}}\int_{\mathcal{Y}} (\overline{\phi}(y) - \phi(y))^2 \rho_t(y;x)\mu(x) \, \mathrm{d}x\mathrm{d}y\right\} \, \mathrm{d}t . \end{aligned}$$

By Fubini's theorem, we can first compute $\int_{\mathcal{X}} \rho_t(y;x)\mu(x)\mathrm{d}x$ and then integrate w.r.t. $\mathcal{Y}$.

$$\begin{aligned}
\int_{\mathcal{X}} \rho_t(y;x)\mu(x)\mathrm{d}x &= \int_{\mathcal{X}} \frac{\exp\left(\tilde{\phi}_t(y) - \frac{c(x,y)}{\varepsilon}\right)\nu(y)}{\int_{\mathcal{Y}} \exp\left(\tilde{\phi}_t(y') - \frac{c(x,y')}{\varepsilon}\right)\nu(y')\mathrm{d}y'}\mu(x)\mathrm{d}x \\
&= \mathbb{E}_{x\sim\mu}\left[\exp\left(\tilde{\phi}_t(y) - \frac{c(x,y)}{\varepsilon}\right)\cdot \mathbb{E}_{y'\sim\nu}\left[\exp\left(\tilde{\phi}_t(y') - \frac{c(x,y')}{\varepsilon}\right)\right]^{-1}\right]\cdot\nu(y) \\
&\overset{(a)}{\le} \mathbb{E}_{(x,y')\sim\mu\otimes\nu}\left[\exp\left(\frac{c(x,y')-c(x,y)}{\varepsilon} + \tilde{\phi}_t(y) - \tilde{\phi}_t(y')\right)\right]\cdot\nu(y) \\
&\overset{(b)}{\le} \underbrace{e^{2B}\cdot\mathbb{E}_{(x,y')\sim\mu\otimes\nu}\left[\exp\left(\frac{c(x,y')}{\varepsilon}\right)\right]}_{\lambda(B)}\cdot\nu(y)\ .
\end{aligned}$$

Step (a) uses Jensen's inequality, and step (b) uses the fact that for y, y', $\tilde{\phi}_t(y) - \tilde{\phi}_t(y') \le 2B$ for y, y' almost everywhere. Substituting this in the result of Lemma 8, we have

$$\begin{aligned}
J(\overline{\phi}) - J(\phi) - \langle \delta J(\phi), \overline{\phi} - \phi\rangle &\ge -\frac{1}{2}\int_0^1 \left\{\int_{\mathcal{Y}} (\overline{\phi}(y) - \phi(y))^2\cdot\lambda(B)\cdot\nu(y)\mathrm{d}y\right\}\,\mathrm{d}t \\
&= -\frac{\lambda(B)\cdot\|\overline{\phi}-\phi\|^2_{L^2(\nu)}}{2}\ .
\end{aligned}$$

■

D.3. Proof of Lemma 8

Proof Recall that the first variation of the semi-dual J is

$$\delta J(\phi)(y) = \nu(y) - \pi(\phi,\phi^+)_{\mathcal{Y}}(y) = \int_{\mathcal{X}} \nu(y)\mu(x)\mathrm{d}x - \int_{\mathcal{X}} \frac{\exp\left(\phi(y) - \frac{c(x,y)}{\varepsilon}\right)\nu(y)}{\int_{\mathcal{Y}} \exp\left(\phi(y') - \frac{c(x,y')}{\varepsilon}\right)\nu(y')\mathrm{d}y'}\mu(x)\mathrm{d}x\ .$$

For a fixed $x \in \mathcal{X}$, consider the function

$$j_x(\phi)(y) := \nu(y) - \frac{\exp\left(\phi(y) - \frac{c(x,y)}{\varepsilon}\right)\nu(y)}{\int_{\mathcal{Y}} \exp\left(\phi(y') - \frac{c(x,y')}{\varepsilon}\right)\nu(y')\mathrm{d}y'}\ .$$

and hence for any $\phi, \overline{\phi} \in \mathcal{S}_B$, we have

$$j_x(\phi)(y) - j_x(\overline{\phi})(y) = \frac{\exp\left(\overline{\phi}(y) - \frac{c(x,y)}{\varepsilon}\right)\nu(y)}{\int_{\mathcal{Y}} \exp\left(\overline{\phi}(y') - \frac{c(x,y')}{\varepsilon}\right)\nu(y')\mathrm{d}y'} - \frac{\exp\left(\phi(y) - \frac{c(x,y)}{\varepsilon}\right)\nu(y)}{\int_{\mathcal{Y}} \exp\left(\phi(y') - \frac{c(x,y')}{\varepsilon}\right)\nu(y')\mathrm{d}y'}$$

Note that $j_x(\phi) - j_x(\overline{\phi}) = \rho_1(.;x) - \rho_0(.;x) = \int_0^1 \dot{\rho}_t(.;x)\mathrm{d}t$. We have by direct calculation that

$$\frac{\mathrm{d}}{\mathrm{d}t}\rho_t(y;x) \equiv \dot{\rho}_t(y;x) = \left\{(\overline{\phi}(y) - \phi(y)) - \int_{\mathcal{Y}} (\overline{\phi}(y') - \phi(y'))\rho_t(y';x)\mathrm{d}y'\right\}\,\rho_t(y;x)\mathrm{d}y\ .$$

Consequently,

$$\begin{aligned}
\langle j_x(\phi) - j_x(\overline{\phi}), \phi - \overline{\phi} \rangle &= \int_{\mathcal{Y}} (\phi(y) - \overline{\phi}(y)) \cdot (\rho_1(y;x) - \rho_0(y;x)) \, \mathrm{d}y \\
&= \int_{\mathcal{Y}} \int_0^1 (\phi(y) - \overline{\phi}(y)) \cdot \dot{\rho}_t(y;x) \, \mathrm{d}t\mathrm{d}y \\
&= - \int_0^1 \int_{\mathcal{Y}} (\overline{\phi}(y) - \phi(y))^2 \, \rho_t(y;x) \, \mathrm{d}y\mathrm{d}t \\
&\quad + \int_0^1 \left\{ \int_{\mathcal{Y}} (\overline{\phi}(y) - \phi(y)) \cdot \rho_t(y;x) \mathrm{d}y \right\}^2 \mathrm{d}t \\
&= - \int_0^1 \mathbb{V}_{\rho_t(.;x)}[\overline{\phi} - \phi] \, \mathrm{d}t \, . \qquad (19)
\end{aligned}$$

Taking the expectation w.r.t. μ on both sides and by Fubini's theorem, we have

$$\langle \delta J(\phi) - \delta J(\overline{\phi}), \phi - \overline{\phi} \rangle = - \int_0^1 \mathbb{E}_{x \sim \mu} \left[\mathbb{V}_{\rho_t(.;x)}[\overline{\phi} - \phi] \right] \, \mathrm{d}t \, . \qquad (20)$$

Define $\tilde{J}_t = J(\tilde{\phi}_t) - \langle \delta J(\phi), \tilde{\phi}_t \rangle$. By the chain rule,

$$\dot{\tilde{J}}_t = \langle \delta J(\tilde{\phi}_t), \overline{\phi} - \phi \rangle - \langle \delta J(\phi), \overline{\phi} - \phi \rangle = \langle \delta J(\tilde{\phi}_t) - \delta J(\phi), \overline{\phi} - \phi \rangle \, .$$

By the fundamental theorem of calculus,

$$\begin{aligned}
J(\overline{\phi}) - J(\phi) - \langle \delta J(\phi), \overline{\phi} - \phi \rangle &= J_1 - J_0 \\
&= \int_0^1 \dot{\tilde{J}}_s \, \mathrm{d}s \\
&= \int_0^1 \langle \delta J(\tilde{\phi}_s) - \delta J(\phi), \overline{\phi} - \phi \rangle \, \mathrm{d}s \\
&= \int_0^1 \frac{1}{s} \cdot \langle \delta J(\tilde{\phi}_s) - \delta J(\phi), \tilde{\phi}_s - \phi \rangle \, \mathrm{d}s \\
&= - \int_0^1 \frac{1}{s} \cdot \int_0^1 \mathbb{E}_{x \sim \mu} \left[\mathbb{V}_{\rho_t(.;x)}[\tilde{\phi}_s - \phi] \right] \mathrm{d}t \, \mathrm{d}s \\
&= - \int_0^1 \int_0^1 s \cdot \mathbb{E}_{x \sim \mu} \left[\mathbb{V}_{\rho_t(.;x)}[\overline{\phi} - \phi] \right] \mathrm{d}t \, \mathrm{d}s \\
&= - \frac{1}{2} \int_0^1 \mathbb{E}_{x \sim \mu} \left[\mathbb{V}_{\rho_t(.;x)}[\overline{\phi} - \phi] \right] \, \mathrm{d}t \, .
\end{aligned}$$

■

Appendix E. Conclusion

In this work, we systemically synthesise a variety of viewpoints on algorithms for eOT – specifically those surrounding the Sinkhorn algorithm. This synthesis, centered around infinite-dimensional optimisation, leads to a collection of novel methods based on the dual formulation of the eOT problem.

We also see how the viewpoints contribute to provable guarantees for these methods, which notably are not based on any strict assumptions on the marginals μ, ν. It would be interesting to see how these guarantees can be improved; going past the bounded kernel condition in Theorem 2, and bounded costs condition to establish a desirable form of smoothness for the semi-dual in Lemma 2. While this not only encourages a new theoretical perspective for solving the eOT problem, we hope that this work also spurs the development of new practical methods other than the Sinkhorn algorithm and its derivatives for the eOT problem.