Designing Algorithms for Entropic Optimal Transport from an Optimisation Perspective

Qijia Jiang
UC Davis
qjang@ucdavis.edu

Vishwak Srinivasan
MIT
vishwaks@mit.edu

Abstract

In this work, we give a collection of novel methods for the entropic-regularised optimal transport problem, which are inspired by existing mirror descent interpretations of the Sinkhorn algorithm used for solving this problem. These are fundamentally proposed from an optimisation perspective: either based on the associated semi-dual problem, or based on solving a non-convex constrained problem over subset of joint distributions. This optimisation viewpoint results in non-asymptotic rates of convergence for the proposed methods under minimal assumptions on the problem structure. We also propose a momentum-equipped method with provable accelerated guarantees through this viewpoint, akin to those in the Euclidean setting. The broader framework we develop based on optimisation over the joint distributions also finds an analogue in the dynamical Schrödinger bridge problem.

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. This also results in a notion of discrepancy between μ and ν that complements the information-theoretic discrepancy measures like the total variation distance or the Kullback-Leibler divergence. The resulting Wasserstein distance finds applications in image processing, operations research and others [San15], while also serving as an important ingredient behind many theoretical developments [Vil03, AGS08]. Nevertheless, due to both computational and statistical reasons, methods based on the OT problem have not seen widespread use in modern machine learning and statistics.

However, these bottlenecks can be alleviated by adding an entropy regularisation to the OT problem, referred to as the eOT problem. One popular algorithm for this problem is the *Sinkhorn* algorithm [SK67], and has seen renewed interest due to its application in large-scale data science problems [Cut13]. The first non-asymptotic rate of convergence for Sinkhorn was given in [FL89] based on a matrix scaling perspective, and it continues to see newer and sharper analyses to this day. More recently, there has been a growing body of literature that view the Sinkhorn algorithm through an alternative (infinite-dimensional) optimisation lens [Mis19, MP20, Lég21, AFKL22, DKPS23, RKHK24]. In this paper, we draw inspiration from these refreshing viewpoints to design new methods for the eOT problem with provable guarantees. The main contributions are summarised below.

(1) We investigate properties of a *semi-dual* formulation of the eOT problem (Section 3.1). These properties naturally lead to two new steepest ascent methods (Section 3.2) for maximizing the semi-dual, and we show that they converge to the optimal dual solution in terms of objective function value at a $1/N$ rate under minimal assumptions where N is the number of iterations.

(2) We show that the particular form of the updates taken by these two methods admits a primal interpretation (Section 3.3), which in turn lead to a general framework Φ-match that allow us to design iterative algorithms that can be viewed as minimising a discrepancy metric between the $\mathcal{Y}$ marginal of a joint distribution and ν while ensuring its $\mathcal{X}$ marginal is μ. Specializing to the case

Preprint. Under review.

of MMD, we establish a rate of convergence of $1/N$ for these new algorithms (Section 4.1). This is analogous to an interpretation of the Sinkhorn algorithm that works with the KL divergence.

(3) We propose two extensions of our findings from (1) and (2): (a) an accelerated scheme based on one of the steepest ascent methods in (1) that we show converges at a $1/N^2$ rate (Section 5.1); (b) a generalization of the class of methods in (2) to path space (Section 5.2).

Related work Traditional analyses view the Sinkhorn algorithm as either alternating projection on the two marginals or block maximization on the two dual potentials. Based on this, classical analysis of the Sinkhorn algorithm with Hilbert's projective metric and Birkhoff's theorem [FL89] renders a linear convergence with contraction rate $1 - \Theta(e^{-\|c\|_\infty/\varepsilon})$, where ε is the regularisation parameter. An important limitation of this analysis is that the rate becomes exponentially slow as either $\|c\|_\infty \to \infty$ or as $\varepsilon \to 0$. More recent analyses have focused on the Sinkhorn algorithm in the setting where $\mathcal{X}$ and $\mathcal{Y}$ are discrete spaces [ANWR17, LHJ22, DGK18]. More relevant to us, in the continuous setting, many probabilistic approaches have been taken for analyzing the algorithm: [CCGT24] leverages 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 [CDV24] for semi-concave bounded costs while still maintaining a contraction rate that only deteriorates polynomially in ε. While the literature is continuously growing, most recent results place assumptions on the growth of the cost or decay of the tails of μ, ν to obtain exponential convergence guarantees.

In contrast, the advantage of taking the optimisation route, specifically viewing the Sinkhorn algorithm as performing ∞-dimensional mirror descent [Lég21, AFKL22, RKHK24], is that it provides guarantee under *minimal* assumptions. For comparison, these works furnish a discrete-time iteration complexity of $\Theta(1/N\varepsilon)$, which if additionally assuming bounded cost [AFKL22] also recovers the same rate as classical Hilbert metric analysis. We achieve the same rates as these works while significantly expanding the scope of algorithm design and shed more light on the eOT problem, unifying both the primal and dual perspectives. There has also been interest in designing alternative algorithms for the eOT problem, among them [CLP23] that design Wasserstein gradient flow dynamics over the submanifold of $\Pi(\mu, \nu)$ which borrow tools from SDE/PDE for analysis.

2 Background

We begin by stating notation that we adopt throughout this work.

Let $\mathcal{X}, \mathcal{Y}$ be subsets of $\mathbb{R}^d$ whose product is $\mathcal{X} \times \mathcal{Y}$. For a set $\mathcal{Z}$, the set of probability measures over $\mathcal{Z}$ is denoted by $\mathcal{P}(\mathcal{Z})$, and for $\rho \in \mathcal{P}(\mathcal{Z})$, the density function of ρ w.r.t. to the Lebesgue measure of $\mathcal{Z}$ if it exists is denoted by $\mathrm{d}\rho$. 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 and ρ' is denoted by $\mathrm{d}_{\mathrm{KL}}(\rho\|\rho')$. For a function $f : \mathcal{Z} \to \mathbb{R}$, its L^p-norm w.r.t. ρ is denoted by $\|f\|_{L^p(\rho)}$, and when ρ is replaced with $\mathcal{Z}$, the underlying measure is understood to be the Lebesgue measure. For a functional $\mathcal{F} : \mathcal{P}(\mathcal{Z}) \to \mathbb{R}$, we call $\delta\mathcal{F}(\rho)$ as its first variation at $\rho \in \mathcal{P}(\mathcal{Z})$, and this is the function that satisfies [San15, Def. 7.12]

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

For another function $g : \mathcal{Z} \to \mathbb{R}$, $\langle f, g\rangle_{L^2(\rho)} = \int_{\mathcal{Z}} fg \,\mathrm{d}\rho$. When the subscript in the norm (and the inner product) is omitted, it corresponds to $L^2(\mathcal{Z})$-norm (and inner product, resp.). Following the convention in [Rud87], we overload L^p to also denote the space of functions whose L^p norm is finite. 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. Given two distributions $\mu \in \mathcal{P}(\mathcal{X}), \nu \in \mathcal{P}(\mathcal{Y})$, $\pi \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ is a coupling of μ and ν if $\pi_{\mathcal{X}} = \mu$ and $\pi_{\mathcal{Y}} = \nu$. The set of all couplings of μ and ν is denoted by $\Pi(\mu, \nu)$.

The entropic optimal transport problem

Let $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ be a lower semi-continuous cost function. The primal formulation of the entropic optimal transport (abbrev. eOT) problem is stated below.

$$\inf_{\pi\in\Pi(\mu,\nu)} \iint c(x, y)\mathrm{d}\pi(x, y) + \varepsilon \cdot \mathrm{d}_{\mathrm{KL}}(\pi\|\mu \otimes \nu) =: \mathrm{OT}_\varepsilon(\mu, \nu; c)\,. \tag{1}$$

In the definition above, $\varepsilon > 0$ is the regularisation parameter. When $\varepsilon = 0$, this corresponds to the (canonical) optimal transport problem (see Appendix A.1 for a more detailed overview). On the other hand, it can be seen that as $\varepsilon \to \infty$, the solution to $\mathrm{OT}_\varepsilon(\mu, \nu; c)$ tends to the product measure $\mu \otimes \nu$. The eOT problem is also directly related to the *static Schrödinger bridge problem* [Lé14, Def. 2.2]. Specifically, let R be the *reference* measure whose density is $\mathrm{d}R \propto \exp\left(-c/\varepsilon\right) \mathrm{d}(\mu \otimes \nu)$. Then,

$$\mathrm{OT}_\varepsilon(\mu, \nu; c) = \inf_{\pi \in \Pi(\mu,\nu)} \varepsilon \cdot \mathrm{d}_{\mathrm{KL}}(\pi \| R)\,. \tag{2}$$

The above problem is a (strictly) convex minimisation problem since $\Pi(\mu, \nu)$ is convex, and $\pi \mapsto \mathrm{d}_{\mathrm{KL}}(\pi \| R)$ is strictly convex. Moreover, the eOT problem has a unique solution $\pi^\star \in \Pi(\mu, \nu)$ [Lé14, NW22] 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*, 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[1]. The eOT problem has a dual formulation with zero duality gap, and is *unconstrained* as stated below.

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

Any solution of the dual problem above corresponds to a pair of Schrödinger potentials and vice versa [Nut21, 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 $\mathrm{OT}_1(\mu, \nu; c/\varepsilon)$.

We highlight that $\mathrm{OT}_\varepsilon(\mu, \nu; c)$ is *biased* relative to $\mathrm{OT}_0(\mu, \nu; c)$. To address this, a debiasing strategy is proposed in [FSV+19] which results in the *Sinkhorn divergence*, and also metrizes convergence in law. In addition to the computational benefits from the Sinkhorn algorithm, the sample complexity of estimating the Sinkhorn divergence using samples from μ and ν scales much better than $\mathrm{OT}_0(\mu, \nu; c)$ for a variety of costs [GCB+19, MNW19, CRL+20].

3 The semi-dual problem in eOT

In this section, we introduce the semi-dual problem. This was originally discussed in [GCPB16] to motivate the use of stochastic algorithms to solve the eOT problem, and was investigated in more detail by [CP18]. Following [Lég21], for any $\phi \in L^1(\nu)$ and $\psi \in L^1(\mu)$, we define ϕ^+ and ψ^- as

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

From Eq. (3), we see that any pair of Schrödinger potentials $(\phi^\star, \psi^\star)$ corresponding to $\mathrm{OT}_1(\mu, \nu; c/\varepsilon)$ satisfies $\psi^\star = \phi^{\star +}$ and $\phi^\star = \psi^{\star -}$. Therefore, it would be sufficient to solve one of $\sup_{\phi \in L^1(\nu)} D(\phi^+, \phi)$ or $\sup_{\psi \in L^1(\mu)} D(\psi, \psi^-)$. Without loss of generality, we work with the maximisation problem over ϕ, and this is referred to as the *semi-dual problem* for eOT. Additionally, we note that for any $\phi \in L^1(\nu)$, the objective J of the semi-dual satisfies

$$J(\phi) := D(\phi^+, \phi) = \sup_{\psi \in L^1(\nu)} D(\psi, \phi) = \int_{\mathcal{Y}} \phi(y)\mathrm{d}\nu(y) - \int_{\mathcal{X}} \phi^+(x)\mathrm{d}\mu(x) \tag{5}$$

[1] If $\psi^\star$ and $\phi^\star$ are Schrödinger potentials, then $\psi^\star + \beta$ and $\phi^\star + \beta$ are also Schrödinger potentials for any constant $\beta \in \mathbb{R}$.

and hence the semi-dual problem for eOT can be viewed as explicitly eliminating ψ via a partial maximisation of $D(\psi, \phi)$ in the dual problem Eq. (4). The objective J is referred to as the *semi-dual*, and is the focus of this work. For $\phi \in L^1(\nu)$, define the joint distribution $\pi(\phi, \phi^+)$ with density

$$\mathrm{d}\pi(\phi, \phi^+)(x, y) = \exp\left(\phi(y) - \phi^+(x) - \frac{c(x, y)}{\varepsilon}\right) \mathrm{d}\nu(y)\mathrm{d}\mu(x) \,. \tag{6}$$

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 for $\phi \leftarrow \phi^\star$ corresponds to the unique solution of the eOT problem Eq. (3) by virtue of $(\phi^\star)^+ = \psi^\star$.

3.1 Properties of the semi-dual J

In this subsection, we state desirable properties of the semi-dual J which underlie the methods that we propose and study in this paper. As a prelude to this, we state two aspects of the semi-dual J that are more general. First, note that the shift property of the Schrödinger potentials is captured by the semi-dual as well; in particular, $J(\phi) = J(\phi + C \cdot \mathbf{1})$ where $\mathbf{1}$ is the all-ones function. Second, the first variation δJ of the semi-dual J can be succinctly expressed in terms of $\mathrm{d}\pi(\phi, \phi^+)_{\mathcal{Y}}$ [Lég21, Lem. 1]; for any $\phi \in L^1(\nu)$,

$$\delta J(\phi) = \mathrm{d}\nu - \mathrm{d}\pi(\phi, \phi^+)_{\mathcal{Y}} \,. \tag{7}$$

The first desirable property that we discuss is the concavity of J.

Lemma 1. *Let $\phi, \overline{\phi} \in L^1(\nu)$. Then, the semi-dual J (Eq. (5)) satisfies*

$$J(\overline{\phi}) \leq J(\phi) + \left\langle \delta J(\phi), \overline{\phi} - \phi \right\rangle \,.$$

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. However, guided by the principles in finite-dimensional optimisation, the "right" gradient ascent algorithm depends on the growth properties of the semi-dual which also enables derivation of non-asymptotic rates of convergence for the algorithm. Motivated by this, we state certain growth properties of the semi-dual J that we derive in this paper.

Lemma 2. *Let $\phi, \overline{\phi} \in L^1(\nu)$. Then, the semi-dual J (Eq. (5)) satisfies*

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

Lemma 3. *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,*

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

Proposition 1 demonstrates how Lemmas 2 and 3 imply growth properties for the semi-dual J.

Proposition 1. *Let $\mathcal{S} \subseteq L^1(\nu)$. If $\left\langle \delta J(\overline{\phi}) - \delta J(\phi), \overline{\phi} - \phi \right\rangle \geq \omega(\overline{\phi}, \phi)$ for all $\overline{\phi}, \phi \in \mathcal{S}$ for some $\omega : \mathcal{S} \times \mathcal{S} \to \mathbb{R}$, then*

$$J(\overline{\phi}) \geq J(\phi) + \left\langle \delta J(\phi), \overline{\phi} - \phi \right\rangle + \int_0^1 \frac{\omega(\phi + t \cdot (\overline{\phi} - \phi), \phi)}{t} \mathrm{d}t \,.$$

In the setting of Lemma 3, it remains to understand what is a reasonable choice of B. For B that is too small, it is possible that the Schrödinger potential would not satisfy $\|\phi^\star\|_{L^\infty(\mathcal{Y})} \leq B$. Interestingly, the Schrödinger potentials $\phi^\star$ and $\psi^\star = (\phi^\star)^+$ inherit properties from the cost function $c(\cdot, \cdot)$. This is formalised in the following proposition.

Proposition 2 ([MG20, 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 $\phi \in L^2(\nu)$ such that $\|\phi\|_{L^\infty(\mathcal{Y})} \leq \frac{3\|c\|_{L^\infty(\mathcal{X}\times\mathcal{Y})}}{2}$ will recover a Schrödinger potential.

3.2 Two steepest ascent methods for solving the semi-dual problem

Following the discussion in the previous subsection, we propose a collection of novel methods to solve the eOT problem here. These methods fundamentally perform different versions of gradient ascent on the semi-dual J while adapting to the growth condition implied by Lemmas 2 and 3. From a suitable initialisation ϕ^0, these methods generate a sequence of functions $\{\phi^n\}_{n\geq 1}$ that satisfy the recurrence $\phi^{n+1} = \mathsf{M}(\phi^n; \eta)$ for a choice of map M and step size $\eta > 0$.

Signed Semi-dual Gradient Ascent This method is based on the map $\mathsf{M} \leftarrow \mathsf{M}^{\text{sign-SGA}}$ defined as

$$\mathsf{M}^{\text{sign-SGA}}(\phi; \eta) := \phi + \eta \cdot \|\delta J(\phi)\|_{L^1(\mathcal{Y})} \cdot \operatorname{sign}(\delta J(\phi)) \,. \tag{sign-SGA}$$

Note that for any $\phi \in L^1(\nu)$, $\mathsf{M}^{\text{sign-SGA}}(\phi; \eta) \in L^1(\nu)$. Also, for a sufficiently small $\eta > 0$, the growth property implied by Lemma 2 and Proposition 1 asserts that $J(\mathsf{M}^{\text{sign-SGA}}(\phi; \eta)) > J(\phi)$. The concavity of J enables us to prove a stronger version of this assertion in the form of a non-asymptotic convergence rate. Let $\phi^0 \in L^1(\nu)$ such that $\phi^0(y_{\texttt{anc}}) = 0$ for a fixed anchor point $y_{\texttt{anc}} \in \mathcal{Y}$.

Lemma 4. *Define* $\mathcal{T}_0 := \{\phi \in L^1(\nu) : \phi(y_{\texttt{anc}}) = 0, J(\phi) \geq J(\phi^0)\}$. *Let* $\{\phi^n\}_{n\geq 1}$ *be the sequence of potentials generated according to*

$$\phi^{n+1} = \phi^{n+1/2} - (\phi^{n+1/2}(y_{\texttt{anc}}) - \phi^n(y_{\texttt{anc}})) \cdot \mathbf{1} \qquad \textit{where } \phi^{n+1/2} = \mathsf{M}^{\text{sign-SGA}}(\phi^n; 1)$$

for $n \geq 0$. *Then, for all* $N \geq 1$, $\phi^N \in \mathcal{T}_0$ *and for* $\widetilde{\phi}^\star = \operatorname{argmax}\ \{J(\phi) : \phi \in \mathcal{T}_0\}$ *we have*

$$J(\phi^N) - J(\widetilde{\phi}^\star) \geq -\frac{\operatorname{diam}(\mathcal{T}_0; L^\infty(\mathcal{Y}))^2}{2(N+1)} \,; \qquad \operatorname{diam}(\mathcal{T}_0; L^\infty(\mathcal{Y})) = \sup_{\overline{\phi}, \phi \in \mathcal{T}_0} \|\overline{\phi} - \phi\|_{L^\infty(\mathcal{Y})} \,.$$

If $\phi^\star$ is a Schrödinger potential, then $\phi^\star - \phi^\star(y_{\texttt{anc}}) \cdot \mathbf{1}$ is also a maximiser of J that lies in $\mathcal{T}_0$. This fact shows that the sequence $\{J(\phi^n)\}_{n\geq 1}$ converges to the maximum of the semi-dual J.

Projected Semi-dual Gradient Ascent The map $\mathsf{M} \leftarrow \mathsf{M}_{\mathcal{S}}^{\text{proj-SGA}}$ associated with this method is additionally dependent on a set $\mathcal{S} \subset L^1(\nu)$

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

While this map may appear fortuitous, it is actually the solution to a truncated quadratic approximation of the semi-dual centered at ϕ, inspired by ISTA [BT09]. Let $\mathcal{S}_B := \{\phi \in L^2(\nu) : \|\phi\|_{L^\infty(\mathcal{Y})} \leq B\} \subset L^1(\nu)$. From Lemma 3 and Proposition 1, when the cost $c(\cdot, \cdot)$ is non-negative and the step size $\eta \leq \lambda(B)^{-1}$, $J(\mathsf{M}_{\mathcal{S}_B}^{\text{proj-SGA}}(\phi; \eta)) > J(\phi)$. Analogous to sign-SGA, the concavity of J results in the following non-asymptotic convergence guarantee for proj-SGA.

Lemma 5. *Let* $\{\phi^n\}_{n\geq 1}$ *be the sequence of potentials generated as* $\phi^{n+1} = \mathsf{M}_{\mathcal{S}_B}^{\text{proj-SGA}}(\phi^n; \lambda(B)^{-1})$ *for* $n \geq 0$ *from* $\phi^0 \in \mathcal{S}_B$. *Then, for all* $N \geq 1$, $\phi^N \in \mathcal{S}_B$ *and for* $\widetilde{\phi}^\star = \operatorname{argmax}\ \{J(\phi) : \phi \in \mathcal{S}_B\}$,

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

From the discussion at the end of Section 3.1, we know that for a bounded cost $c(\cdot, \cdot)$ setting $B = \frac{3\|c\|_{L^\infty(\mathcal{X}\times\mathcal{Y})}}{2}$ would result in $\widetilde{\phi}^\star$ coinciding with a Schrödinger potential $\phi^\star$.

3.3 Ordinary gradient ascent of the semi-dual as matching marginals

In light of sign-SGA and proj-SGA, a natural question to ask is: can we perform ordinary gradient ascent of the semi-dual J and establish rates of convergence for this method? In the notation of the preceding subsection, define $\mathsf{M} \leftarrow \mathsf{M}^{\text{SGA}}$ where

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

The update $\mathsf{M}^{\text{SGA}}(\phi)$ was previously considered by [GCPB16] for discrete spaces $\mathcal{X}$ and $\mathcal{Y}$, where ϕ can be represented as a finite-dimensional vector. In this setting, Lemma 2 implies a standard notion of smoothness for the semi-dual J [Nes18, Chap. 2] by the monotonicity of norms, and

consequently results in an *assumption-free* non-asymptotic convergence guarantee. 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 true. To circumvent this, we view the update as minimising the "difference" between $\pi(\phi, \phi^+)_\mathcal{Y}$ and ν. Specifically, from the form of $\delta J(\phi)$ for any $L^1(\nu)$ in Eq. (7), we can rewrite the update as

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

Note that a fixed point $\phi^\star$ of this update satisfies $\pi(\phi^\star, \phi^{\star+})_\mathcal{Y} = \nu$. This corresponds to a maximiser of J as this is equivalent to $\delta J(\phi^\star) = 0$. Relatedly, the Sinkhorn algorithm can be expressed in terms of an update map $\mathsf{M} \leftarrow \mathsf{M}^{\mathsf{Sinkhorn}}$ as shown in [RKHK24] that is defined as

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

In fact, both $\mathsf{M}^{\mathsf{SGA}}$ and $\mathsf{M}^{\mathsf{Sinkhorn}}$ can be captured as instantiations of a more general update described below. Let $\Phi : L^1(\mathcal{Y}) \to L^\infty(\mathcal{Y})$ be an operator that returns a positive function. Then, the update $\mathsf{M}^{\Phi\text{-match}}$ is defined as

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

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 Sinkhorn with a step size η (also called η-Sinkhorn in [RKHK24]). In the next section, we show how Φ-match can be viewed as an update for minimising a discrepancy between marginal distributions, and lead to convergence rates for SGA.

4 Interpretations of Φ-match

The interpretations of Φ-match that we discuss here are motivated by recent work in understanding the Sinkhorn algorithm (Sinkhorn) [AFKL22, RKHK24]. In essence, these prior work view the Sinkhorn algorithm as minimising $\mathrm{d}_{\mathrm{KL}}(\pi_\mathcal{Y}, \nu)$ over the domain $\mathcal{Q}$ formed by joint distributions $\pi \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ that take the form in Eq. (6). Note that $\mathcal{Q}$ is *not* a convex subset of $\mathcal{P}(\mathcal{X} \times \mathcal{Y})$ owing to the factorisation structure. Despite this, an intriguing observation about the Sinkhorn algorithm is that it ensures that the iterates lie in this set $\mathcal{Q}$.

Here, we show that Φ-match also operates in the same way while minimising an objective that is not the KL divergence but a discrepancy which depends on Φ. We specifically show that Φ-match can be interpreted in the following ways: as an alternating projection scheme and as a local greedy method analogous to gradient descent. Notably, these discussions do not involve the semi-dual, and are solely based in $\mathcal{P}(\mathcal{X} \times \mathcal{Y})$. These interpretations allow us to not only obtain rates of convergence for SGA, but also its kernelised version (k-SGA).

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

$$\mathsf{project}_\mathcal{Y}(\pi; \Phi) := \operatorname*{argmin}_{\bar{\pi}} \left\{ \mathrm{d}_{\mathrm{KL}}(\bar{\pi} \| \pi) : \mathrm{d}\bar{\pi}_\mathcal{Y} \propto \mathrm{d}\pi_\mathcal{Y} \cdot \frac{\Phi(\mathrm{d}\nu)}{\Phi(\mathrm{d}\pi_\mathcal{Y})} \right\}, \tag{8a}$$

$$\mathsf{project}_{\mathcal{X},\mu}(\pi', \pi; \eta) := \operatorname*{argmin}_{\bar{\pi}} \left\{ \eta \cdot \mathrm{d}_{\mathrm{KL}}(\bar{\pi} \| \pi') + (1 - \eta) \cdot \mathrm{d}_{\mathrm{KL}}(\bar{\pi} \| \pi) : \bar{\pi}_\mathcal{X} = \mu \right\} . \tag{8b}$$

For a given $\pi \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$, $\mathsf{project}_\mathcal{Y}$ can be seen as correcting the $\mathcal{Y}$-marginal of π towards ν (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 η) 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. (8a) and (8b)) successively is equivalent to Φ-match. Notably, this ensures that the factorisation property of joint distributions in $\mathcal{Q}$ is preserved. We give the proof of this lemma in Appendix D.2.1.

Lemma 6. *Let $\phi^0 \in L^1(\nu)$ and let $\{\phi^n\}_{n \geq 1}$ be the sequence of potentials obtained as $\phi^{n+1} = \mathsf{M}^{\Phi\text{-match}}(\phi^n; \eta)$. Then, the sequence of distribution $\{\pi^n\}_{n \geq 0}$ where $\pi^n = \pi(\phi^n, (\phi^n)^+)$ satisfy for every $n \geq 0$*

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

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

$$\mathsf{root}_{\mathcal{X},\mu}(\pi; \mathcal{F}) := \operatorname*{argmin}_{\bar{\pi}} \left\{ \langle \mathcal{F}(\pi), \bar{\pi} - \pi \rangle + \eta^{-1} \cdot \mathrm{d}_{\mathrm{KL}}(\bar{\pi} \| \pi) : \bar{\pi}_{\mathcal{X}} = \mu \right\} . \tag{9}$$

Note that $\mathsf{root}_{\mathcal{X},\mu}(\pi; \mathcal{F})$ is defined for a general vector field $\mathcal{F}$. When $\mathcal{F}$ is the first variation of a functional that measures the discrepancy between $\pi_{\mathcal{Y}}$ and ν, $\mathsf{root}_{\mathcal{X},\mu}(\pi; \mathcal{F})$ can be viewed as minimising a local first-order approximation of this functional, thereby approximately matching the $\mathcal{Y}$-marginal, while maintaining the $\mathcal{X}$ marginal. The following lemma states that the successive projections defined in Eqs. (8a) and (8b) correspond to a root finding procedure for a specific $\mathcal{F}$ that depends on Φ and ν.

Lemma 7. *Let $\pi \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ be such that $\pi_{\mathcal{X}} = \mu$. Define the map $\mathcal{V}_\Phi : \pi \mapsto \log \Phi(\mathrm{d}\pi_{\mathcal{Y}}) - \log \Phi(\mathrm{d}\nu)$. Then,*

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

The proof of Lemma 7 are given in Appendix D.2.2. This leads to the following straightforward corollary about Φ-match in the manner of Lemma 6.

Corollary 1. *Let $\phi^0 \in L^1(\nu)$ and let $\{\phi^n\}_{n \geq 0}$ be the sequence of potentials obtained from Φ-match. Then, the sequence of distributions $\{\pi^n\}_{n \geq 0}$ where $\pi^n \equiv \pi(\phi^n, (\phi^n)^+)$ satisfies for every $n \geq 0$*

$$\pi^{n+1} = \mathsf{root}_{\mathcal{X},\mu}(\pi^n; \mathcal{V}_\Phi) .$$

This shows that while the updates are themselves expressed in terms of the potentials ϕ, the joint distributions are automatically are contained in the set $\mathcal{Q}$ and solve Eq. (9).

Recall from before that Φ-match with $\Phi : f \mapsto f$ and $\eta = 1$ coincides with Sinkhorn. In this setting, Lemma 6 recovers the classical iterative Bregman projection interpretation of the Sinkhorn method [PC19, Remark 4.8]. The equivalence in Corollary 1 for the Sinkhorn method was originally derived in [AFKL22, Prop. 5] and generalised to an arbitrary step size $\eta \in (0, 1)$ in [RKHK24, Lem. 1]. In this setting, the map $\mathcal{V}_\Phi$ corresponds to the first variation of the functional $\rho \mapsto \mathrm{d}_{\mathrm{KL}}(\rho_{\mathcal{Y}} \| \nu)$. We provide an interpretation of SGA similarly, for which we first introduce the notion of the *maximum mean discrepancy* (MMD) between distributions [GBR+06].

Let $k : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ be a positive-definite kernel, and let $\mathcal{H}_k$ be its RKHS. We refer the reader to [SC08, Chap. 4] for a more detailed exposition about kernels and their RKHS. The MMD between $\xi \in \mathcal{P}(\mathcal{Y})$ and ν is defined as

$$\mathrm{MMD}_k(\xi, \nu) = \| \mathfrak{m}_k(\xi) - \mathfrak{m}_k(\nu) \|_{\mathcal{H}_k} ; \qquad \mathfrak{m}_k(\xi)(y) := \int_{\mathcal{Y}} k(y, y') \cdot \mathrm{d}\xi(y') .$$

The first variation of $\xi \mapsto \mathcal{L}_k(\xi; \nu) := \frac{1}{2}\mathrm{MMD}_k(\xi, \nu)^2$ is given by $\mathfrak{m}_k(\xi) - \mathfrak{m}_k(\nu)$ [MSR19, Lem. 1]. Going back to Φ-match, recall that $\Phi(f) : f \mapsto e^f$ corresponds to SGA. The map $\mathcal{V}_\Phi$ in Lemma 7 in this case is given by $\mathcal{V}_\Phi(\rho) = \mathrm{d}\rho_{\mathcal{Y}} - \mathrm{d}\nu$, which coincides with the first variation of $\mathcal{L}_k(\rho_{\mathcal{Y}}, \nu)$ for the kernel k defined as $k(y, y') = 1$ iff $y = y'$. This is especially interesting as this kernel is *characteristic* [FBJ04], which means that $\mathrm{MMD}_k(\rho_{\mathcal{Y}}, \nu) = 0$ iff $\rho_{\mathcal{Y}} = \nu$, sharing similarities to the functional $\rho \mapsto \mathrm{d}_{\mathrm{KL}}(\rho_{\mathcal{Y}} \| \nu)$ that Sinkhorn is optimizing for.

One can hence more generally consider the update $\mathsf{M}^{k\text{-SGA}}$ for minimising $\mathcal{L}_k(\cdot; \nu)$ for a general characteristic kernel k. This is defined as

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

k-SGA is also 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, Corollary 1 shows that the sequence of iterates $\{\phi^n\}_{n \geq 1}$ formed by k-SGA result in a sequence of distributions $\{\pi(\phi^n, (\phi^n)^+)\}_{n \geq 1}$ formed by iteratively applying $\mathsf{root}_{\mathcal{X},\mu}$ with $\mathcal{F} \leftarrow \mathcal{V}_{\Phi_k}$ where $\mathcal{V}_\Phi$ is defined in Lemma 7. By the additivity of $\mathfrak{m}_k$, this update can also be viewed as a kernelised semi-dual gradient ascent.

4.1 Deriving non-asymptotic rates for k-SGA

With the new MMD_k objective, we derive non-asymptotic rates for k-SGA for a bounded, positive-definite kernels k here. We state a key lemma from [AFKL22] 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 3 ([AFKL22, 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 $\xi, \bar{\xi} \in \mathcal{P}(\mathcal{Y})$,*

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

Consequently,

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

We now state the formal guarantee for k-SGA; the proof is given in Appendix D.2.3.

Lemma 8. *Let $\phi^0 \in L^1(\nu)$, and let $\{\phi^n\}_{n \geq 1}$ be the sequence of potentials obtained as $\phi^{n+1} = \mathsf{M}^{k\text{-SGA}}(\phi^n; \frac{1}{2c_k})$ for a bounded, positive-definite kernel k. Then, the sequence of distributions $\{\pi^n\}_{n \geq 0}$ where $\pi^n = \pi(\phi^n, (\phi^n)^+)$ satisfies for any $N \geq 1$ that*

$$\mathrm{MMD}_k(\pi^N_{\mathcal{Y}}, \nu)^2 \leq \frac{4c_k}{N} \cdot \mathrm{d}_{\mathrm{KL}}(\pi^\star \| \pi^0) \, . \tag{10}$$

Note that every π^n in the sequence of distributions generated by k-SGA maintains the factorisation form of the optimal coupling in Eq. (3). When the kernel k is characteristic, Lemma 8 shows that $\pi^N_{\mathcal{Y}}$ approaches ν and consequently establishes a rate of convergence to the optimal coupling $\pi^\star$. As discussed previously, SGA corresponds to the special case of kernel given by $k(y, y') = 1$ iff $y = y'$, and the above lemma gives a $\frac{1}{N}$ rate of convergence *without* additional assumptions on the marginal ν or the cost function $c(\cdot, \cdot)$. This is fundamentally different to the rate shown in Lemmas 4 and 5, which is given in terms of the semi-dual J, and demonstrates the benefits of this alternate viewpoint.

5 Extensions

In this section, we build on the methods developed in the preceding sections and give two extensions. These are (1) an accelerated method for the eOT problem based on the semi-dual optimisation in Section 3, and (2) adaption of Φ-match for the dynamical Schrödinger bridge problem.

5.1 An accelerated method for the eOT problem

In light of the developments in the preceding sections where we give algorithms that achieve a $1/N$ rate of convergence for the eOT problem, a natural question to ask is if this rate can be improved. In this subsection, we give an accelerated version of proj-SGA that provably achieves a rate of $1/N^2$ for solving the eOT problem. This is particularly facilitated by the fact that the growth condition which proj-SGA is based on, is given in terms of $L^2(\nu)$ – a Hilbert space. This consequently enables us to adopt the structure of FISTA [BT09] with minor adjustments.

Lemma 9. *Let $\phi^1 = \overline{\phi}^0 \in \mathcal{S}_B$, and $t_1 = 1$. Consider the sequences $\{\phi^n\}_{n \geq 2}, \{\overline{\phi}^n\}_{n \geq 1}$ generated according to the recursion for $n \geq 1$*

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

Then, for any $N \geq 1$

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

Recollecting the discussion about the choice of B after Lemma 5: when the cost $c(\cdot, \cdot)$ is bounded and $B = \frac{3\|c\|_{L^\infty(\mathcal{X} \times \mathcal{Y})}}{2}$, the sequence $\{J(\overline{\phi}^N)\}_{N \geq 1}$ provably converges to $\mathrm{OT}_1(\mu, \nu; c/\varepsilon)$ at a rate that scales as $1/N^2$. We give the proof of Lemma 9 in Appendix D.3.1. Alternatively, one could possibly also design accelerated algorithms for minimising $\rho \mapsto \mathcal{L}_k(\cdot; \nu)$ over $\rho \in \mathcal{Q}$; however, this might prove to be challenging owing to the non-convexity of the set $\mathcal{Q}$.

5.2 Path-space Φ-match for the Schrödinger Bridge problem

The eOT problem in the static Schrödinger bridge form (Eq. (2)) can be generalised to path measures. Let $\mathbb{P}^{\mathsf{ref}}$ be a reference path measure defined by solutions of the SDE

$$\mathrm{d}X_t = u^{\mathsf{ref}}(X_t)\mathrm{d}t + \sqrt{2}\cdot \mathrm{d}B_t, \quad \mathbb{P}_0^{\mathsf{ref}} = \mu\,. \tag{11}$$

Above, $(B_t)_{t\geq 0}$ is the Brownian motion. The (dynamical) Schrödinger bridge problem is given as the following minimisation problem over all stochastic processes over $[0,T]$

$$\min_{\mathbb{P}}\ \mathrm{d}_{\mathrm{KL}}(\mathbb{P}\|\mathbb{P}^{\mathsf{ref}}) \quad \text{such that } \mathbb{P}_0 = \mu, \mathbb{P}_T = \nu\,. \tag{12}$$

The disintegration property of d_{KL} implies that the optimal path measure can be de-coupled into $\mathbb{P}^*(X_{t\in[0,T]}) = \pi^*(X_0, X_T)\mathbb{P}^{\mathsf{ref}}(X_{t\in(0,T)}|X_0, X_T)$ [Föl88] where $\pi^*(X_0, X_T)$ is the optimal coupling from Eq. (2) when $R = \mathbb{P}^{\mathsf{ref}}_{0,T}$. One can view Eq. (12) as a dynamic formulation of the eOT problem over the trajectory space $\mathcal{C}([0,T],\mathbb{R}^d)$ that generalises the marginal coupling case over $\mathcal{P}(\mathbb{R}^d\times\mathbb{R}^d)$ at the two terminals (c.f. Eq. (2)). We provide more background on the quantities and definitions introduced in this subsection in Appendix A.2 for the convenience of the reader.

The path-space version of the Sinkhorn algorithm is the *Iterative Proportional Fitting* (IPF) algorithm. It generates a sequence of iterates $\{\mathbb{P}^n\}_{n\geq 1}$ initialised with $\mathbb{P}^0 = \mathbb{P}^{\mathsf{ref}}$ that satisfy

$$\mathbb{P}^{n+1/2} = \operatorname*{argmin}_{\overline{\mathbb{P}}:\,\overline{\mathbb{P}}_T=\nu} \mathrm{d}_{\mathrm{KL}}(\overline{\mathbb{P}}\parallel\mathbb{P}^n)\,, \quad \mathbb{P}^{n+1} = \operatorname*{argmin}_{\overline{\mathbb{P}}:\,\overline{\mathbb{P}}_0=\mu} \mathrm{d}_{\mathrm{KL}}(\overline{\mathbb{P}}\parallel\mathbb{P}^{n+1/2})\,. \tag{13}$$

The Φ-match analogue of this alternating projection scheme is given by

$$\mathbb{P}^{n+1/2} = \operatorname*{argmin}_{\overline{\mathbb{P}}}\left\{\mathrm{d}_{\mathrm{KL}}(\overline{\mathbb{P}}\|\mathbb{P}^n)\ :\ \mathrm{d}\overline{\mathbb{P}}_T \propto \mathrm{d}\mathbb{P}^n_T\cdot\frac{\Phi(\mathrm{d}\nu)}{\Phi(\mathrm{d}\mathbb{P}^n_T)}\right\} \tag{14a}$$

$$\mathbb{P}^{n+1} = \operatorname*{argmin}_{\overline{\mathbb{P}}}\left\{\eta\cdot\mathrm{d}_{\mathrm{KL}}(\overline{\mathbb{P}}\|\mathbb{P}^{n+1/2}) + (1-\eta)\cdot\mathrm{d}_{\mathrm{KL}}(\overline{\mathbb{P}}\|\mathbb{P}^n)\ :\ \overline{\mathbb{P}}_0 = \mu\right\}\,. \tag{14b}$$

Eq. (14a)-Eq. (14b) can be implemented as updates on the drifts of a sequences of SDEs whose corresponding solutions are given by $\{\mathbb{P}^n\}_{n\geq 1}$. This relies on time-reversals and stochastic optimal control similar to those in [RKHK24, Sec. 4.3]. We also obtain a relation between the iterates generated by Φ-match and a sequence of path measures that converge to their respective optimals $\phi^*, \mathbb{P}^*$ at the same rate; this is stated as the following informal lemma.

Lemma 10 (Informal)**.** *Let $\{\phi^n\}_{n\geq 1}$ be the sequence of potentials generated from Φ-match. For any $n\geq 0$, define the solution $(g_t)_{t\in[0,T]}$ that solves for each $y\in\mathcal{Y}$*

$$\partial_t g^n_t(y) + u^{\mathsf{ref}}(y)^\top\nabla g^n_t(y) + \Delta g^n_t(y) = 0, \quad g^n_T(y) = \exp(\phi^n(y))\,.$$

Let $\mathbb{P}^n$ be the path measure defined by solutions of the SDE

$$\mathrm{d}X^n_t = u^{\mathsf{ref}}(X^n_t)\mathrm{d}t + \nabla\phi^n_t(X^n_t)\mathrm{d}t + \sqrt{2}\cdot\mathrm{d}B_t; \quad X^n_0\sim\mu$$

where $\phi^n_t = \log g^n_t$. Then, the sequence $\{\mathbb{P}^n\}_{n\geq 1}$ converges to $\mathbb{P}^\star$ at the same rate as $\{\phi^n\}_{n\geq 1}$ converges to $\phi^\star$ for the eOT problem.

We give a more detailed version of this statement and the preceding claims in Appendix C. The key takeaway of this subsection is that the Φ-match abstraction finds a natural analogue in the path space and can be seamlessly generalized to the dynamical Schrödinger bridge problem.

6 Conclusion

In this work, we systemically synthesise a variety of viewpoints on algorithms for eOT – specifically those involving the Sinkhorn algorithm. This synthesis, centered around infinite-dimensional optimization, leads to a collection of novel methods based on either the primal or dual formulation of the eOT problem, allowing us to go beyond the classical Sinkhorn algorithm. 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; in particular, going past the bounded cost condition that we assume for proj-SGA and its accelerated version. While our work is theoretical in nature, we hope that the guarantees provided will encourage implementations of the methods proposed here.

References

[AFKL22] Pierre-Cyril Aubin-Frankowski, Anna Korba, and Flavien Léger. Mirror descent with relative smoothness in measure spaces, with application to Sinkhorn and EM. *Advances in Neural Information Processing Systems*, 35:17263–17275, 2022.

[AGS08] Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient flows: in metric spaces and in the space of probability measures*. Springer Science & Business Media, 2008.

[ANWR17] Jason Altschuler, Jonathan Niles-Weed, and Philippe Rigollet. Near-linear time approximation algorithms for optimal transport via Sinkhorn iteration. *Advances in neural information processing systems*, 30, 2017.

[BT09] 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.

[CCGT24] 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.

[CDV24] Lénaïc Chizat, Alex Delalande, and Tomas Vaškevičius. Sharper Exponential Convergence Rates for Sinkhorn's Algorithm in Continuous Settings. *arXiv preprint arXiv:2407.01202*, 2024.

[CH21] Kenneth F Caluya and Abhishek Halder. Wasserstein proximal algorithms for the Schrödinger bridge problem: Density control with nonlinear drift. *IEEE Transactions on Automatic Control*, 67(3):1163–1178, 2021.

[CLP23] Giovanni Conforti, Daniel Lacker, and Soumik Pal. Projected Langevin dynamics and a gradient flow for entropic optimal transport. *arXiv preprint arXiv:2309.08598*, 2023.

[CP18] Marco Cuturi and Gabriel Peyré. Semidual regularized optimal transport. *SIAM Rev.*, 60(4):941–965, 2018.

[CRL+20] Lenaic Chizat, Pierre Roussillon, Flavien Léger, François-Xavier Vialard, and Gabriel Peyré. Faster wasserstein distance estimation with the sinkhorn divergence. *Advances in Neural Information Processing Systems*, 33:2257–2269, 2020.

[Cut13] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26, 2013.

[CWVB09] Rick Chartrand, Brendt Wohlberg, Kevin Vixie, and Erik Bollt. A gradient descent solution to the monge-kantorovich problem. *Applied Mathematical Sciences*, 3(22):1071–1080, 2009.

[DGK18] Pavel Dvurechensky, Alexander Gasnikov, and Alexey Kroshnin. Computational optimal transport: Complexity by accelerated gradient descent is better than by Sinkhorn's algorithm. In *International Conference on Machine Learning*, pages 1367–1376. PMLR, 2018.

[DKPS23] 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.

[DP91] Paolo Dai Pra. A stochastic control approach to reciprocal diffusion processes. *Applied mathematics and Optimization*, 23(1):313–329, 1991.

[FBJ04] Kenji Fukumizu, Francis R Bach, and Michael I Jordan. Dimensionality reduction for supervised learning with reproducing kernel hilbert spaces. *Journal of Machine Learning Research*, 5(Jan):73–99, 2004.

[FL89] Joel Franklin and Jens Lorenz. On the scaling of multidimensional matrices. *Linear Algebra and its applications*, 114:717–735, 1989.

[Föl88] Hans Föllmer. Random fields and diffusion processes. In *École d'Été de Probabilités de Saint-Flour XV–XVII, 1985–87*, volume 1362 of *Lecture Notes in Math.*, pages 101–203. Springer, Berlin, 1988.

[FSV+19] Jean Feydy, Thibault Séjourné, François-Xavier Vialard, Shun-ichi Amari, Alain Trouvé, and Gabriel Peyré. Interpolating between optimal transport and mmd using sinkhorn divergences. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2681–2690. PMLR, 2019.

[GBR+06] Arthur Gretton, Karsten Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alex Smola. A kernel method for the two-sample-problem. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems*, volume 19. MIT Press, 2006.

[GCB+19] Aude Genevay, Lénaic Chizat, Francis Bach, Marco Cuturi, and Gabriel Peyré. Sample complexity of sinkhorn divergences. In *The 22nd international conference on artificial intelligence and statistics*, pages 1574–1583. PMLR, 2019.

[GCPB16] Aude Genevay, Marco Cuturi, Gabriel Peyré, and Francis Bach. Stochastic optimization for large-scale optimal transport. *Advances in neural information processing systems*, 29, 2016.

[GM96] Wilfrid Gangbo and Robert J. McCann. The geometry of optimal transportation. *Acta Math.*, 177(2):113–161, 1996.

[JL20] Matt Jacobs and Flavien Léger. A fast approach to optimal transport: The back-and-forth method. *Numerische Mathematik*, 146(3):513–544, 2020.

[KBB15] Walid Krichene, Alexandre Bayen, and Peter L Bartlett. Accelerated mirror descent in continuous and discrete time. *Advances in neural information processing systems*, 28, 2015.

[Lég21] Flavien Léger. A gradient descent perspective on Sinkhorn. *Applied Mathematics & Optimization*, 84(2):1843–1855, 2021.

[LHJ22] 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.

[Lé14] Christian Léonard. A survey of the schrödinger problem and some of its connections with optimal transport. *Discrete and Continuous Dynamical Systems*, 34(4):1533–1574, 2014.

[MG20] 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):27, 2020.

[Mis19] Konstantin Mishchenko. Sinkhorn algorithm as a special case of stochastic mirror descent. *arXiv preprint arXiv:1909.06918*, 2019.

[MNW19] Gonzalo Mena and Jonathan Niles-Weed. Statistical bounds for entropic optimal transport: sample complexity and the central limit theorem. *Advances in neural information processing systems*, 32, 2019.

[MP20] Arthur Mensch and Gabriel Peyré. Online sinkhorn: Optimal transport distances from sample streams. *Advances in Neural Information Processing Systems*, 33:1657–1667, 2020.

[MSR19] Youssef Mroueh, Tom Sercu, and Anant Raj. Sobolev descent. In Kamalika Chaudhuri and Masashi Sugiyama, editors, *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 2976–2985, 2019.

[Nes18] Yurii Nesterov. *Lectures on Convex Optimization*, volume 137 of *Springer Optimization and Its Applications*. Springer, Cham, second edition, 2018.

[Nut21] Marcel Nutz. Introduction to entropic optimal transport. *Lecture notes, Columbia University*, 2021.

[NW22] Marcel Nutz and Johannes Wiesel. Entropic optimal transport: Convergence of potentials. *Probability Theory and Related Fields*, 184(1):401–424, 2022.

[Pav14] Grigorios A. Pavliotis. *Stochastic processes and applications*, volume 60 of *Texts in Applied Mathematics*. Springer, New York, 2014. Diffusion processes, the Fokker-Planck and Langevin equations.

[PC19] Gabriel Peyré and Marco Cuturi. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11:355–607, 2019.

[RKHK24] 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.

[Rud87] Walter Rudin. *Real and complex analysis, 3rd ed.* McGraw-Hill, Inc., USA, 1987.

[San15] Filippo Santambrogio. Optimal transport for applied mathematicians. *Birkäuser, NY*, 55(58-63):94, 2015.

[SC08] Ingo Steinwart and Andreas Christmann. *Support vector machines*. Information Science and Statistics. Springer, New York, 2008.

[SK67] Richard Sinkhorn and Paul Knopp. Concerning nonnegative matrices and doubly stochastic matrices. *Pacific Journal of Mathematics*, 21(2):343–348, 1967.

[Vil03] Cédric Villani. *Topics in optimal transportation*, volume 58 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2003.

Organisation: We give some additional background that is pertinent to this work in Appendix A. This includes a brief discussion of the standard optimal transport problem for completeness in Appendix A.1 and more details about the (dynamical) Schrödinger bridge problem in Appendix A.2. We give an additional mirror-descent interpretation of Φ-match which we excluded from the main text due to space constraints in Appendix B. We also provide an extended discussion continuing on Section 5.2 in Appendix C. Appendix D is dedicated solely to the proofs of the lemmas and propositions that appear in the main text, as well as intermediate results that are used in proving these assertions. At the end, we give a brief discussion about implementing k-SGA in Appendix E.

A Additional Background

A.1 The optimal transport problem

The Kantorovich formulation of the optimal transport (abbrev. OT) problem is given by the following program

$$\inf_{\pi \in \Pi(\mu,\nu)} \iint c(x,y) \mathrm{d}\pi(x,y) =: \mathrm{OT}(\mu,\nu;c) \,. \tag{15}$$

An important case is when $c(x,y) = \|x-y\|^p$, the value $\mathrm{OT}(\mu,\nu;c)$ is $W_p(\mu,\nu)^p$ where W_p is the p-Wasserstein distance between μ and ν. The OT problem admits a dual formulation with zero duality gap called the Kantorovich dual [Vil03, Thm. 1.3] and is given by

$$\mathrm{OT}(\mu,\nu;c) = \sup \left\{ \int f(x) d\mu(x) + \int g(y) \mathrm{d}\nu(y) : f(x) + g(y) \leq c(x,y) \right\} \,. \tag{16}$$

The semi-dual problem has its analogue for the OT problem as well. From [GM96], any solution $(f^\star, g^\star)$ that solves Eq. (16) is a pair of functions that satisfies

$$f^\star = (g^\star)^c \qquad g^\star = (f^\star)^c \,.$$

A function ϕ is said to be c-concave if it is the c-transform of a continuous function, and where the c-transform of a continuous function φ is defined as

$$\varphi^c(x) = \inf_y \; c(x,y) - \varphi(y) \,.$$

Consequently, the dual problem in Eq. (16) can be reduced to a problem over just one of $f \in L^1(\mu)$ or $g \in L^1(\nu)$ (c.f. [JL20, Thm. 1]) as

$$\mathrm{OT}(\mu,\nu;c) = \sup \int f(x) \mathrm{d}\mu(x) + \int f^c(y) \mathrm{d}\nu(y) \,. \tag{17}$$

Analogous to the methods we propose in Section 3 for maximising the semi-dual in eOT, iterative methods for maximising the semi-dual have also been considered for the standard optimal transport problem [CWVB09, JL20]. [CWVB09] consider the L^2-gradient (equivalently, the Fréchet derivative) of the OT semi-dual, while recent work by [JL20] use the $\dot{H}^1$-gradient defined with respect to the homogeneous Sobolev space $\dot{H}^1$.

A.2 Dynamical Schrödinger bridge problem

Recall from Eq. (12) that the dynamical Schrödinger problem over the path space is given by

$$\min_{\mathbb{P}} \; \mathrm{d}_{\mathrm{KL}}(\mathbb{P} \| \mathbb{P}^{\mathsf{ref}}) \quad \text{such that } \mathbb{P}_0 = \mu, \mathbb{P}_T = \nu \,.$$

This is a (strictly) convex problem defined on the space of probability distributions over curves $\mathcal{C}([0,T],\mathbb{R}^d)$ i.e., $\mathcal{P}(\mathcal{C}([0,T],\mathbb{R}^d))$. Above $\mathbb{P}^{\mathsf{ref}}$ is a path measure defined by the solutions of the following SDE

$$\mathrm{d}X_t = u^{\mathsf{ref}}(X_t)\mathrm{d}t + \sqrt{2} \cdot \mathrm{d}B_t, \qquad X_0 \sim \mu \,.$$

By Girsanov's theorem [Pav14, Chap. 3], the dynamical SB problem can be reduced to a minimisation problem over path measures $\mathbb{P}$, each of which is induced by an SDE with an additional drift $(v_t)_{t \geq 0}$

$$\mathrm{d}X_t = u^{\mathsf{ref}}(X_t)\mathrm{d}t + v_t(X_t)\mathrm{d}t + \sqrt{2} \cdot \mathrm{d}B_t \,, \qquad X_0 \sim \mu \,. \tag{18}$$

In this work, we assume that the reference path measure $\mathbb{P}^{\text{ref}}$ is Markovian i.e., the solution $\{X_t\}_{t\in[0,T]}$ that defines $\mathbb{P}^{\text{ref}}$ is Markov process.

As noted in Section 5.2, the disintegration property of $\mathrm{d_{KL}}$ implies that the optimal path measure can be decoupled as

$$\mathbb{P}^*(X_{t\in[0,T]}) = \pi^*(X_0, X_T)\mathbb{P}^{\text{ref}}(X_{t\in(0,T)}|X_0, X_T)\ .$$

where $\pi^*(X_0, X_T)$ is the optimal coupling from Eq. (2) when $R = \mathbb{P}^{\text{ref}}_{0,T}$. As a result, letting $\mathbb{P}^{\text{ref}}$ be the law of a standard reversible Brownian motion gives rise to an entropy-regularised OT problem with quadratic cost $c(x,y) = \frac{\|x-y\|^2}{2T}$ in Eq. (1) since $\mathrm{d}\mathbb{P}^{\text{ref}}_{0,T}(X_0, X_T) \propto \exp\left(-\frac{\|X_0-X_T\|^2}{2T}\right)$. This implies that to recover the optimal path measure, one can solve a static OT involving 2 marginals for $c(X_0, X_T) = \frac{1}{2}\|X_0 - X_T\|^2$ as the cost and $\varepsilon = T$ as the entropic regularisation parameter, followed by sampling from the conditional Brownian bridge process whose endpoints are given by X_0, X_T.

Under certain regularity conditions, there is a factorisation analogous to Eq. (3) that holds in the trajectory space, which is given by

$$\frac{\mathbb{P}^*}{\mathbb{P}^{\text{ref}}}(X_t)_{0\le t\le T} = \frac{\mathbb{P}^*_{0,T}}{\mathbb{P}^{\text{ref}}_{0,T}}(X_0, X_T) = \exp(\boldsymbol{\phi}_T(X_T) + \boldsymbol{\psi}_0(X_0)), \quad \mathbb{P}^*\text{-}a.e. \tag{19}$$

for bounded continuous functions $\boldsymbol{\psi}_0, \boldsymbol{\phi}_T$ over $\mathbb{R}^d$ that are only functions of the initial and the final variable under certain regularity conditions; we refer the reader to [Lé14, Thms 2.8 & 2.9] for these. Eq. (19) is necessary and almost sufficient for characterizing $\mathbb{P}^*$.

Analogous to the dual formulation of the eOT problem in Eq. (4), there exists a dual formulation for Eq. (12). This involves maximising over the space of all bounded continuous functions on $\mathbb{R}^d$ subject to a constraint given by a Hamilton-Jacobi-Bellman (HJB) equation.

$$\sup_{\{\boldsymbol{\phi}_t\}_{t\in[0,T]}} \int \boldsymbol{\phi}_T(y)\mathrm{d}\nu(y) - \int \boldsymbol{\phi}_0(x)\mathrm{d}\mu(x) \quad \text{s.t.} \quad \frac{1}{2}\|\nabla\boldsymbol{\phi}_t\|^2 + \Delta\boldsymbol{\phi}_t + \partial_t\boldsymbol{\phi}_t + \nabla\boldsymbol{\phi}_t^\top u^{\text{ref}} = 0\ . \tag{20}$$

The HJB constraint can be folded into the dual using the semigroup associated with the solution of the HJB equation. This results in the problem

$$\sup_{\phi:\mathbb{R}^d\mapsto\mathbb{R}} \int \phi(y)\mathrm{d}\nu(y) - \int H_T[\phi](x)\mathrm{d}\mu(x)\ ; \qquad H_t[\phi] := \log(h_t e^{\phi})\ .$$

Above, the operator H_t is a semigroup providing us with solutions of the backward Hamilton-Jacobi-Bellman equation, and $h_t f$ is the flow corresponding to the PDE

$$\partial_t f = -\nabla f^\top u^{\text{ref}} - \Delta f$$

starting at f and evaluated backward at time t. In particular, if $u^{\text{ref}} = 0$, it reduces to a backward heat equation.

B Another interpretation of Φ-match: Mirror Descent

The interpretation of Φ-match presented here is a generalisation of the mirror method perspective of η-Sinkhorn introduced in [RKHK24]. Define the reference measure $\pi^{\text{ref}}_\varepsilon$ whose density is

$$\mathrm{d}\pi^{\text{ref}}_\varepsilon(x,y) = \frac{1}{Z}\cdot\exp\left(-\frac{c(x,y)}{\varepsilon}\right)\mathrm{d}\mu(x)\mathrm{d}\nu(y)\ ; \quad Z = \iint_{\mathcal{X}\times\mathcal{Y}} \exp\left(-\frac{c(x,y)}{\varepsilon}\right)\mathrm{d}\mu(x)\mathrm{d}\nu(y)\ .$$

Let $\varphi : \mathcal{P}(\mathcal{X}\times\mathcal{Y}) \to \mathbb{R}$ be given by $\varphi(\pi) = \mathrm{d_{KL}}(\pi\|\pi^{\text{ref}}_\varepsilon)$. Then,

$$\delta\varphi(\pi)(x,y) = \log\frac{\mathrm{d}\pi(x,y)}{\mathrm{d}\pi^{\text{ref}}_\varepsilon(x,y)}\ .$$

In [RKHK24, Lemma 3] it is noted that for $\varphi^\star$ given by the convex conjugate of φ over the domain $\{\pi : \pi_{\mathcal{X}} = \mu\}$,

$$\delta\varphi^\star(h)(x,y) = \mathrm{d}\mu(x)\cdot\frac{\mathrm{d}\pi^{\text{ref}}_\varepsilon(x,y)\exp(h(x,y))}{\int_{\mathcal{Y}}\mathrm{d}\pi^{\text{ref}}_\varepsilon(x,y')\exp(h(x,y'))}\ . \tag{21}$$

This implies that $\delta\varphi^\star$ forms a map from $h \in L^1(\mathcal{X}\times\mathcal{Y}) \mapsto \{\pi : \pi_\mathcal{X} = \mu\}$. To see that it maps onto the set $\{\pi : \pi_\mathcal{X} = \mu\}$, we can rewrite the RHS as $\mu(x)\cdot\tilde{\pi}(y|x)$ for some probability distribution $\tilde{\pi}(x,y) \ll \pi_\epsilon^{\text{ref}}(x,y)$. This mapping gives us a mirror map between the dual potential and the primal measure that manifests its importance in the following lemma.

Lemma 11. *Let $\{\phi^n\}_{n\geq 0}$ be the sequence of potentials obtained from Φ-match. Then the sequence of distributions $\{\pi^n\}_{n\geq 0}$ where $\pi^n = \pi(\phi^n, (\phi^n)^+)$ satisfies the recursion*

$$\mathrm{d}\pi^{n+1} = \delta\varphi^\star\left(\delta\varphi(\pi^n) - \eta\cdot\mathcal{V}_\Phi(\pi^n)\right)\ ; \qquad \mathcal{V}_\Phi(\pi) = \log\Phi(\mathrm{d}\pi_\mathcal{Y}) - \log\Phi(\mathrm{d}\nu)\ . \tag{22}$$

We give the proof of Lemma 11 in Appendix D.4.1. Lemma 11 offers a mirror descent interpretation for SGA (obtained when $\Phi : f \mapsto e^f$), in the sense that we start with the current primal iterate, map it to the dual space, perform a gradient step there and map back to get the updated primal iterate. The judicious choice of φ makes sure that the primal iterate always stays in the set $\mathcal{Q}$, and the gradient step gradually match the $\mathcal{Y}$-marginal.

With this particular choice of mirror map $\delta\varphi$, one can show that $D_\varphi(\pi\|\pi') = \mathrm{d}_{\mathrm{KL}}(\pi\|\pi')$ when both $\pi_\mathcal{X}, \pi'_\mathcal{X} = \mu$ (c.f. Lemma B.1 in [RKHK24]). Therefore φ, as a Bregman potential, induces a valid distance between measures with constrained marginal.

B.1 Accelerating SGA through a mirror flow

Recall that the constraint set we are interested in is

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

Following the discussion in Section 2, $\mathcal{Q}\subset\{\pi : \pi_\mathcal{X} = \mu\}$. Throughout this subsection we will adopt the following notation. The probability density w.r.t. the Lebesgue measure of a measure ρ will be identified by ρ as well. We also use f, g in lieu of $-\psi, \phi$ for the $\mathcal{X}, \mathcal{Y}$ potentials. For convenience, we use the shorthand $f\oplus g$ to denote $(f\oplus g)(x,y) = f(x) + g(y)$ where $f : \mathcal{X}\to\mathbb{R}$ and $g : \mathcal{Y}\to\mathbb{R}$. By direct calculation, we see that $\delta\varphi^\star(f\oplus g)\in\mathcal{Q}$:

$$\begin{aligned}\delta\varphi^\star(f\oplus g)(x,y) &= \mathrm{d}\mu(x)\cdot\exp\left(g(y) - \frac{c(x,y)}{\varepsilon}\right)\mathrm{d}\nu(y)\cdot\left(\int_\mathcal{Y}\exp\left(g(y) - \frac{c(x,y)}{\varepsilon}\right)\mathrm{d}\nu(y)\right)^{-1}\\ &= \exp\left(g(y) - g^+(x) - \frac{c(x,y)}{\varepsilon}\right)\mathrm{d}\mu(x)\mathrm{d}\nu(y)\ .\end{aligned} \tag{23}$$

If h^0 is of the form $h^0 = f^0\oplus g^0$, then the continuous-time analogue of the update in Eq. (22) is given by

$$\dot{h}^t = \mathbf{0}\oplus(-(\hat{\pi}^t_\mathcal{Y} - \nu)), \quad \hat{\pi}^t = \delta\varphi^\star(h^t)\in\mathcal{Q}\,.$$

This is also equivalent to the ODE

$$\dot{g}^t = -(\hat{\pi}^t_\mathcal{Y} - \nu)\ , \quad \hat{\pi}^t = \delta\varphi^\star(h^t)\in\mathcal{Q} \tag{24}$$

with $\dot{f}^t$ fixed by the dynamics on $\dot{g}^t$ by the $\hat{\pi}^t_\mathcal{X} = \mu$ constraint. The formal equivalence is established in [RKHK24, Theorem 3.1], which also results in the rate

$$\mathrm{d}_{\mathrm{KL}}(\hat{\pi}^t_y\|\nu) \leq \frac{\mathrm{d}_{\mathrm{KL}}(\pi(g^0,(g^0)^+),\pi^\star)}{t}\ . \tag{25}$$

For building momentum into the dynamics, we build on [KBB15], together with our interpretation of Φ-match in Lemma 11. Let $r\geq 2$ be a constant. Define the ODE system below

$$\dot{\hat{\pi}}^t = \frac{r}{t}(\delta\varphi^\star(h^t) - \hat{\pi}^t)\ , \tag{26a}$$

$$\dot{h}^t = \mathbf{0}\oplus\left(-\frac{t}{r}\cdot(\hat{\pi}^t_\mathcal{Y} - \nu)\right)\ . \tag{26b}$$

The initial conditions to this system is $\hat{\pi}^0 = \delta\varphi^*(h^0(x,y))$ where $h^0 = f^0\oplus g^0$ and hence

$$\hat{\pi}^0(x,y) = \exp\left(-\frac{c(x,y)}{\varepsilon}\right)\mu(x)\nu(y)\exp((g^0)^+(x) + g^0(y))\in\mathcal{Q}\,. \tag{27}$$

We reduce the pair of ODEs to a single ODE. This is done by rewriting Eq. (26a) as

$$t^r \dot{\hat{\pi}}^t + r t^{r-1} \hat{\pi}^t = r t^{r-1} \delta\varphi^*(h^t)$$

and upon time integration,

$$t^r \hat{\pi}^t = r \int_0^t \tau^{r-1} \delta\varphi^*(h^\tau) \mathrm{d}\tau \;\Leftrightarrow\; \hat{\pi}^t = \frac{\int_0^t \tau^{r-1} \delta\varphi^*(h^\tau) \mathrm{d}\tau}{\int_0^t \tau^{r-1} \mathrm{d}\tau} \,. \tag{28}$$

Eq. (28) implies that the solution $\hat{\pi}^t$ at each t is weighted "sum" of $(\delta\varphi^\star(h^\tau))_{\tau\in(0,t)}$ that is based on Eq. (26b). With the initial condition as in $h^0 = f^0 \oplus g^0$, the second update Eq. (26b) says the dual "variable" g^t accumulates gradient at a certain rate. More precisely, Eq. (26b) can be equivalently written as

$$\dot{g}^t = -\frac{t}{r} \cdot (\hat{\pi}^t_{\mathcal{Y}} - \nu) \qquad h^t = f^0 \oplus g^t \,. \tag{29}$$

Hence, to implement this, we only require access to $\hat{\pi}^t_{\mathcal{Y}}$, which can be computed from Eq. (28) as

$$\begin{aligned} \hat{\pi}^t_{\mathcal{Y}} &= \int_{\mathcal{X}} \frac{\int_0^t \tau^{r-1} \delta\varphi^\star(h^\tau) \mathrm{d}\tau}{\int_0^t \tau^{r-1} \mathrm{d}\tau} \mathrm{d}x \\ &= \left(\int_0^t \tau^{r-1} \mathrm{d}\tau \right)^{-1} \cdot \int_0^t \tau^{r-1} \left\{ \int_{\mathcal{X}} \exp\left(g^\tau(y) - (g^\tau)^+(x) - \frac{c(x,y)}{\varepsilon} \right) \mathrm{d}\mu(x) \mathrm{d}\nu(y) \right\} \mathrm{d}\tau \end{aligned} \tag{30}$$

where we used $\delta\varphi^\star(h^\tau) \in \mathcal{Q}$ from Eq. (23). Note that Eq. (30) is only a function of g. Consequently, the original accelerated mirror descent dynamics in Eqs. (26a) and (26b) are reduced to Eq. (29)-Eq. (30) in terms of $g^t, \hat{\pi}^t_{\mathcal{Y}}$ only. However, the non-convexity of $\mathcal{Q}$ doesn't guarantee that $\hat{\pi}^t \in \mathcal{Q}$ despite being a convex combination of $\rho^\tau := \delta\varphi^\star(h^\tau)$ which individually lie in $\mathcal{Q}$. This makes it somewhat difficult to argue if $\hat{\pi}^t$ is approaching an optimal coupling.

However, this doesn't preclude us from giving a rate of convergence for the continuous-time dynamics Eqs. (26a) and (26b) for the $\mathcal{Y}$-marginal.

Lemma 12. *Let $(\hat{\pi}^t)_{t\geq 0}$ be the solution to the system Eqs. (26a) and (26b). For kernel $k(y, y') = 1$ iff $y = y'$, it holds for any $r \geq 2$ that*

$$\mathcal{L}_k(\hat{\pi}^t_{\mathcal{Y}}, \nu) \leq \frac{r^2}{t^2} \cdot \mathrm{d}_{\mathrm{KL}}(\pi(g^0, (g^0)^+), \pi^\star) \,,$$

where $\mathcal{L}_k(\cdot; \nu) = \frac{1}{2}\mathrm{MMD}_k(\cdot, \nu)^2$.

We give the proof of Lemma 12 in Appendix D.4.2. Although the rate above is $\mathcal{O}(1/t^2)$, faster than MD ODE with $\mathcal{O}(1/t)$ rate in Eq. (25), the lemma does not establish convergence of $\hat{\pi}^t \to \pi^*$, only the convergence for the $\mathcal{X}$ and the $\mathcal{Y}$ marginals. However, from Eq. (30) we have that $\hat{\pi}^t$ is a weighted sum of $\rho^\tau \in \mathcal{Q}$ and since the weights increase very fast with τ after running long enough we'd expect $\hat{\pi}^t \approx \rho^t \in \mathcal{Q}$ for large t. What we can conclude from this discussion is that despite the niceness of $\mathcal{L}_k$ from a mirror descent standpoint (e.g., smooth and convex), the non-convexity of the constraint set we are optimizing over poses a fundamental challenge for acceleration, which involves interpolation between past iterates. This also highlights the benefit of the semi-dual perspective for building momentum, as we do in Section 5.1.

C Details on path-space Φ-match from Section 5.2

The IPF updates (Eq. (13)) can be viewed as a root finding procedure like Eq. (9), but now over the space of path measures. More precisely, in [RKHK24] it was shown that $\mathbb{P}^{n+1}$ obtained from Eq. (13) also solves (the inner product here is given by $L^2([0,T] \times \mathbb{R}^d)$)

$$\mathbb{P}^{n+1} = \operatorname*{argmin}_{\overline{\mathbb{P}}} \; \left\{ \langle \delta \mathrm{d}_{\mathrm{KL}}(\mathbb{P}^n_T \| \nu), \overline{\mathbb{P}} - \mathbb{P}^n \rangle + \eta^{-1} \cdot \mathrm{d}_{\mathrm{KL}}(\overline{\mathbb{P}} \| \mathbb{P}^n) : \overline{\mathbb{P}}_0 = \mu \right\} \,, \tag{31}$$

where $\delta \mathrm{d}_{\mathrm{KL}}(\mathbb{P}^n_T \| \nu) = \log \frac{d\mathbb{P}^n_T}{d\nu}$ with stepsize $\eta = 1$. The IPF updates can be carried out via a sequence of time-reversals over the path space by initialising alternatively the end-points of $\overline{\mathbb{P}}$ as $\overline{\mathbb{P}}_T = \nu$ and $\overline{\mathbb{P}}_0 = \mu$. As a specific example, the solution to the first half step of Eq. (13) is given by

$$\mathbb{P}^{n+1/2}(X_t)_{0\leq t\leq T} = \nu(X_T) \mathbb{P}^n((X_t)_{0\leq t<T} | X_T) \,.$$

Written as SDEs, this corresponds to (assuming $\mathbb{P}^n$ has time marginal p_t^n and drift v_t^n)

$$\mathrm{d}Y_t = -v_{T-t}^n(Y_t)\mathrm{d}t + 2\nabla \log p_{T-t}^n(Y_t)\mathrm{d}t + \sqrt{2}\cdot \mathrm{d}B_t\,, \quad Y_0 \sim \nu.$$

We build on the optimisation interpretation Eq. (31) and generalize our newly proposed Φ-match algorithm to the path space below. We propose 2 different implementations, both resulting in a recursive update on the drift $\{v_t^n(\cdot)\}_{n\geq 1}$ in Eq. (18), as a way to represent the (equivalent) path measures updates $\{\mathbb{P}^n\}_{n\geq 1}$.

Proofs for this section can be found in Appendix D.3.2 and Appendix D.3.3.

C.1 Pathspace Φ-match: Primal

We base the first one on the alternating projection interpretation of Φ-match. Very similar to the proof in Lemma 7, one can show that for a generic Φ and stepsize $\eta > 0$,

$$\mathbb{P}^{n+1} = \operatorname*{argmin}_{\overline{\mathbb{P}}} \left\{ \langle \log \Phi(\mathrm{d}\mathbb{P}_T^n) - \log \Phi(\mathrm{d}\nu), \overline{\mathbb{P}} - \mathbb{P}^n \rangle + \eta^{-1} \cdot \mathrm{d}_{\mathrm{KL}}(\overline{\mathbb{P}} \| \mathbb{P}^n) : \overline{\mathbb{P}}_0 = \mu \right\} \tag{32}$$

coincides with $\mathbb{P}^{n+1}$ obtained from Eqs. (14a) and (14b). and generalises the equivalence in Lemma 7 for Eqs. (8a) and (8b). Rewriting the $\{\mathbb{P}^n\}_{n\geq 1}$ updates as such allows us to give drift updates $v_t^n \to v_t^{n+1}$ very similar to that in Theorem 4.2 from [RKHK24].

Lemma 13. *Under the assumption of initialization with $(v_t^0)_{t\leq T} = (u_t^{\mathsf{ref}})_{t\leq T}$, the update Eq. (32) can be written as a sequence of SDEs*

$$\mathrm{d}X_t = v_t^n(X_t)\mathrm{d}t + \sqrt{2}\mathrm{d}B_t, \quad X_0 \sim \mu \tag{33}$$

for $n = 1, 2, \ldots,$ with drift updates

$$v_t^{n+1}(x) = v_t^n(x) - 2\eta\nabla \log p_t^n(x) + 2\eta\nabla \log p_t^{n+1/2}(x) - 2\nabla V_t(x)\,.$$

Above $p_t^n, p_t^{n+1/2}$ are the marginal densities of $\mathbb{P}^n, \mathbb{P}^{n+1/2}$ and $\nabla V_t(x)$ can be evaluated explicitly.

Note that when $\eta = 1$, we simply get $v_t^{n+1} = v_t^n - 2(\nabla \log p_t^n - \nabla \log p_t^{n+1/2})$ as an update for the drift in Eq. (33), which is time reversal applied twice and $p_t^n, p_t^{n+1/2}$ are the time-marginals corresponding to $\mathbb{P}^n, \mathbb{P}^{n+1/2}$. More generally, Lemma 13 offers a primal measure space implementation with iterative application of score matching time-reversals (for learning $p_t^{n+1/2}, p_t^n$) and stochastic optimal controls (for learning V_t) similar to [RKHK24, Proposition C.2].

Because of the projection in Eq. (14b), each path measure always has the marginal $\mathbb{P}_0^n = \mu$ constrained. With a very similar proof as Lemma 6, one can show iterates $(\mathbb{P}^n)_{n\geq 1}$ from Eq. (14a)-Eq. (14b) remain in the factorized form Eq. (19) if initialized as such, and constitutes an update on the $\boldsymbol{\phi}_T$ function as

$$\boldsymbol{\phi}_T^{n+1} = \boldsymbol{\phi}_T^n - \eta \cdot (\log \Phi(\mathrm{d}\mathbb{P}_T^n) - \log \Phi(\mathrm{d}\nu))\ .$$

This implies that the intermediate path measures $\{\mathbb{P}^n\}_{n\geq 1}$ solves the SB problem for its respective current marginal $(\mu, \mathbb{P}_T^n)$ with $\mathbb{P}^n(X_{t\in(0,T)}|X_0, X_T) = \mathbb{P}^{\mathsf{ref}}(X_{t\in(0,T)}|X_0, X_T)$. Therefore as $\mathbb{P}_T^n \to \nu$, we have $\mathbb{P}^n \to \mathbb{P}^\star$.

C.2 Pathspace Φ-match: Dual

In this section, we propose a dual potential space implementation of Φ-match that is similar to the continuous flow of SDE viewpoint in [RKHK24, Section 4.4]. In contrast to Appendix C.1, which is based purely on projections / optimisation of path measures, this is built on the dual potential updates from static eOT. Interestingly, both result in the same sequence of path measures $\{\mathbb{P}^n\}_{n\geq 1}$.

In order to bridge the two primal and dual spaces, we leverage Eq. (19). Importantly for $g^\star(y) = e^{\phi^\star(y)}$, $f^\star(x) = e^{\psi^\star(x)}$ [Lé14, Theorem 3.4] where each evolves as Kolmogorov Forward/Backward Equation under Eq. (11) as

$$\partial_t f_t^\star + \nabla \cdot (u^{\mathsf{ref}} f_t^\star) - \Delta f_t^\star = 0, \quad f_0^\star(x) = e^{\psi^\star(x)}$$

$$\partial_t g_t^\star + u^{\mathsf{ref}\top} \nabla g_t^\star + \Delta g_t^\star = 0, \quad g_T^\star(y) = e^{\phi^\star(y)}\,,$$

we have

$$\frac{\mathbb{P}^\star_{k:l}}{\mathbb{P}^{\mathsf{ref}}_{k:l}}(x_{k:l}) = f^\star_k(x_k) g^\star_l(x_l) \quad \forall\, k < l \in [0, T]\,.$$

Therefore with their dynamics fixed, knowing the two boundary functions $\boldsymbol{\phi}^\star_T, \boldsymbol{\psi}^\star_0$ gives us all the information about the optimal path measure $\mathbb{P}^\star$. Taking hints, we keep our sequence of path measures $\{\mathbb{P}^n\}_{n\geq 1}$ in the factorized form involving $\boldsymbol{\phi}^n_T, \boldsymbol{\psi}^n_0$ only:

$$\frac{\mathbb{P}^n}{\mathbb{P}^{\mathsf{ref}}}(X^n_t)_{0\leq t\leq T} = \frac{\mathbb{P}^n_{0,T}}{\mathbb{P}^{\mathsf{ref}}_{0,T}}(X^n_0, X^n_T) = \exp(\boldsymbol{\phi}^n_T(X^n_T) + \boldsymbol{\psi}^n_0(X^n_0))\,. \tag{34}$$

This implies that they always solve the SB problem for their own marginals, and consequently remain within the reciprocal and Markovian class. If moreover, $\boldsymbol{\psi}^n_0 = (\boldsymbol{\phi}^n_T)^+$ then we always have $\mathbb{P}^n_0 = \mu$. In the following lemma, which formalises Lemma 10, we derive updates on the SDE drift by leverging the corresponding Schrödinger potentials from the static setting.

Lemma 14 (Formal version of Lemma 10)**.** *Let $\{\phi^n\}_{n\geq 1}$ be the sequence of potentials obtained from Φ-match. For any $n \geq 1$, let $(g^n_t)_{t\in[0,T]}$ be the solution to the PDE*

$$\partial_t g^n_t + \nabla g^{n\top}_t u^{\mathsf{ref}} + \Delta g^n_t = 0, \;\; g^n_T(y) = e^{\phi^n(y)}\,.$$

Define $e^{\phi^n_t} = g^n_t$. Simulating the resulting SDE as

$$\mathrm{d}X^n_t = u^{\mathsf{ref}}(X^n_t) + \nabla\phi^n_t(X^n_t)\mathrm{d}t + \sqrt{2}\cdot \mathrm{d}B_t,\; X^n_0 \sim \mu\,, \tag{35}$$

the corresponding path measure $\mathbb{P}^n$ factorizes as Eq. (34) and converges to $\mathbb{P}^\star$ at the same rate as the sequence $\{\phi^n\}_{n\geq 1}$ converges to $\phi^\star$ for the eOT problem.

Note that the additional drift in Eq. (35) implies that v_t in Eq. (18) is necessarily a gradient vector field. Moreover, via a Cole-Hopf transform, one can also write a PDE dynamics on ϕ_t directly (instead of the g_t that we have above) similar to that in Eq. (20). This is also evident from the Feynman-Kac formula, which prescribes that $(\phi^n_t)_{t\in[0,T]}$ can be expressed as

$$\phi^n_t(y) = \log \mathbb{E}^{\mathsf{ref}}[e^{\phi^n(x_T)}|x_t = y]$$

with respect to the the reference process Eq. (11).

D Proofs

D.1 Proofs of Lemmas in Section 3

D.1.1 Proofs of Lemmas 1 to 3

We obtain the Lemmas 1 to 3 as corollaries of the following broader result.

Lemma 15. *Let $\overline{\phi}, \phi \in \mathcal{S}$ where $\mathcal{S} \subseteq L^1(\nu)$. 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*

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

Then,

$$\langle \delta J(\phi) - \delta J(\overline{\phi}), \phi - \overline{\phi}\rangle = -\int_0^1 \mathrm{Var}_{\rho_t(.;x)}[\overline{\phi} - \phi]\;\mathrm{d}t\,.$$

We give the proof of Lemma 15 in Appendix D.1.4.

Proof of Lemma 1. From Lemma 15, we have that for any $\phi, \overline{\phi} \in L^1(\nu)$

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

Since the variance is non-negative, this results in

$$\langle \delta J(\phi) - \delta J(\overline{\phi}), \phi - \overline{\phi}\rangle \leq 0\,.$$

In the notation of Lemma 15, define $\tilde{J}_t = J(\tilde{\phi}_t) - \langle \delta J(\phi), \tilde{\phi}_t \rangle$. Note that

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

By the fundamental theorem of calculus,

$$\begin{aligned} J(\overline{\phi}) - J(\phi) - \langle \delta J(\phi), \overline{\phi} - \phi \rangle &= J_1 - J_0 \\ &= \int_0^1 \dot{\tilde{J}}_t \, \mathrm{d}t \\ &= \int_0^1 \langle \delta J(\tilde{\phi}_t) - \delta J(\phi), \overline{\phi} - \phi \rangle \, \mathrm{d}t \\ &\leq \int_0^1 \frac{1}{t} \cdot \langle \delta J(\tilde{\phi}_t) - \delta J(\phi), \tilde{\phi}_t - \phi \rangle \, \mathrm{d}t \leq 0 \ . \end{aligned}$$

This completes the proof. □

Proof of Lemma 2. From Lemma 15, we have that for any $\phi, \overline{\phi} \in L^1(\nu)$

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

By definition of the variance

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

The final inequality is by Hölder's inequality. Substituing this in the result of Lemma 15, we get

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

□

Proof of Lemma 3. From the convexity of $\mathcal{S}_B$, note that $\tilde{\phi}_t = (1-t)\phi + t\overline{\phi} \in \mathcal{S}_B$. Since $\mathcal{S}_B \subset L^1(\nu)$, we have by Lemma 15 that

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

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

Step (a) uses Jensen's inequality, and step (b) uses the fact that for y, y', $\tilde{\phi}_t(y) - \tilde{\phi}_t(y') \leq 2B$ for y, y' almost everywhere. Therefore,

$$\langle \delta J(\phi) - \delta J(\overline{\phi}), \phi - \overline{\phi} \rangle \geq -\lambda(B) \cdot \|\overline{\phi} - \phi\|^2_{L^2(\nu)} \ .$$

□

D.1.2 Proofs of Lemma 4

Proof. From Lemma 2 and Proposition 1, we have for any $\phi, \overline{\phi} \in L^1(\nu) \cap L^\infty(\mathcal{Y})$ that

$$\begin{aligned} J(\overline{\phi}) - J(\phi) - \langle \delta J(\phi), \overline{\phi} - \phi \rangle &= \int_0^1 \langle \delta J(\phi + t(\overline{\phi} - \phi)) - \delta J(\phi), \overline{\phi} - \phi \rangle \mathrm{d}t \\ &\geq -\frac{\|\overline{\phi} - \phi\|_\infty^2}{2} \,. \end{aligned}$$

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

$$\begin{aligned} J(\phi^{n+1/2}) &\geq J(\phi^n) + \langle \delta J(\phi^n), \phi^{n+1/2} - \phi^n \rangle - \frac{\|\phi^{n+1/2} - \phi^n\|_\infty^2}{2} \\ &= J(\phi^n) + \|\delta J(\phi^n)\|_{L^1(\mathcal{Y})} \cdot \|\delta J(\phi^n)\|_{L^1(\mathcal{Y})} - \frac{1}{2} \cdot \|\delta J(\phi^n)\|_{L^1(\mathcal{Y})}^2 \\ &= J(\phi^n) + \frac{\|\delta J(\phi^n)\|_{L^1(\mathcal{Y})}^2}{2} \,. \end{aligned} \tag{36}$$

The second equality above is due to the fact that $\langle \mathrm{sign}(\delta J(\phi^n)), \delta J(\phi^n) \rangle = \|\delta J(\phi^n)\|_1$. Since the semi-dual is shift-invariant i.e., $J(\phi - C \cdot \mathbf{1}) = J(\phi)$ for any constant $C \in \mathbb{R}$, we have that $\phi \leftarrow \phi^{n+1/2}$ and $C \leftarrow \phi^{n+1/2}(y_{\mathtt{anc}}) - \phi^n(y_{\mathtt{anc}})$

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

Hence $\phi^{n+1} \in \mathcal{T}_0$ as $\phi^{n+1}(y_{\mathtt{anc}}) = \phi^n(y_{\mathtt{anc}})$. Next, by concavity of J (Lemma 1) that

$$J^\star \leq J(\phi^n) + \langle \delta J(\phi^n), \phi^\star - \phi^n \rangle \,. \tag{37}$$

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

$$\begin{aligned} E_{n+1} - E_n &= n(n+1) \cdot (J(\phi^{n+1}) - J(\phi^n)) + n \cdot (J(\phi^n) - J^\star) \\ &\overset{(a)}{\geq} n \cdot \left\{ \frac{n+1}{2} \cdot \|\delta J(\phi^n)\|_1^2 + \langle \delta J(\phi^n), \phi^n - \phi^\star \rangle \right\} \\ &\overset{(b)}{\geq} -\frac{n}{2(n+1)} \cdot \|\phi^n - \phi^\star\|_\infty^2 \\ &\overset{(c)}{\geq} -\frac{\mathrm{diam}(\mathcal{T}_0; L^\infty(\mathcal{Y}))^2}{2} \,. \end{aligned}$$

Above, step (a) applies Eqs. (36) and (37), and step (b) applies the Hölder-Young inequality. Finally, step (c) uses the fact that $\phi^n \in \mathcal{T}_0$ shown previously. Summing the above inequality from $n = 0$ to $n = N - 1$, we get

$$\begin{aligned} E_N - E_0 &\geq -\frac{N}{2} \cdot \mathrm{diam}(\mathcal{T}_0; L^\infty(\mathcal{Y}))^2 \\ \Rightarrow J(\phi^n) - J^\star &\geq -\frac{\mathrm{diam}(\mathcal{T}_0; L^\infty(\mathcal{Y}))^2}{2(N+1)} \,. \end{aligned}$$

□

D.1.3 Proof of Lemma 5

Before we give the proof, we lay out some preliminaries and intermediate results that will come in handy to prove Lemma 9 later in Appendix D.3.1.

We begin with the following definitions which are based on [BT09]. Let $\mathcal{S}$ be a convex subset of $L^2(\nu)$. The $\mathcal{S}$-truncated quadratic approximation to J centered at a given $\phi \in L^2(\nu)$ is

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

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

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

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

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

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

Lemma 16. *Let $\phi \in \mathcal{S}_{\bar{B}}$. Then, for $\eta \le \frac{1}{\lambda(\max\{B,\bar{B}\})}$,*

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

Lemma 17. *Let $\phi \in \mathcal{S}_{\bar{B}}$. For any $\overline{\phi} \in L^2(\nu)$ and $\eta \le \frac{1}{\lambda(\max\{B,\bar{B}\})}$, we have that*

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

The proof of the above two lemmas are given in Appendix D.1.4.

Proof of Lemma 5. For $\phi^0 \in \mathcal{S}_B$, each step according to proj-SGA ensures that $\phi^n \in \mathcal{S}_B$ for all $n \ge 1$. For $\eta \le \frac{1}{\lambda(B)}$, we have from Lemma 17 applied to $\overline{\phi} \leftarrow \widetilde{\phi}^\star$ and $\phi \leftarrow \phi^n$ for an arbitrary $n \ge 0$ that

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

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

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

Additionally from Lemma 16, we have for the choice of η,

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

Hence,

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

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

D.1.4 Proofs of intermediate lemmas in this subsection

Proof of Lemma 15. Recall that the first variation of the semi-dual J is

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

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

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

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

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

Define the interpolation $\tilde{\phi}_t(y) = (1-t)\phi(y) + t\overline{\phi}(y)$. We consider the following density (for a fixed $x \in \mathcal{X}$)

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

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

$$\mathrm{d}\dot{\rho}_t(y;x) = \left\{(\overline{\phi}(y) - \phi(y)) - \int_{\mathcal{Y}} (\overline{\phi}(y') - \phi(y'))\mathrm{d}\rho_t(y';x)\right\} \mathrm{d}\rho_t(y;x) \,.$$

Consequently,

$$\begin{aligned}
\langle j_x(\phi) - j_x(\overline{\phi}), \phi - \overline{\phi}\rangle &= \int_{\mathcal{Y}} (\phi(y) - \overline{\phi}(y)) \cdot (\mathrm{d}\rho_1(y;x) - \mathrm{d}\rho_0(y;x)) \\
&= \int_{\mathcal{Y}} \int_0^1 (\phi(y) - \overline{\phi}(y)) \cdot \mathrm{d}\dot{\rho}_t(y;x)\mathrm{d}t \\
&= -\int_0^1 \int_{\mathcal{Y}} (\overline{\phi}(y) - \phi(y))^2 \mathrm{d}\rho_t(y;x)\mathrm{d}t \\
&\quad + \int_0^1 \left\{\int_{\mathcal{Y}} (\overline{\phi}(y) - \phi(y)) \cdot \mathrm{d}\rho_t(y;x)\right\}^2 \mathrm{d}t \\
&= -\int_0^1 \mathrm{Var}_{\rho_t(.;x)}[\overline{\phi} - \phi] \,. \qquad (38)
\end{aligned}$$

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

$$\langle \delta J(\phi) - \delta J(\overline{\phi}, \phi - \overline{\phi})\rangle = -\int_0^1 \mathbb{E}_{x\sim\mu}\left[\mathrm{Var}_{\rho_t(.;x)}[\overline{\phi} - \phi]\right] \,. \qquad (39)$$

□

Proof of Lemma 16. Note that $\mathsf{M}_{\mathcal{S}_B}^{\text{proj-SGA}}(\phi;\eta) \in \mathcal{S}_B$, and therefore $\phi, \mathsf{M}_{\mathcal{S}_B}^{\text{proj-SGA}}(\phi;\eta) \in \mathcal{S}_{\max\{B,\bar{B}\}}$. By Lemma 3 and Proposition 1,

$$\begin{aligned}
\overline{J}_{\mathcal{S}_B}(\mathsf{M}_{\mathcal{S}_B}^{\text{proj-SGA}}(\phi;\eta)) &= J(\mathsf{M}_{\mathcal{S}_B}^{\text{proj-SGA}}(\phi;\eta)) - \mathbb{I}_{\mathcal{S}_B}(\mathsf{M}_{\mathcal{S}_B}^{\text{proj-SGA}}(\phi;\eta)) \\
&\geq J(\phi) + \left\langle \frac{\delta J(\phi)}{\mathrm{d}\nu}, \mathsf{M}_{\mathcal{S}_B}^{\text{proj-SGA}}(\phi;\eta) - \phi \right\rangle_\nu \\
&\quad - \frac{\lambda(\max\{B,\bar{B}\})}{2} \cdot \|\mathsf{M}_{\mathcal{S}_B}^{\text{proj-SGA}}(\phi;\eta) - \phi\|_\nu^2 - \mathbf{1}_{\mathcal{S}_B}(\mathsf{M}_{\mathcal{S}_B}^{\text{proj-SGA}}(\phi;\eta)) \\
&= \widetilde{J}_{\eta,\mathcal{S}_B}(\mathsf{M}_{\mathcal{S}_B}^{\text{proj-SGA}}(\phi;\eta);\phi) \\
&\quad + \left(\frac{1}{2\eta} - \frac{\lambda(\max\{\bar{B},B\})}{2}\right) \cdot \|\mathsf{M}_{\mathcal{S}_B}^{\text{proj-SGA}}(\phi;\eta) - \phi\|_{L^2(\nu)}^2 \\
&\geq \widetilde{J}_{\eta,\mathcal{S}_B}(\mathsf{M}_{\mathcal{S}_B}^{\text{proj-SGA}}(\phi;\eta);\phi) \,.
\end{aligned}$$

□

Proof of Lemma 17. By optimality, note that for any $\phi \in L^2(\nu)$

$$\delta J(\phi) - \frac{1}{\eta} \cdot \mathrm{d}\nu \cdot (\mathsf{M}_{\mathcal{S}}^{\text{proj-SGA}}(\phi;\eta) - \phi) - \gamma(\phi) = 0; \qquad \gamma(\phi) \in \partial\mathbb{I}_{\mathcal{S}}(\mathsf{M}_{\mathcal{S}}^{\text{proj-SGA}}(\phi;\eta)) \,. \qquad (40)$$

From Lemma 16, we know that

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

Consequently,

$$\begin{aligned}
\overline{J}(\mathsf{M}_{\mathcal{S}_B}^{\text{proj-SGA}}(\phi;\eta)) - \overline{J}(\overline{\phi}) &\geq \widetilde{J}_{\eta,\mathcal{S}_B}(\mathsf{M}_{\mathcal{S}_B}^{\text{proj-SGA}}(\phi;\eta);\phi) - \overline{J}(\overline{\phi}) \\
&= \widetilde{J}_{\eta,\mathcal{S}_B}(\mathsf{M}_{\mathcal{S}_B}^{\text{proj-SGA}}(\phi;\eta);\phi) - J(\overline{\phi}) + \mathbb{I}_{\mathcal{S}_B}(\overline{\phi}) \\
&\overset{(a)}{=} J(\phi) + \left\langle \frac{\delta J(\phi)}{\mathrm{d}\nu}, \mathsf{M}_{\mathcal{S}_B}^{\text{proj-SGA}}(\phi;\eta) - \phi \right\rangle_{L^2(\nu)} - J(\overline{\phi}) \\
&\quad - \mathbb{I}_{\mathcal{S}_B}(\mathsf{M}_{\mathcal{S}_B}^{\text{proj-SGA}}(\phi;\eta)) - \frac{1}{2\eta}\|\mathsf{M}_{\mathcal{S}_B}^{\text{proj-SGA}}(\phi;\eta) - \phi\|_\nu^2 + \mathbb{I}_{\mathcal{S}_B}(\overline{\phi}) \\
&\overset{(b)}{\geq} \left\langle \frac{\delta J(\phi)}{\mathrm{d}\nu}, \mathsf{M}_{\mathcal{S}_B}^{\text{proj-SGA}}(\phi;\eta) - \overline{\phi} \right\rangle_{L^2(\nu)} + \langle \gamma(\phi), \overline{\phi} - \mathsf{M}_{\mathcal{S}_B}^{\text{proj-SGA}}(\phi;\eta)\rangle \\
&\quad - \frac{1}{2\eta}\|\mathsf{M}_{\mathcal{S}_B}^{\text{proj-SGA}}(\phi;\eta) - \phi\|_{L^2(\nu)}^2 \\
&= \left\langle \frac{\delta J(\phi)}{\mathrm{d}\nu} - \frac{\gamma(\phi)}{\mathrm{d}\nu}, \mathsf{M}_{\mathcal{S}_B}^{\text{proj-SGA}}(\phi;\eta) - \overline{\phi} \right\rangle_{L^2(\nu)} \\
&\quad - \frac{1}{2\eta}\|\mathsf{M}_{\mathcal{S}_B}^{\text{proj-SGA}}(\phi;\eta) - \phi\|_{L^2(\nu)}^2 \\
&\overset{(c)}{=} \frac{1}{\eta} \cdot \langle \mathsf{M}_{\mathcal{S}_B}^{\text{proj-SGA}}(\phi;\eta) - \phi, \mathsf{M}_{\mathcal{S}_B}^{\text{proj-SGA}}(\phi;\eta) - \overline{\phi}\rangle_{L^2(\nu)} \\
&\quad - \frac{1}{2\eta}\|\mathsf{M}_{\mathcal{S}_B}^{\text{proj-SGA}}(\phi;\eta) - \phi\|_{L^2(\nu)}^2 \\
&= \frac{1}{2\eta} \cdot \|\mathsf{M}_{\mathcal{S}_B}^{\text{proj-SGA}}(\phi;\eta) - \phi\|_{L^2(\nu)}^2 \\
&\quad + \frac{1}{\eta} \cdot \langle \mathsf{M}_{\mathcal{S}_B}^{\text{proj-SGA}}(\phi;\eta) - \phi, \phi - \overline{\phi}\rangle_{L^2(\nu)}\ .
\end{aligned}$$

Step (a) uses the definition of $\widetilde{J}_{\eta,\mathcal{S}_B}$, step (b) uses the concavity of J and the convexity of $\mathbb{I}_{\mathcal{S}_B}$ as

$$\begin{aligned}
-J(\overline{\phi}) + J(\phi) &\geq \langle \delta J(\phi), \phi - \overline{\phi}\rangle\ , \\
\mathbb{I}_{\mathcal{S}_B}(\overline{\phi}) - \mathbb{I}_{\mathcal{S}_B}(\mathsf{M}_{\mathcal{S}_B}^{\text{proj-SGA}}(\phi;\eta)) &\geq \langle \gamma(\phi), \overline{\phi} - \mathsf{M}_{\mathcal{S}_B}^{\text{proj-SGA}}(\phi;\eta)\rangle\ .
\end{aligned}$$

Finally, we use the optimality condition (Eq. (40)) in step (c). □

D.2 Proofs of Lemmas in Section 4

D.2.1 Proof of Lemma 6

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

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

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

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

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 $\phi^{n+1/2}(y) = \phi^n(y) + (\log \Phi(\mathrm{d}\nu)(y) - \log \Phi(\mathrm{d}\pi^n_{\mathcal{Y}})(y))$ and $\psi^{n+1/2}(x) = \psi^n(x) + \log Z$. From [RKHK24, Corr. B.1], we have that $\mathsf{project}_{\mathcal{X},\mu}(\pi^{n+1/2}, \pi^n; \eta)$ satisfies

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

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,

$$\begin{aligned} \bar{\phi}(y) &= \eta \cdot \phi^{n+1/2}(y) + (1-\eta) \cdot \phi^n(y) \\ &= \phi^n(y) + \eta \cdot (\log \Phi(\mathrm{d}\nu)(y) - \log \Phi(\mathrm{d}\pi^n_{\mathcal{Y}})(y)) \; . \end{aligned} \tag{41}$$

Since $\mathsf{project}_{\mathcal{X},\mu}(\pi^{n+1/2}, \pi^n; \eta) = \mu$, 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. (41) 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. □

D.2.2 Proof of Lemma 7

Proof. For convenience, we use the shorthand notation $\tilde{\pi} = \mathsf{project}_{\mathcal{Y}}(\pi; \Phi)$. As in the proof of Lemma 6, Eq. (8a) ensures that

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

The objective in Eq. (9) with $\mathcal{F} \leftarrow \mathcal{V}_\Phi$ can be simplified as

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

$$+ c(\pi) \ .$$

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

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

thus establishing the equivalence in the statement as $\mathsf{project}_{\mathcal{X},\mu}$ also minimises over the set $\{\pi : \pi_{\mathcal{X}} = \mu\}$. □

D.2.3 Proof of Lemma 8

Proof. For an arbitrary $n \geq 0$, we have the following identity for any $\overline{\pi}$ such that $\overline{\pi}_{\mathcal{X}} = \mu$ that

$$\begin{aligned} &\eta \cdot \langle \mathcal{V}_{\Phi_k}(\pi^n), \mathrm{d}\overline{\pi} - \mathrm{d}\pi^n \rangle + \mathrm{d_{KL}}(\overline{\pi}\|\pi^n) \\ &\qquad \geq \eta \cdot \langle \mathcal{V}_{\Phi}(\pi^n), \mathrm{d}\pi^{n+1} - \mathrm{d}\pi^n \rangle + \mathrm{d_{KL}}(\pi^{n+1}\|\pi^n) + \mathrm{d_{KL}}(\overline{\pi}\|\pi^{n+1}) \ . \end{aligned} \tag{42}$$

This is obtained by the three-point identity [AFKL22, Lem. 3] with

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

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

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

From Proposition 3, we know that that in this case

$$\begin{aligned} \mathcal{L}_k(\pi^{n+1}_{\mathcal{Y}}; \nu) &\leq \mathcal{L}_k(\pi^n_{\mathcal{Y}}; \nu) + \langle \delta\mathcal{L}_k(\pi^n_{\mathcal{Y}}; \nu), \mathrm{d}\pi^{n+1}_{\mathcal{Y}} - \mathrm{d}\pi^n_{\mathcal{Y}} \rangle + 2c_k \cdot \mathrm{d_{KL}}(\pi^{n+1}_{\mathcal{Y}}\|\pi^n_{\mathcal{Y}}) \\ &= \mathcal{L}_k(\pi^n_{\mathcal{Y}}; \nu) + \langle \mathcal{V}_{\Phi}(\pi^n), \mathrm{d}\pi^{n+1} - \mathrm{d}\pi^n \rangle + 2c_k \cdot \mathrm{d_{KL}}(\pi^{n+1}_{\mathcal{Y}}\|\pi^n_{\mathcal{Y}}) \ . \end{aligned} \tag{43}$$

With the data-processing inequality,

$$\begin{aligned} \mathcal{L}_k(\pi^{n+1}_{\mathcal{Y}}; \nu) &\leq \mathcal{L}_k(\pi^n_{\mathcal{Y}}; \nu) + \langle \mathcal{V}_{\Phi}(\pi^n), \mathrm{d}\pi^{n+1} - \mathrm{d}\pi^n \rangle + 2c_k \cdot \mathrm{d_{KL}}(\pi^{n+1}\|\pi^n) \\ &\overset{(a)}{\leq} \mathcal{L}_k(\pi^n_{\mathcal{Y}}; \nu) + \left(2c_k - \frac{1}{\eta}\right) \cdot \mathrm{d_{KL}}(\pi^{k+1}\|\pi^k) - \frac{1}{\eta} \cdot \mathrm{d_{KL}}(\pi^k\|\pi^{k+1}) \\ &\leq \mathcal{L}_k(\pi^n_{\mathcal{Y}}; \nu) \ . \end{aligned}$$

Step (a) applies Eq. (42) for $\overline{\pi} \leftarrow \pi^n$, and step (b) uses the fact that $\eta = \frac{1}{2c_k}$ and the non-negativity of the KL divergence. We also have by Proposition 3 that

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

Substituing the above and Eq. (43) in Eq. (42), we get

$$\frac{1}{4c_k}\mathrm{MMD}(\pi^{n+1}_{\mathcal{Y}}, \nu)^2 - \frac{1}{4c_k}\mathrm{MMD}(\overline{\pi}_{\mathcal{Y}}, \nu)^2 \leq \mathrm{d_{KL}}(\overline{\pi}\|\pi^n) - \mathrm{d_{KL}}(\overline{\pi}\|\pi^{n+1}) \ .$$

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

$$\frac{1}{4c_k} \sum_{n=1}^{N} \{\mathrm{MMD}(\pi^n_{\mathcal{Y}}, \nu)^2 - \mathrm{MMD}(\overline{\pi}_{\mathcal{Y}}, \nu)^2\} \leq \mathrm{d_{KL}}(\overline{\pi}\|\pi^0) - \mathrm{d_{KL}}(\overline{\pi}\|\pi^N) \ .$$

Since we know that $\frac{1}{2}\mathrm{MMD}(\pi^{n+1}_{\mathcal{Y}}, \nu)^2 \leq \frac{1}{2}\mathrm{MMD}(\pi^n_{\mathcal{Y}}, \nu)^2$ and that there exists $\overline{\pi}$ with $\overline{\pi}_{\mathcal{Y}} = \nu$ given by $\pi^\star$

$$\mathrm{MMD}(\pi^N_{\mathcal{Y}}, \nu)^2 \leq \frac{4c_k}{N} \cdot \mathrm{d_{KL}}(\pi^\star\|\pi^0) \ .$$

□

Remark 1. Suppose $\phi^0 = \mathbf{0}$, and $\pi^0 = \pi(\phi^0, (\phi^0)^+)$. Then from [Lég21, Cor. 1], we know that

$$\mathrm{d_{KL}}(\pi^\star\|\pi^0) \leq \mathrm{d_{KL}}(\pi^\star\|R)$$

where R is the reference measure Eq. (2). Consequently, the rate we obtain is

$$\frac{4c_k}{N} \cdot \frac{\mathrm{OT}_\varepsilon(\mu, \nu; c)}{\varepsilon} \ .$$

This highlights the better dependence on ε compared to the more classical analyses of Sinkhorn for the eOT problem where the dependence on ε is of the form $e^{-\varepsilon^{-1}}$.

D.3 Proofs of Lemmas in Section 5

D.3.1 Proof of Lemma 9

Prior to stating the proof for Lemma 9, 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 $n \geq 1$, $\frac{t_n - 1}{t_{n+1}} \in (0, 1)$ (Lemma 19).

A key step towards the proof of Lemma 9 is the following lemma, analogous to [BT09, Lem. 4.1].

Lemma 18. *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,*

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

We give the proof of Lemma 18 in Appendix D.3.4.

Proof of Lemma 9. Since $\lambda(3B) \geq \lambda(B)$ and $\phi^1, \overline{\phi}^1 \in \mathcal{S}_B$, Lemma 17 with $\phi \leftarrow \phi^1, \overline{\phi} \leftarrow \widetilde{\phi}^\star$ gives

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

In the notation of Lemma 18,

$$-v_1 \geq \frac{\lambda(3B)}{2} \cdot \|u_1\|_{L^2(\nu)}^2 - \frac{\lambda(3B)}{2} \cdot \|\overline{\phi}^0 - \widetilde{\phi}^\star\|_{L^2(\nu)}^2 \,. \tag{44}$$

Telescoping the identity from Lemma 18 for $n = 1$ to $N - 1$ gives

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

Rearranging the terms, we have

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

where the last step follows from Eq. (44). Since $t_N \geq \frac{N+1}{2}$, we have

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

□

D.3.2 Proof of Lemma 13

Proof. The first half-step simply involves changing the initial condition of the SDE from $\mathbb{P}_T^n$ (assumed to have drift v^n) to

$$\mathrm{d}Y_t = [-v_{T-t}^n(Y_t) + 2\nabla \log p_{T-t}^n(Y_t)]\mathrm{d}t + \sqrt{2} \cdot \mathrm{d}B_t, \quad Y_0 \sim \mathbb{P}_T^n \cdot \frac{\Phi(\mathrm{d}\nu)}{\Phi(\mathrm{d}\mathbb{P}_T^n)}$$

since Eq. (14a) is simply a time reversal on the path space with a different marginal at time T. Using Theorem 4.2 from [RKHK24], we can represent the second half step Eq. (14b) as

$$\begin{aligned}
\mathrm{d}X_t &= \left(\eta \left[v_t^n(X_t) - 2\nabla \log p_t^n(X_t) + 2\nabla \log p_t^{n+1/2}(X_t) \right] + (1-\eta) v_t^n(X_t) - 2\nabla V_t(X_t) \right) \mathrm{d}t \\
&\quad + \sqrt{2} \cdot \mathrm{d}B_t \,, \\
&= [v_t^n(X_t) - 2\eta\nabla \log p_t^n(X_t) + 2\eta\nabla \log p_t^{n+1/2}(X_t) - 2\nabla V_t(X_t)]\mathrm{d}t + \sqrt{2} \cdot \mathrm{d}B_t
\end{aligned} \tag{45}$$

where $X_0 \sim \mu$. The extra drift V_t is given by

$$V_t(x) = -\log \mathbb{E}\left[\exp\left(-\frac{2\eta(1-\eta)}{2}\int_t^T \|\nabla \log p_s^{n+1/2}(Z_s) - \nabla \log p_s^n(Z_s)\|^2\, ds\right)\Big| Z_t = x\right]$$

where the expectation is taken over an SDE starting at $Z_t = x$ and following

$$dZ_s = [v_s^n(Z_s) + 2\eta\nabla \log p_s^{n+1/2}(Z_s) - 2\eta\nabla \log p_s^n(Z_s)]ds + \sqrt{2}\mathrm{d}B_t,\ s \geq t\,.$$

□

D.3.3 Proof of Lemma 14

Proof. We make note of two equivalences:

1. between the dual potential ϕ^n from eOT and backward dynamics on $\{\phi_t^n\}_t$ determined by the reference transition: this follows from [CH21, Lé14] in the case of nonlinear drift (i.e., $u^{\mathsf{ref}} \neq 0$).
2. between the updates on the drifts of the SDE $v_t^n = u^{\mathsf{ref}} + \nabla\phi_t^n$ and the path measures $\mathbb{P}^n$ factorized as Eq. (34): this follows from classical result on Doob's h-transform, which implies that the optimal additional drift for $\mathbb{P}^*$ should be in the form of $(\nabla \log g_t^*)_{t\in[0,T]}$ built from the optimal potential $g_T^* = e^{\phi^\star}$ from eOT. The fact that Eq. (34) is the same as the law of the SDE Eq. (35) is also a consequence of the same twisted kernel argument [DP91].

The rate claim follows from an application of chain rule that gives $\mathrm{d}_{\mathrm{KL}}(\mathbb{P}^n\|\mathbb{P}^*) = \mathrm{d}_{\mathrm{KL}}(\pi^n\|\pi^*)$ for this particular set of path measures that satisfies $\mathbb{P}^n(X_{t\in(0,T)}|X_0, X_T) = \mathbb{P}^{\mathsf{ref}}(X_{t\in(0,T)}|X_0, X_T)$. And since we have the $\mathbb{P}_0^n = \mu$ constrained in Eq. (35), and factorized as Eq. (34), this rate is determined by $\mathbb{P}_T^n \to \nu$ (or equivalently $\phi_T^n = \phi^n \to \phi^*$), similar to the static two-marginal case for Φ-match, where we also maintain the coupling π^n so they form the optimal coupling for its current marginals. □

D.3.4 Proofs of intermediate lemmas instantiated in this subsection

Proof of Lemma 18. First, since $\overline{\phi}^n \in \mathcal{S}_B$ for all $n \geq 1$ and $\frac{t_n-1}{t_{n+1}} \leq 1$, by the triangle inequality for the semi-norm $L^\infty(\mathcal{Y})$, we have that $\phi^n \in \mathcal{S}_{3B}$ for all $n \geq 0$. Now, we apply Lemma 17 to two settings. First, with $\phi \leftarrow \phi^{n+1}, \overline{\phi} \leftarrow \overline{\phi}^n, \bar{B} \leftarrow 3B$, we have

$$\overline{J}(\overline{\phi}^{n+1}) - \overline{J}(\overline{\phi}^n) \geq \frac{\lambda(3B)}{2}\cdot\|\overline{\phi}^{n+1} - \phi^{n+1}\|_{L^2(\nu)}^2 + \lambda(3B)\cdot\langle\overline{\phi}^{n+1} - \phi^{n+1}, \phi^{n+1} - \overline{\phi}^n\rangle_{L^2(\nu)}\,.$$

Second, with $\phi \leftarrow \phi^{n+1}, \overline{\phi} \leftarrow \widetilde{\phi}^\star, \bar{B} \leftarrow 3B$, we have

$$\overline{J}(\overline{\phi}^{n+1}) - \overline{J}(\widetilde{\phi}^\star) \geq \frac{\lambda(3B)}{2}\cdot\|\overline{\phi}^{n+1} - \phi^{n+1}\|_{L^2(\nu)}^2 + \lambda(3B)\cdot\langle\overline{\phi}^{n+1} - \phi^{n+1}, \phi^{n+1} - \widetilde{\phi}^\star\rangle_{L^2(\nu)}\,.$$

With the definition of v_k, the left hand sides of both inequalities are $v_k - v_{k+1}$ and $-v_{k+1}$ respectively. The remainder of the proof follows from the proof of [BT09, Lem. 4.1]. □

Lemma 19. *Consider the recursion*

$$t_{k+1} = \frac{1+\sqrt{1+4t_k^2}}{2} \qquad k \geq 1\,.$$

If $t_1 \geq 1$, *then* $0 \leq \frac{t_k-1}{t_{k+1}} \leq 1$.

Proof. Note that for any t_k, $\frac{1+\sqrt{1+4t_k^2}}{2} \geq \frac{1+1}{2} = 1$. Hence $\frac{t_k-1}{t_{k+1}} \geq 0$. Algebraically,

$$\begin{aligned} t_k - 1 \leq t_{k+1} &\Leftrightarrow t_k \leq t_{k+1} + 1 \\ &\Leftrightarrow t_k - \frac{3}{2} \leq \frac{\sqrt{1+4t_k^2}}{2} \\ &\Leftrightarrow 4t_k^2 + 9 - 12t_k \leq 1 + 4t_k^2 \\ &\Leftrightarrow \frac{2}{3} \leq t_k\,. \end{aligned}$$

Since we know that $t_k \geq 1 \geq \frac{2}{3}$, we have $\frac{t_k-1}{t_{k+1}} \leq 1$. □

D.4 Proofs of Lemmas in Appendix B

D.4.1 Proof of Lemma 11

Proof. We begin by writing the definition of $\delta\varphi(\pi^n)$

$$\delta\varphi(\pi^n) = \log \frac{\mathrm{d}\pi^n}{\mathrm{d}\pi_\varepsilon^{\mathrm{ref}}} = (-(\phi^n)^+) \oplus \phi^n \,.$$

By definition of $\mathcal{V}_\Phi$ and Φ-match,

$$\delta\varphi(\pi^n) - \eta \cdot \mathcal{V}_\Phi(\pi^n) = (-(\phi^n)^+) \oplus (\phi^n - \eta \cdot (\log \Phi(\mathrm{d}\pi_{\mathcal{Y}}^n) - \log \Phi(\mathrm{d}\nu))) = (-(\phi^n)^+) \oplus \phi^{n+1} \,.$$

Applying the mapping $\delta\varphi^*$, we have

$$\begin{aligned}\delta\varphi^*(\delta\varphi(\pi^n) - \eta \cdot \mathcal{V}_\Phi(\pi^n)) &= \delta\varphi^*((-(\phi^n)^+) \oplus \phi^{n+1}) \\ &\overset{(a)}{=} \delta\varphi^*(0 \oplus \phi^{n+1}) \,.\end{aligned}$$

In step (a), we use the observation that $\delta\varphi^*(h + (f \oplus 0)) = \delta\varphi^*(h)$ for any $h \in L^1(\mathcal{X} \times \mathcal{Y})$ and $f \in L^1(\mathcal{X})$ as noted in Eq. (23). By definition of $\delta\varphi^\star$, we note that for any $\phi \in L^1(\nu)$, we have

$$\begin{aligned}\delta\varphi^\star(0 \oplus \phi)(x,y) &= \mathrm{d}\mu(x) \cdot \frac{\mathrm{d}\pi_\varepsilon^{\mathrm{ref}}(x,y)\exp(\phi(y))}{\int_{\mathcal{Y}} \mathrm{d}\pi_\varepsilon^{\mathrm{ref}}(x,y')\exp(\phi(y'))} \\ &= \mathrm{d}\mu(x)\mathrm{d}\nu(y)\exp\left(\phi(y) - \frac{c(x,y)}{\varepsilon}\right) \cdot \left(\int_{\mathcal{Y}} \exp\left(\phi(y') - \frac{c(x,y')}{\varepsilon}\right)\mathrm{d}\nu(y')\right)^{-1} \\ &= \mathrm{d}\mu(x)\mathrm{d}\nu(y)\exp\left(\phi(y) - \phi^+(x) - \frac{c(x,y)}{\varepsilon}\right) \\ &= \mathrm{d}\pi(\phi, \phi^+) \,.\end{aligned}$$

As a result, we obtain

$$\delta\varphi^*(\delta\varphi(\pi^n) - \eta \cdot \mathcal{V}_\Phi(\pi^n)) = \mathrm{d}\pi(\phi^{n+1}, (\phi^{n+1})^+) = \mathrm{d}\pi^{n+1} \,,$$

showing the equivalence to Φ-match. □

D.4.2 Proof of Lemma 12

Proof. The proof is based on showing the decay of the following Lyapunov functional

$$V(\hat{\pi}_y^t, g^t, t) = \frac{t^2}{r}\mathcal{L}_k(\hat{\pi}_y^t; \nu) + r D_{\varphi^\star}((f^0 \oplus g^t); (f^0 \oplus g^\star)) \,,$$

where $D_{\varphi^\star}(h'; h) = \varphi^\star(h') - \varphi^\star(h) - \langle \delta\varphi^\star(h), h' - h\rangle$. The time derivative of V is

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}V(\hat{\pi}_y^t, g^t, t) &= \frac{2t}{r}\mathcal{L}_k(\hat{\pi}_y^t; \nu) + \frac{t^2}{r}\langle \delta\mathcal{L}_k(\hat{\pi}_y^t; \nu), \dot{\hat{\pi}}^t\rangle \\ &\quad + r\langle \dot{g}^t, \nabla\varphi^\star((f^0 \oplus g^t)) - \nabla\varphi^\star((f^0 \oplus g^\star))\rangle \\ &= \frac{2t}{r}\mathcal{L}_k(\hat{\pi}_y^t; \nu) + t\langle \delta\mathcal{L}_k(\hat{\pi}_y^t; \nu), \frac{t}{r}\dot{\hat{\pi}}^t - \nabla\varphi^\star((f^0 \oplus g^t)) + \nabla\varphi^\star((f^0 \oplus g^\star))\rangle \\ &= \frac{2t}{r}\mathcal{L}_k(\hat{\pi}_y^t; \nu) - t\langle \delta\mathcal{L}_k(\hat{\pi}_y^t; \nu), \hat{\pi}^t - \pi^\star\rangle \\ &= -\frac{t(r-2)}{r}\mathcal{L}_k(\hat{\pi}_y^t; \nu) \,.\end{aligned}$$

Therefore if $r \geq 2$, V is decreasing with time, and we have

$$\begin{aligned}\frac{t^2}{r}\mathcal{L}_k(\hat{\pi}_y^t; \nu) &\leq V(\hat{\pi}_y^t, g^t, t) \\ &\leq V(\hat{\pi}_y^0, g^0, 0) \\ &= r D_{\varphi^\star}((f^0 \oplus g^t), (f^0 \oplus g^\star)) \\ &= r \cdot \mathrm{d}_{\mathrm{KL}}(\pi(g^0, (g^0)^+), \pi^\star) \,.\end{aligned}$$

where in the last step, we used [RKHK24, Lem. B.1]. This completes the proof. □

E Remarks on k-SGA

E.1 Reparameterisation

Alternatively, one can also interpret k-SGA as gradient ascent on a re-parametrised dual objective. Let $\widetilde{J} : L^1(\nu) \to \mathbb{R}$ be defined as

$$\widetilde{J}(\varphi) = J\left(\int_{\mathcal{Y}} k(z,\cdot)\varphi(z)\mathrm{d}z\right) .$$

Let $M_k : L^1(\nu) \to L^\infty(\mathcal{Y})$ be defined as $M_k(g) = \int_{\mathcal{Y}} k(z,.)g(z)\mathrm{d}z$, and note that this is linear. Therefore,

$$\begin{aligned}\delta\widetilde{J}(\varphi)(y) &= \delta J(M_k(\varphi))[\delta M_k(\varphi)(y)] \\ &= \int_{\mathcal{Y}} \delta J(M_k(\varphi))(z)\cdot k(z,y)\mathrm{d}z .\end{aligned}$$

A gradient ascent procedure for $\widetilde{J}$ results in the iteration

$$\begin{aligned}\varphi^{n+1} &= \varphi^n + \eta\cdot\delta\widetilde{J}(\varphi^n) \qquad (46)\\ \Rightarrow M_k(\varphi^{n+1}) &= M_k\left(\varphi^n + \eta\cdot\delta\widetilde{J}(\varphi^n)\right)\\ &= M_k(\varphi^n) + \eta\cdot\int_{\mathcal{Y}} k(w,.)\cdot\delta\widetilde{J}(\varphi^n)(w)\mathrm{d}w\\ &= M_k(\varphi^n) + \eta\cdot\int_{\mathcal{Y}} k(w,.)\int_{\mathcal{Y}}\delta J(M_k(\varphi^n))(z)\cdot k(w,z)\mathrm{d}z\mathrm{d}w\\ &= M_k(\varphi^n) + \eta\cdot\int_{\mathcal{Y}}\delta J(M_k(\varphi^n))(z)\cdot k(z,.)\mathrm{d}z .\end{aligned}$$

This shows that a sequence of $\{\varphi^n\}_{n\geq 0}$ obtained through Eq. (46) can be mapped to $\{\phi^n\}_{n\geq 0}$ obtained from k-SGA as $\phi^n = M_k(\varphi^n)$ for all $n \geq 0$.

E.2 Particle Implementation

Given an oracle to draw samples from $\pi(\phi,\phi^+)_{\mathcal{Y}}$ and ν, the kernelised update of J defined in k-SGA allows a finite-particle approximation defined as

$$\widehat{m}_{\pi(\phi,\phi^+)_{\mathcal{Y}}}(y) - \widehat{m}_\nu(y) = \frac{1}{N}\sum_{i=1}^N \{k(z_i,y) - k(z_i',y)\} \quad \text{for} \quad \{z_i\}_{i=1}^N \sim \pi(\phi,\phi^+)_{\mathcal{Y}},\ \{z_i'\}_{i=1}^N \sim \nu .$$

Two choices of the kernel k that are notable here are the Stein kernels w.r.t. ν and $\pi(\phi,\phi^+)_{\mathcal{Y}}$. The Stein kernel for a distribution ρ and with a base kernel k is given by

$$\begin{aligned}k_\rho(x,x') = k(x,x')\cdot\langle\nabla\log\rho(x),\nabla\log\rho(x')\rangle + \langle\nabla\log\rho(x),\nabla_2 k(x,x')\rangle\\ + \langle\nabla\log\rho(x'),\nabla_1 k(x,x')\rangle + \nabla_1\cdot(\nabla_2 k(x,x')) .\end{aligned}$$

The key property of the Stein kernel is that

$$\int k_\rho(z,.)\rho(z)\mathrm{d}z = 0 .$$

When given an oracle to sample from $\pi(\phi,\phi^+)_{\mathcal{Y}}$ but not ν, it would be useful to consider the Stein kernel k_ν, and with this choice of $k \leftarrow k_\nu$, k-SGA becomes

$$\phi^{n+1} = \phi^n - \eta\cdot\int_{\mathcal{Y}} k_\nu(z,.)\pi(\phi^n,(\phi^n)^+)_{\mathcal{Y}}(z)\mathrm{d}z ,$$

and this is amenable to estimation by sampling from $\pi(\phi,\phi^+)_{\mathcal{Y}}$. The definition of the kernel requires knowledge of $\nabla\log\nu$ oblivious to potentially unknown normalizing constant. Alternatively, it is also possible to consider the Stein kernel w.r.t. $\pi(\phi,\phi^+)_{\mathcal{Y}}$, which requires access to $\nabla\log\pi(\phi,\phi^+)_{\mathcal{Y}}$, and k-SGA becomes

$$\phi^{n+1} = \phi^n + \eta\cdot\int_{\mathcal{Y}} k_{\pi(\phi^n,\phi^{n+})_{\mathcal{Y}}}(z,.)\nu(z)\mathrm{d}z .$$

This integral can be approximated with access to samples from ν, which can be obtained with access to $\nabla\log\nu$ using Langevin-style methods.

Special case of a Gaussian kernel When an oracle to draw samples from either $\pi(\phi, \phi^+)_{\mathcal{Y}}$ or ν is unavailable, a Gaussian kernel would be useful to consider. Recall that a Gaussian kernel (with bandwidth $h > 0$) is defined as

$$k_{\mathrm{Gauss}}(x, y) = \exp\left(-\frac{\|x - y\|^2}{2h}\right) .$$

Note that

$$\int_{\mathcal{Y}} k_{\mathrm{Gauss}}(z, y)\phi(z)\mathrm{d}z = \frac{\sqrt{2\pi h^d}}{\sqrt{2\pi h^d}} \int_{\mathcal{Y}} e^{-\frac{\|z-y\|_2^2}{2h}} \phi(z)\mathrm{d}z = \sqrt{2\pi h^d} \cdot \mathbb{E}_{\mathsf{z}\sim\mathcal{N}(y,hI_d)}[\phi(\mathsf{z})] .$$

Hence with samples $\{z_i\}_{i=1}^N$ and $\{z_i'\}_{i=1}^N$ from $\mathcal{N}(y, h \cdot I_d)$, we have

$$\widehat{m}_{\pi(\phi,\phi^+)_{\mathcal{Y}}}(y) - \widehat{m}_{\nu}(y) = \frac{\sqrt{2\pi h^d}}{N} \sum_{i=1}^{N} \left\{\pi(\phi, \phi^+)_{\mathcal{Y}}(z_i) - \nu(z_i')\right\} .$$