Designing Algorithms for Entropic Optimal Transport from an Optimisation Perspective

Vishwak Srinivasan VISHWAKS@MIT.EDU
Department of Electrical Engineering and Computer Science, MIT

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

Editors: Matus Telgarsky and Jonathan Ullman

Abstract

In this work, we develop 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.

Keywords: Entropic optimal transport, infinite-dimensional optimisation.

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. The OT problem was originally proposed by Gaspard Monge in the 1780s to address the problem of finding a method to transport resources between a collection of sources and sinks, and was rediscovered in the early 1900s by Hitchcock, Kantorovich, and Koopman with applications in designing transportation systems, coinciding with the birth of linear programming. Recent advances in computing resources has renewed interest in the OT problem, both in the design of approximate methods for this problem suited for large-scale settings (Peyré and Cuturi, 2019), and in the development of a theoretical understanding of its properties (Villani, 2003; Santambrogio, 2015).

The "optimality" in the OT problem is defined in terms of a cost function $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$. The optimal value of the problem results in a notion of discrepancy between μ and ν that complements information-theoretic discrepancy measures like the total variation distance or the Kullback-Leibler (KL) divergence. Formally, let $\Pi(\mu, \nu)$ be the set of all joint distributions whose marginals are μ and ν. The Kantorovich formulation of the OT problem is given by the following program:

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

. Both authors contributed equally.

A notable case is when $c(x, y) = \|x - y\|_p^p$ where $\|\cdot\|_p$ is the ℓ_p-norm, and this corresponds to the p-Wasserstein distance between μ and ν raised to the p^{th} power. The space of probability measures endowed with the p-Wasserstein distance takes on a rich Riemannian structure with implications in developing the metric theory of gradient flows (Ambrosio et al., 2008). Wasserstein distances have also found applications in image processing and operations research, which has motivated the design of efficient methods to compute these distances.

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

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

Above, $\mathrm{d}_{\mathrm{KL}}(\pi\|\mu \otimes \nu)$ is the KL divergence between π and the product distribution $\mu \otimes \nu$. We highlight that $\mathrm{OT}_\varepsilon(\mu, \nu; c)$ is *biased* relative to $\mathrm{OT}(\mu, \nu; c)$ as $\mathrm{OT}_\varepsilon(\mu, \mu; c) \neq 0$. To address this, a debiasing strategy is proposed in Feydy et al. (2019) which results in the *Sinkhorn divergence* which is shown to metricise convergence in law. The sample complexity to estimate the Sinkhorn divergence using samples from μ and ν scales much better than $\mathrm{OT}(\mu, \nu; c)$ for a variety of costs (Genevay et al., 2019; Mena and Niles-Weed, 2019; Chizat et al., 2020).

An algorithm for solving the eOT problem is the *Sinkhorn* algorithm (Sinkhorn and Knopp, 1967), and this was popularised by Cuturi (2013) by demonstrating it as a viable solution for solving the eOT problem over large datasets, and by extension for solving the OT problem approximately. We refer the reader to Peyré and Cuturi (2019, Remark 4.5) for a historical overview of the method. While Sinkhorn and Knopp (1967) proved that the method converges asymptotically, the first known non-asymptotic rate of convergence was given by Franklin and Lorenz (1989) by viewing the Sinkhorn algorithm as a matrix scaling method and leverage Hilbert's projective metric in conjunction with Birkhoff's theorem to arrive at their result. Other interpretations of the Sinkhorn algorithm have led to both asymptotic and non-asymptotic guarantees; see Rüschendorf (1995); Di Marino and Gerolin (2020); Carlier (2022) for instance. A more thorough discussion of prior work in this regard is given in Section 2. Of relevance to this paper is a more recent interpretation of the Sinkhorn algorithm as an *(infinite-dimensional) optimisation procedure*, which originated with the mirror descent interpretation of the Sinkhorn algorithm in Mishchenko (2019) for discrete spaces $\mathcal{X}$ and $\mathcal{Y}$. This interpretation has been instrumental in obtaining *assumption-free guarantees* for the Sinkhorn algorithm and has led to a growing body of literature since its inception (Mensch and Peyré, 2020; Léger, 2021; Aubin-Frankowski et al., 2022; Deb et al., 2023; Reza Karimi et al., 2024).

Summary of contributions

In this work, we draw inspiration from this refreshing viewpoint and propose methods for the eOT problem that we give non-asymptotic guarantees for. A key mathematical object that we consider in this design is the *semi-dual formulation* of the eOT problem that we describe in more detail in Section 3. This semi-dual formulation translates the eOT problem from the primal space of couplings $\Pi(\mu, \nu)$ into a concave unconstrained program over the space of suitably integrable functions (referred to as the dual space). We propose a new class of methods called Φ-match, which generalises

the dual space interpretation of the Sinkhorn algorithm. In Section 4, through an appropriate map from the dual space onto the primal space of joint distributions $\overline{\mathcal{Q}} = \{\pi \in \mathcal{P}(\mathcal{X}\times\mathcal{Y}) : \pi_{\mathcal{X}} = \mu\}$, we demonstrate how Φ-match operates over $\overline{\mathcal{Q}}$ by extending existing interpretations of the Sinkhorn algorithm by Reza Karimi et al. (2024).

A subclass of methods captured by Φ-match – which we refer to as k-SGA – can be seen as iteratively minimising the squared maximum mean discrepancy (Gretton et al., 2006) between the $\mathcal{Y}$-marginals for the joint distributions defined by the potentials and ν. More specifically, *without any assumptions* on the domains $\mathcal{X}, \mathcal{Y}$ and the marginal distributions μ, ν, we find that the $\mathcal{Y}$-marginal of the sequence of joint distributions resulting from these methods converges to ν at a rate that scales as $\frac{1}{N}$ where N is the number of iterations, and consequently leads to an optimal coupling for the eOT problem due to the form of the joint distribution that is maintained. A corollary of this broader result is the *first assumption-free guarantees* for the semi-dual gradient ascent (SGA) method that was originally proposed in Genevay et al. (2016) for discrete spaces $\mathcal{X}$, $\mathcal{Y}$.

By continuing to draw on the optimisation perspective, in Section 5.1, we reveal how principles in finite-dimensional optimisation can be adapted to growth conditions of the semi-dual, which leads to other steepest descent methods, along with an accelerated variant that extends Φ-match. By virtue of these principles, we also obtain non-asymptotic rates of convergence for these other methods under minimal assumptions on the problem structure. We conclude by discussing a path-space generalisation of Φ-match for solving the dynamical Schrödinger bridge problem in Section 5.2.

2. Background

Notation For a set $\mathcal{Z}$, the set of probability measures over $\mathcal{Z}$ is denoted by $\mathcal{P}(\mathcal{Z})$. Given a distribution $\pi \in \mathcal{P}(\mathcal{X}\times\mathcal{Y})$, $\pi_{\mathcal{X}}$ and $\pi_{\mathcal{Y}}$ denote the $\mathcal{X}$-marginal and $\mathcal{Y}$-marginal, respectively. Given two distributions $\mu \in \mathcal{P}(\mathcal{X}), \nu \in \mathcal{P}(\mathcal{Y})$, we say that $\pi \in \mathcal{P}(\mathcal{X}\times\mathcal{Y})$ is a coupling of μ and ν if $\pi_{\mathcal{X}} = \mu$ and $\pi_{\mathcal{Y}} = \nu$. For $\rho \in \mathcal{P}(\mathcal{Z}$, we use $\mathrm{d}\rho$ to represent its density. If ρ has a density w.r.t. the Lebesgue measure, we denote it by ρ as well.

For a functional $\mathcal{F} : \mathcal{P}(\mathcal{Z}) \to \mathbb{R}$, we call $\delta\mathcal{F}(\rho)$ its first variation at $\rho \in \mathcal{P}(\mathcal{Z})$, and this is the function (up to additive constants) that satisfies (Santambrogio, 2015, Def. 7.12)

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

For a measurable function $f : \mathcal{Z} \to \mathbb{R}$, we use L^p-norm w.r.t. ρ is denoted by $\|f\|_{L^p(\rho)}$, which is defined as $\left(\int_{\mathcal{Z}} |f(z)|^p \mathrm{d}\rho(z)\right)^{1/p}$. If ρ is replaced with $\mathcal{Z}$ as in $\|f\|_{L^p(\mathcal{Z})}$, then this is understood to be the L^p-norm of f w.r.t. the Lebesgue measure of $\mathcal{Z}$. For another measurable function $g : \mathcal{Z} \to \mathbb{R}$, the $L^2(\rho)$ inner product is defined as $\langle f, g\rangle_{L^2(\rho)} = \int_{\mathcal{Z}} f(z)g(z)\mathrm{d}\rho(z)$. As a special case, when the subscript in the norm and inner product are omitted as in $\|\cdot\|$ and $\langle\cdot,\cdot\rangle$, then these correspond to the $L^2(\mathcal{Z})$-norm and inner product respectively. Following the notation in Rudin (1987, Chap. 3), we use L^p to also denote the space of measurable functions whose L^p norm is finite.

2.1. The entropic optimal transport problem

The eOT problem defined in eq. (2) is directly related to the *static Schrödinger bridge problem* $\mathrm{SB}(\mu, \nu; \pi^{\mathsf{ref}})$ (Léonard, 2014, Def. 2.2) for the reference measure π^{ref} with density $\mathrm{d}\pi^{\mathsf{ref}} \propto$

$\exp\left(-c/\varepsilon\right) \mathrm{d}(\mu \otimes \nu)$ and marginals μ and ν and normalisation constant Z_{ref}. To be precise,

$$\mathrm{OT}_\varepsilon(\mu, \nu; c) = \varepsilon \cdot \Big\{ \underbrace{\inf_{\pi \in \Pi(\mu,\nu)} \mathrm{d}_{\mathrm{KL}}(\pi \| \pi^{\mathsf{ref}})}_{\mathrm{SB}(\mu,\nu;\pi^{\mathsf{ref}})} - \log Z_{\mathsf{ref}} \Big\} . \tag{3}$$

The above problem is a (strictly) convex minimisation problem as (a) $\Pi(\mu, \nu)$ is convex, and (b) $\pi \mapsto \mathrm{d}_{\mathrm{KL}}(\pi \| \pi^{\mathsf{ref}})$ is strictly convex. Moreover, under certain regularity conditions (Léonard, 2014; Nutz and Wiesel, 2022), the eOT problem admits a unique solution $\pi^\star \in \Pi(\mu, \nu)$ of the form

$$\mathrm{d}\pi^\star(x, y) = \exp\left(\phi^\star(y) - \psi^\star(x) - \frac{c(x, y)}{\varepsilon}\right) \mathrm{d}\mu(x)\mathrm{d}\nu(y) \tag{4}$$

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

$$\begin{aligned} \mathrm{OT}_\varepsilon(\mu, \nu; c) = \varepsilon \cdot \sup \Bigg\{ & \int \phi(y)\mathrm{d}\nu(y) - \int \psi(x)\mathrm{d}\mu(x) \\ & - \log \iint \exp\left(\phi(y) - \psi(x) - \frac{c(x, y)}{\varepsilon}\right) \mathrm{d}\mu(x)\mathrm{d}\nu(y) \Bigg\} . \end{aligned} \tag{5}$$

Any solution of the dual problem above corresponds to a pair of Schrödinger potentials and vice versa (Nutz, 2021, Thm. 3.2). Note that $\mathrm{OT}_\varepsilon(\mu, \nu; c) = \varepsilon \cdot \mathrm{OT}_1(\mu, \nu; c/\varepsilon)$. 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 $\mathrm{OT}_1(\mu, \nu; c/\varepsilon)$.

2.2. Related work

Traditional analyses of the Sinkhorn algorithm view it as either (a) alternating projection on the two marginals μ, ν or (b) block maximization on the two dual potentials ψ, ϕ. These render a linear convergence with a contraction rate of the form $1 - e^{-\|c\|_\infty/\varepsilon}$. An important limitation of this analysis is that the rate becomes *exponentially slower* with growing $\|c\|_\infty$ or decreasing ε. More recent analyse have targeted the setting where $\mathcal{X}$ and $\mathcal{Y}$ are discrete spaces (Altschuler et al., 2017; Lin et al., 2022; Dvurechensky et al., 2018) and derive better rates in this setting.

More relevant to our work is the setting where $\mathcal{X}$ and $\mathcal{Y}$ are continuous spaces. Here, existing analyses have taken a probabilistic approach for analysing the Sinkhorn algorithm – a summary of these approach is that they place assumptions on the decay of the tails of μ, ν and / or log-concavity, or assume growth properties for the cost function. For instance, Chiarini et al. (2024) leverage the stability of optimal plans with respect to the marginals to obtain exponential convergence with unbounded cost for all $\varepsilon > 0$, albeit under various sets of conditions on the marginals. This relaxes the assumptions made in Chizat et al. (2024) for semi-concave bounded costs while still maintaining a contraction rate that only deteriorates *polynomially* in ε. We refer the reader to Chiarini et al. (2024, Sec. 1.5) for a more comprehensive literature review for analyses of the Sinkhorn algorithm.

In contrast, the advantage of taking the optimisation route i.e., viewing the Sinkhorn algorithm as performing infinite-dimensional mirror descent (Léger, 2021; Aubin-Frankowski et al., 2022;

Reza Karimi et al., 2024) is that it provides a guarantee under *minimal* assumptions. From non-asymptotic guarantee standpoint, these aforementioned works furnish a discrete-time iteration complexity that scales as $1/N\varepsilon$. If the costs are additionally assumed to be bounded, then Aubin-Frankowski et al. (2022) recover a contractive rate reminiscent of the classical Hilbert analysis. In this work, we achieve the similar rates while significantly expanding the scope of algorithm design and shed more light on the eOT problem, unifying both the primal and dual perspectives. Mensch and Peyré (2020) gives another mirror descent interpretation of Sinkhorn but the change of variable results in a non-convex objective which is hard to prove convergence for. There has also been interest in designing alternative algorithms for the eOT problem, among them Conforti et al. (2023) that designs Wasserstein gradient flow dynamics over the submanifold of $\Pi(\mu, \nu)$ which borrow tools from SDEs and PDEs in its analysis.

3. A new class of methods for solving the eOT problem

In this section, we introduce the class of methods Φ-match. We begin by first introducing this semi-dual problem in Section 3.1 and present this class of methods in Section 3.2.

3.1. The semi-dual problem in eOT

The semi-dual problem was originally discussed in Genevay et al. (2016), and later in more detail by Cuturi and Peyré (2018). To introduce the semi-dual, we first define the following operation (in the notation of Léger (2021)). Let $\phi \in L^1(\nu)$, define

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

This transformation is not unnatural; from eq. (4), we can see that a pair of Schrödinger potentials $\phi^\star, \psi^\star$ for $\mathrm{OT}_1(\mu, \nu; c/\varepsilon)$ satisfies $\psi^\star = (\phi^\star)^+$. Therefore, instead of solving the dual problem (eq. (5)) in two variables ϕ, ψ, it is sufficient to solve

$$\sup_{\phi \in L^1(\nu)} D(\phi^+, \phi) \,.$$

The above problem is referred to as the *semi-dual problem* associated with eq. (2). Additionally, we note that for any $\phi \in L^1(\nu)$, the objective J of the semi-dual satisfies

$$J(\phi) := D(\phi^+, \phi) = \sup_{\psi \in L^1(\mu)} D(\psi, \phi) = \int_{\mathcal{Y}} \phi(y) \mathrm{d}\nu(y) - \int_{\mathcal{X}} \phi^+(x) \mathrm{d}\mu(x) \,. \tag{6}$$

The fact that $D(\phi^+, \phi) = \sup_{\psi \in L^1(\mu)} D(\psi, \phi)$ can be derived by using the fact that $\psi \mapsto D(\psi, \phi)$ is concave and its first variation at ϕ^+ is 0. This leads to viewing the semi-dual problem for eOT as explicitly eliminating ψ via a partial maximisation of $D(\psi, \phi)$ in the dual problem (eq. (5)). To translate a dual potential $\phi \in L^1(\nu)$ to a joint distribution over $\mathcal{X} \times \mathcal{Y}$, define $\pi(\phi, \phi^+)$ with density w.r.t. $\mu \otimes \nu$ as

$$\mathrm{d}\pi(\phi, \phi^+)(x, y) = \exp\left(\phi(y) - \phi^+(x) - \frac{c(x,y)}{\varepsilon}\right) \mathrm{d}\nu(y)\mathrm{d}\mu(x) \,. \tag{7}$$

This is a valid probability density function over $\mathcal{X} \times \mathcal{Y}$ as evidenced by the fact that its $\mathcal{X}$-marginal is always μ. Recall that this joint distribution with $\phi \leftarrow \phi^\star$ in eq. (7) corresponds to the unique solution of the eOT problem (eq. (4)) by virtue of $(\phi^\star)^+ = \psi^\star$.

3.1.1. PROPERTIES OF THE SEMI-DUAL J

We henceforth assume that μ and ν have densities w.r.t. the Lebesgue measure. In this subsection, we state properties of the semi-dual J that underlie the methods that we propose and study in this section. As a prelude, we state two general observations about the semi-dual J.

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

Fact 2 (First variation) *The first variation δJ of the semi-dual J can be succinctly expressed in terms of the marginal $\pi(\phi, \phi^+)_\mathcal{Y}$ (Léger, 2021, Lem. 1): for any $\phi \in L^1(\nu)$,*

$$\delta J(\phi)(y) = \nu(y) - \pi(\phi, \phi^+)_\mathcal{Y}(y) = \nu(y) - \int_\mathcal{X} \exp\left(\phi(y) - \phi^+(x) - \frac{c(x,y)}{\varepsilon}\right) \nu(y)\mu(x)\mathrm{d}x\,. \quad (8)$$

The following two properties play a crucial role in our analysis. The first is that the Bregman divergence induced by J is non-positive – in other words J is a concave functional, and the second is a growth property of the semi-dual J. These are proven in Section D.1.

Lemma 1 *Let $\phi, \overline{\phi} \in L^1(\nu)$. Then, $J(\overline{\phi}) - J(\phi) - \langle \delta J(\phi), \overline{\phi} - \phi \rangle \leq 0$.*

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

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

Lemma 1 and Fact 2 implies that one could ostensibly use a gradient ascent-like procedure to find a maximiser of the semi-dual J and consequently solve the eOT problem to within a desired tolerance. More precisely, from an initialisation $\phi^0 \in L^1(\nu)$, we can obtain a sequence of iterates $\{\phi^n\}_{n \geq 1}$ based on the recursion

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

The update $\mathsf{M}^{\mathsf{SGA}}$ was previously considered by Genevay et al. (2016) 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 (Nesterov, 2018, Chap. 2) by the monotonicity of finite-dimensional norms. This consequently results in an *assumption-free* non-asymptotic convergence guarantee for SGA with $\eta < 2$. When generalising to continuous spaces, a temporary setback towards establishing such rates of convergence for this update is that $\|\phi\|_{L^\infty(\mathcal{Y})} \leq \|\phi\|_{L^2(\mathcal{Y})}$ is not generally true, thus leading to an "incompatibility". We rectify this by instead adopting an

alternate perspective on SGA as minimising the "discrepancy" between the $\mathcal{Y}$-marginal of $\pi(\phi, \phi^+)$ and ν. From the form of $\delta J(\phi)$ in Equation (8), we can rewrite the update SGA as

$$\mathsf{M}^{\mathsf{SGA}}(\phi; \eta) = \phi + \eta \cdot \left(\nu - \pi(\phi, \phi^+)_{\mathcal{Y}}\right) \ . \tag{9}$$

Note that a fixed point $\phi^\star$ of this update satisfies $\pi(\phi^\star, (\phi^\star)^+)_{\mathcal{Y}} = \nu$. This also corresponds to a maximiser of J since $\pi(\phi^\star, (\phi^\star)^+)_{\mathcal{Y}} = \nu$ which is equivalent to $\delta J(\phi^\star) = 0$. We draw a visual comparison to the dual-space version of the Sinkhorn algorithm (with a step size $\eta > 0$) (Reza Karimi et al., 2024) given by the following map

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

This leads us to propose the Φ-match class of methods that generalises beyond SGA and Sinkhorn, and corresponds to the following update

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

When $\Phi(f) : f \mapsto e^f$, this recovers SGA; and (2) when $\Phi(f) : f \mapsto f$, this recovers the dual-space interpretation of the Sinkhorn algorithm in Reza Karimi et al. (2024). The choice of the composition $\log \circ \Phi$ will become more apparent in the following section where we provide different interpretations of Φ-match in the primal space $\overline{\mathcal{Q}}$ to complement the dual form here.

4. Interpretations of Φ-match

The interpretations of Φ-match that we discuss here are motivated by recent work in understanding the Sinkhorn algorithm (Aubin-Frankowski et al., 2022; Reza Karimi et al., 2024). In essence, these prior works view the Sinkhorn algorithm as minimising $\mathrm{d}_{\mathrm{KL}}(\pi_{\mathcal{Y}}, \nu)$ over a subset $\mathcal{Q}$ of joint distributions over $\mathcal{X} \times \mathcal{Y}$. This is defined as

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

While $\pi^\star$ belongs to $\mathcal{Q}$, this constrained set is *not* a convex subset of $\mathcal{P}(\mathcal{X} \times \mathcal{Y})$ owing to the factorisation structure. Despite this, an intriguing observation about the Sinkhorn algorithm is that it ensures 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 instead a discrepancy which depends on Φ. We specifically show that Φ-match can be interpreted in the following two ways: (1) as an alternating projection scheme and (2) as a local greedy method analogous to gradient descent / mirror descent. While Φ-match is derived from the semi-dual, these interpretations do not involve the semi-dual and solely operate in $\mathcal{P}(\mathcal{X} \times \mathcal{Y})$. This leads to rates of convergence for SGA and its kernelised version (k-SGA) that we introduce later on.

Φ-match as iterative projections Let $\eta \in [0, 1]$, and define the following operations

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

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

For a given $\pi \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$, $\mathsf{project}_{\mathcal{Y},\nu}$ (eq. (11a)) can be seen as correcting the $\mathcal{Y}$-marginal of π towards ν, and the nature of this correction depends on Φ. On the other hand, $\mathsf{project}_{\mathcal{X},\mu}(\pi', \pi; \eta)$ finds a "midpoint" between π' and π, while fixing the $\mathcal{X}$-marginal to be μ. Note that neither π' nor π need to have $\mathcal{X}$-marginal as μ.

Φ-match as a local greedy method Consider the following operation

$$\mathsf{root}_{\mathcal{X},\mu}(\pi; \mathcal{F}, \eta) := \operatorname*{argmin}_{\bar{\pi}} \left\{ \langle \mathcal{F}(\pi), \bar{\pi} - \pi \rangle + \eta^{-1} \cdot \mathrm{d}_{\mathrm{KL}}(\bar{\pi} \| \pi) : \bar{\pi}_{\mathcal{X}} = \mu \right\} . \tag{12}$$

This is termed as local greedy method owing to the following observation: Suppose $\mathcal{F}$ is the first variation of a functional that measures the discrepancy between the $\mathcal{Y}$-marginal of π and ν, then $\mathsf{root}_{\mathcal{X},\mu}(\pi; \mathcal{F}, \eta)$ can be viewed as minimising a local first-order approximation of this functional, thereby approximately matching the $\mathcal{Y}$-marginal, while restricting the $\mathcal{X}$-marginal to be μ. Define the map $\mathcal{V}_\Phi$ as

$$\mathcal{V}_\Phi(\pi)(x, y) = \log \Phi(\pi_{\mathcal{Y}})(y) - \log \Phi(\nu)(y) . \tag{13}$$

While the iterative projections and the local greedy method may appear disparate, the following theorem shows how Φ-match are related to both the iterative projections eqs. (11a) and (11b), and the local greedy method $\mathsf{root}_{\mathcal{X},\mu}$ when $\mathcal{F} \leftarrow \mathcal{V}_\Phi$.

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

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

We give the proof of Theorem 3 in Section D.3.1. This theorem generalises Reza Karimi et al. (2024, Lem. 1), and also recovers the classical iterative Bregman projection interpretation of the Sinkhorn method (Peyré and Cuturi, 2019, Remark 4.8) by setting $\Phi \leftarrow \mathrm{Id}$ and $\eta \leftarrow 1$. The map $\mathcal{V}_\Phi$ in eq. (13) is $\log \frac{\pi_{\mathcal{Y}}}{\nu}$ in Sinkhorn, which corresponds to the first variation of the functional $\rho \mapsto \mathrm{d}_{\mathrm{KL}}(\rho_{\mathcal{Y}} \| \nu)$. The connection to eq. (12) in this setting was originally derived in Aubin-Frankowski et al. (2022, Prop. 5) and generalised to an arbitrary step size $\eta \in (0, 1)$ in Reza Karimi et al. (2024, Lem. 1). Instead of $\Phi \leftarrow \mathrm{Id}$, if one were to consider Φ as $\Phi(\rho) = \exp\left(f'\left(\frac{\rho}{\nu}\right)\right)$, then eq. (13) can be viewed the first variation of a f-divergence $\rho \mapsto D_f(\rho \| \nu)$.

4.1. Connection between SGA and Maximum Mean Discrepancy

The Φ-match abstraction also permits a similar interpretation of SGA, which corresponds to $\Phi : f \mapsto e^f$ in Φ-match. The map $\mathcal{V}_\Phi$ (eq. (13)) in this case is $\pi_{\mathcal{Y}} - \nu$, which is the first variation of another notion of discrepancy between $\pi_{\mathcal{Y}}$ and ν given by a *Maximum Mean Discrepancy* (MMD) (Gretton et al., 2006); we provide a brief overview of this in Section A.1. A crucial fact that leads to the connection between SGA and the MMD is that the first variation of $\xi \mapsto \frac{1}{2}\mathrm{MMD}_k(\xi, \rho)^2$ is given by $\mathfrak{m}_k(\xi) - \mathfrak{m}_k(\rho)$ (Mroueh et al., 2019, Lem. 1) where $\mathfrak{m}_k(\xi)(y) = \int k(y, y') \mathrm{d}\xi(y')$, and it can be immediately seen that the map $\mathcal{V}_\Phi$ for SGA coincides with the first variation of

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

for the identity kernel k_{Id} defined as $k_{\mathrm{Id}}(y, y') = 1$ iff $y = y'$, since $\mathfrak{m}_k(\xi)(y) = \xi(y)$. This also motivates the consideration of a more general update $\mathsf{M}^{k\text{-SGA}}$ for iteratively minimising $\mathcal{L}_k(\cdot; \nu)$ for any characteristic kernel, and this results in the following update:

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

Due to the generality of the abstraction Φ-match, we can also view k-SGA as an instance of Φ-match with the choice $\Phi_k(f) : f \mapsto e^{\mathfrak{m}_k(f)}$ where we overload $\mathfrak{m}_k(f) := \int_{\mathcal{Y}} k(y, y') \cdot f(y') \mathrm{d}y'$. Theorem 3 shows that the sequence of iterates $\{\phi^n\}_{n \geq 1}$ formed by k-SGA results in a sequence of distributions $\{\pi(\phi^n, (\phi^n)^+)\}_{n \geq 1}$ formed by iteratively applying $\mathsf{root}_{\mathcal{X},\mu}$ with $\mathcal{F} \leftarrow \mathcal{V}_{\Phi_k}$. This implication, along with properties of $\mathcal{L}_k$ for bounded, positive definite (PD) kernels, enables us to derive non-asymptotic rates of convergence for k-SGA as stated in the theorem below.

Theorem 4 *Let $\phi^0 \in L^1(\nu)$, and consider a PD kernel k such that $k(y, y) \leq c_k$ for all $y \in \mathcal{Y}$. Define $\{\phi^n\}_{n \geq 1}$ to be the sequence of potentials obtained as $\phi^{n+1} = \mathsf{M}^{k\text{-SGA}}\left(\phi^n; \min\left\{\frac{1}{2c_k}, 1\right\}\right)$ and their corresponding sequence of distributions $\{\pi^n\}_{n \geq 1}$ where $\pi^n = \pi(\phi^n, (\phi^n)^+)$. Then, for any $N \geq 1$*

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

The proof of the above theorem is given in Section D.3.2. Note that every π^n in the sequence of distributions generated by k-SGA stays in $\mathcal{Q}$ by construction. When the kernel k is characteristic, Theorem 4 shows that $\pi^n_{\mathcal{Y}}$ approaches ν and consequently establishes a rate of convergence to the optimal coupling $\pi^\star$. Suppose $\phi^0 = \mathbf{0}$, and $\pi^0 = \pi(\phi^0, (\phi^0)^+)$, then from Léger (2021, proof of Cor. 1), we know that $\mathrm{d}_{\mathrm{KL}}(\pi^\star \| \pi^0) \leq \mathrm{d}_{\mathrm{KL}}(\pi^\star \| \pi^{\mathsf{ref}})$ where π^{ref} is the reference measure in eq. (3). Consequently, after N iterations, the rate we obtain from Theorem 4 is

$$\frac{2c_k}{N} \cdot \frac{\mathrm{d}_{\mathrm{KL}}(\pi^\star \| \pi^{\mathsf{ref}})}{\varepsilon} .$$

This highlights the better dependence on ε compared to the more classical analyses of the Sinkhorn algorithm where the dependence on ε is of the form $1 - e^{-\varepsilon^{-1}}$. This theorem also results in the *first known assumption-free guarantees* for SGA by setting $c_k = 1$, since the bound is meaningful as soon as an optimal coupling exists $\mathrm{d}_{\mathrm{KL}}(\pi^\star \| \pi^{\mathsf{ref}}) < \infty$.

5. Extensions

Here, we build on the ideas in the preceding sections and present two extensions. The first extension revisits the smoothness property of the semi-dual discussed previously and illustrates how it can be leveraged to design iterative methods for maximising the semi-dual with provable guarantees, which also allows us to design an accelerated version of one of these algorithms without having to place assumptions on μ and ν. The second extension adapts Φ-match for the dynamical Schrödinger bridge problem which generalises the static Schrödinger bridge problem to path measures.

5.1. Adapting to smoothness of the semi-dual J

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

Lemma 5 *Let $\phi, \overline{\phi} \in L^2(\nu)$ be such that $\|\phi\|_{L^\infty(\mathcal{Y})}, \|\overline{\phi}\|_{L^\infty(\mathcal{Y})} \leq B$ for a given $B > 0$. Assume that the cost $c(\cdot,\cdot) \geq 0$. Then,*

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

The domain $\mathcal{S}_B = \{\phi : \|\phi\|_{L^\infty(\mathcal{Y})} \leq B\}$ considered in the lemma above is of particular interest when the cost function c is uniformly bounded; a result from Di Marino and Gerolin (2020) states that the Schrödinger potentials belong in such $\mathcal{S}_B$ where B depends on the bound on c.

Here, we leverage the smoothness properties established for the semi-dual in Lemmas 2 and 5 to propose two additional methods – sign-SGA and proj-SGA – for the eOT problem that conceptually differ from the Φ-match family. sign-SGA is based on the $L^\infty(\mathcal{Y})$-smoothness of the semi-dual, while proj-SGA is based on the $L^2(\nu)$-smoothness of the semi-dual over $\mathcal{S}_B$. For sign-SGA, y_{anc} is an arbitrary point in $\mathcal{Y}$ termed an "anchor point", and the "recentering" second step in sign-SGA ensures that iterates lie within a bounded superlevel set of J.

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

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

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

Theorem 6 (Informal) *Let $\phi^0 \in L^1(\nu) \cap L^\infty(\mathcal{Y})$, $y_{\mathrm{anc}} \in \mathcal{Y}$. Consider the sequence of potentials $\{\phi^n\}_{n \geq 1}$ generated according to sign-SGA with $\eta = 1$. Then, there exists $C > 0$ depending on ϕ^0, y_{anc} such that for all $N \geq 1$,*

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

Theorem 7 (Informal) *Suppose $c(\cdot,\cdot)$ is a non-negative cost function such that $\lambda(B) < \infty$, and $\phi^0 \in \mathcal{S}_B$. Consider the sequence of potentials $\{\phi^n\}_{n \geq 1}$ generated according to proj-SGA with*

$\eta = \lambda(B)^{-1}$. *Then, there exists* $C > 0$ *depending on* ϕ^0 *such that for all* $N \geq 1$,

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

Note that the growth condition in Lemma 5 that proj-SGA is conceptually based on is given in terms of $L^2(\nu)$ which is a Hilbert space. This inspires us to adopt the structure of FISTA (Beck and Teboulle, 2009) to design an accelerated version of proj-SGA, which leads to a rate of convergence that scales as $1/N^2$. We propose the following accelerated version of proj-SGA based on FISTA to handle the non-uniform growth condition in Lemma 5: for initial values $\phi^1 = \overline{\phi}^0 \in \mathcal{S}_B$, and $t_1 = 1$,

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

Theorem 8 (Informal) *Consider the setting of Theorem 7, and let* $\{\phi^n\}_{n \geq 2}, \{\overline{\phi}^n\}_{n \geq 1}$ *be generated according to* proj-SGA++. *Then, there exists* $C > 0$ *depending on* ϕ^0 *such that for all* $N \geq 1$,

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

When B is sufficiently large (for bounded costs as suggested in (Di Marino and Gerolin, 2020)), the minimiser of J over bounded subset $\mathcal{S}_B$ coincides with $\phi^\star$, which implies that $\min_{\phi \in \mathcal{S}_B} J(\phi) = J(\phi^\star)$. The more detailed versions of Theorems 6 to 8 are given in Appendix B with their proofs. We conclude this discussion by drawing a comparison to *how* the rates for SGA, sign-SGA, and proj-SGA have been established. Specifically, the techniques used to prove Theorems 6 to 8 are based on more standard optimisation arguments (generalised to ∞-dimensional spaces), and are fundamentally different from Theorem 4 as they operate directly on the dual space of potentials. This highlights the benefit of considering alternate viewpoints for establishing provable guarantees.

The quantity $\lambda(B)$ This appears in Theorems 7 and 8, and unfortunately cannot generally be bounded by a uniform constant. However, there are setting where this can be done with relative ease. For instance, in the setting where the cost is bounded as in Di Marino and Gerolin (2020) (which also ensures regularity of the dual potentials), this quantity can be bounded as $e^{B(2+\varepsilon^{-1})}$. Alternatively, when $c(x, y') = \frac{1}{2}\|x - y'\|^2$, this quantity can be bounded in the regime where μ and ν have sub-Gaussian tails. More broadly speaking, one can obtain bounds on $\lambda(B)$ depending on the interplay between properties of the cost function and the marginal distributions μ and ν.

5.2. Path-space Φ-match for the Schrödinger bridge problem

The eOT problem in the static Schrödinger bridge form (Equation (3)) can be viewed as finding a certain distribution in $\mathcal{P}(\mathcal{X} \times \mathcal{Y})$ closest to another reference distribution in $\mathcal{P}(\mathcal{X} \times \mathcal{Y})$. This can also be posed for distributions over curves $\mathcal{C}([0, T]; \mathcal{Z})$ for $\mathcal{Z} \subseteq \mathbb{R}^d$ more generally, and a path measure precisely captures the notion of a probability measure over trajectories. For a stochastic process

$(X_t)_{t\in[0,T]}$ with state space $\mathcal{Z} \subseteq \mathbb{R}^d$, the path measure is a collection of distributions $(\mathbb{P}_t)_{t\in[0,T]}$ where $\mathbb{P}_t$ is the law of X_t. The dynamical Schrödinger bridge problem is formulated as finding the path measure with endpoints $\mathbb{P}_0 = \mu$ and $\mathbb{P}_T = \nu$ that is closest to the reference path measure:

$$\min_{\mathbb{P}} \ \mathrm{d}_{\mathrm{KL}}(\mathbb{P}\|\mathbb{P}^{\mathsf{ref}}) \quad \text{such that } \mathbb{P}_0 = \mu, \mathbb{P}_T = \nu \ . \tag{14}$$

This is a (strictly) convex problem defined in $\mathcal{P}(\mathcal{C}([0,T],\mathcal{Z}))$. We defer a more detailed discussion about the Schrödinger Bridge problem to Section A.2 and focus on algorithms for solving Equation (14) based on Φ-match here.

A popular algorithm for solving the dynamic Schrödinger bridge problem is referred to as the *Iterative Proportional Fitting* (IPF) algorithm given by the iteration below.

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

This can be viewed as the path-space analogue of the alternating projection form of the Sinkhorn algorithm from (11a)-(11b). In Reza Karimi et al. (2024, Prop. 2), it is shown that this iteration is the solution to the following local greedy update:

$$\mathbb{P}^{n+1} = \operatorname{argmin} \left\{\langle \delta \mathrm{d}_{\mathrm{KL}}(\mathbb{P}^n_T\|\nu), \overline{\mathbb{P}} - \mathbb{P}^n\rangle_{L^2(\mathcal{C}([0,T],\mathcal{Z}))} + \mathrm{d}_{\mathrm{KL}}(\overline{\mathbb{P}}\|\mathbb{P}^n) \ : \ \overline{\mathbb{P}}_0 = \mu\right\} \ .$$

From IPF to path-Φ-match The similarities between the local greedy update above and Equation (12) motivate us to propose a path-space analogue of Φ-match. For $\eta \in [0,1]$, define

$$\mathbb{P}^{n+1} = \operatorname*{argmin}_{\overline{\mathbb{P}}} \ \left\{\langle \overline{\mathcal{V}}_\Phi(\mathbb{P}^n), \ \overline{\mathbb{P}} - \mathbb{P}^n\rangle + \eta^{-1} \cdot \mathrm{d}_{\mathrm{KL}}(\overline{\mathbb{P}}\|\mathbb{P}^n) \ : \ \overline{\mathbb{P}}_0 = \mu\right\} \tag{15}$$

where $\overline{\mathcal{V}}_\Phi(\mathbb{P}) = \log\Phi(\mathbb{P}_T) - \log\Phi(\nu)$. When $\Phi(f) = f$, $\overline{\mathcal{V}}_\Phi$ is the first variation of $\rho \mapsto \mathrm{d}_{\mathrm{KL}}(\rho_T\|\nu)$, and Equation (15) recovers the interpretation of IPF stated above with $\eta = 1$. This local greedy update can be expressed a sequence of alternating projections but in the space of path measures as written below

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

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

As it turns out, path-Φ-match (and therefore path measure update Equation (15)) can be implemented as updates on the drifts of a sequence of SDEs whose corresponding solutions are given by $\{\mathbb{P}^n\}_{n\geq 1}$. We formally characterise this recursive update on $\{v^n_t\}_{n\geq 1, t\in[0,T]}$ in the following proposition that extends Reza Karimi et al. (2024, Thm. 4.2) for IPF to path-Φ-match.

Proposition 9 *Let $\{\mathbb{P}^n\}_{n\geq 1}$ be obtained from path-Φ-match with $\mathbb{P}^0 = \mathbb{P}^{\mathsf{ref}}$. Then for every $n \geq 1$, $\mathbb{P}^n$ is the path measure associated with the solution of the following SDE:*

$$\mathrm{d}X_t = v^n_t(X_t)\mathrm{d}t + \sqrt{2}\cdot \mathrm{d}B_t \ , \quad X_0 \sim \mu \ ,$$

where $v^0_t = u^{\mathsf{ref}}$ and for every $n \geq 0$,

$$v^{n+1}_t = v^n_t + 2\eta \cdot (\nabla \log p^{n+1/2}_t - \nabla \log p^n_t) - 2\cdot \nabla V_t$$

with $p_t^{n+1/2}$, p_t^n denoting marginal densities of $\mathbb{P}^{n+1/2}$ and $\mathbb{P}^n$ respectively. Above V_t is defined as

$$V_t(x) := -\log \mathbb{E}\left[\exp\left(-\eta\cdot(1-\eta)\cdot\int_t^T \|\nabla \log p_s^{n+1/2}(Z_s) - \nabla \log p_s^n(Z_s)\|^2 \mathrm{d}s\right)\,\middle|\, Z_t = x\right]$$

where the expectation is taken over the law of the SDE $(Z_s)_{s\geq t}$ given by

$$\mathrm{d}Z_s = [v_s^n(Z_s) + 2\eta\cdot(\nabla \log p_s^{n+1/2}(Z_s) - \nabla \log p_s^n(Z_s))]\,\mathrm{d}s + \sqrt{2}\cdot \mathrm{d}B_s\ ;\ Z_t = x\ .$$

The key difference here is that the path measure $\mathbb{P}^{n+1/2}$ is defined by the time-reversal of the SDE with respect to $\mathbb{P}^n$, but with initial distribution whose density is proportional to $\mathbb{P}_T^n \cdot \frac{\Phi(\nu)}{\Phi(\mathbb{P}_T^n)}$, instead of ν as for the case of η-IPF. Reza Karimi et al. (2024, Sec. C.2) discuss approaches for computing ∇V_t through stochastic optimal control, and the other terms $\nabla \log p_t^{n+1/2}$, $\nabla \log p_t^n$ can be obtained through score-matching time-reversals as in the classical IPF algorithm where $\eta = 1$. We give details of the proof of this proposition in Section D.5.1, and additional perspectives on the connection between path-Φ-match and the dynamical Schrödinger bridge problem in Appendix C.

6. Conclusion

In this work, we systemically synthesise a variety of viewpoints on algorithms for eOT – specifically those surrounding the Sinkhorn algorithm. This synthesis, centered around infinite-dimensional optimisation, leads to a collection of novel methods based on 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 seamlessly, which notably do not rely on any strict assumptions on the marginals μ, ν. We believe exploiting the regularity conditions of the various formulations of the eOT problem established in this paper, this novel perspective opens up new avenues for designing algorithms and provide a way to rethink principled approaches to the eOT / Schrödinger bridge problem.

Acknowledgments

The authors would like to thank the reviewers at ALT 2026 for their valuable feedback, and Andre Wibisono and Ashia Wilson for their involvement during earlier stages of this work. VS would like to thank Pierre-Cyril Aubin for his comments on an earlier draft of this work.

References

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

Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient flows in metric spaces and in the space of probability measures*. Lectures in Mathematics ETH Zürich. Birkhäuser Verlag, Basel, second edition, 2008.

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

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.

Espen Bernton, Jeremy Heng, Arnaud Doucet, and Pierre E Jacob. Schrödinger Bridge Samplers. *arXiv preprint arXiv:1912.13170*, 2019.

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, 2022.

Guillaume Carlier. On the linear convergence of the multimarginal sinkhorn algorithm. *SIAM Journal on Optimization*, 32(2):786–794, 2022.

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.

Lénaïc Chizat, Pierre Roussillon, Flavien Léger, François-Xavier Vialard, and Gabriel Peyré. Faster Wasserstein distance estimation with the Sinkhorn divergence. In *Advances in Neural Information Processing Systems*, volume 33, pages 2257–2269. Curran Associates, Inc., 2020.

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

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

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

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

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

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.

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

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

Jean Feydy, Thibault Séjourné, François-Xavier Vialard, Shun-ichi Amari, Alain Trouve, and Gabriel Peyré. Interpolating between Optimal Transport and MMD using sinkhorn divergences. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 2681–2690. PMLR, 2019.

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.

Joel Franklin and Jens Lorenz. On the scaling of multidimensional matrices. *Linear Algebra and its Applications*, 114-115:717–735, 1989.

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.

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

Aude Genevay, Lénaïc Chizat, Francis Bach, Marco Cuturi, and Gabriel Peyré. Sample Complexity of Sinkhorn Divergences. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 1574–1583. PMLR, 2019.

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

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

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

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.

Gonzalo Mena and Jonathan Niles-Weed. Statistical bounds for entropic optimal transport: sample complexity and the central limit theorem. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

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

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

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

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

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

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

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.

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

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.

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

Ludger Rüschendorf. Convergence of the iterative proportional fitting procedure. *The Annals of Statistics*, pages 1160–1174, 1995.

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

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

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

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

Appendix A. Additional background

A.1. Maximum Mean Discrepancy (MMD)

Let $k : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$ be a positive-definite kernel, and let $\mathcal{H}_k$ be its RKHS. We refer the reader to Steinwart and Christmann (2008, Chap. 4) for a more detailed exposition about kernels and their RKHS. Recall that the mean function of $\xi \in \mathcal{P}(\mathcal{Z})$ w.r.t. kernel k is defined as

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

The MMD (for a kernel k) (Gretton et al., 2006) between two distributions measures the discrepancy between the mean function of two distributions, and is formally defined as

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

For a *characteristic* kernel (Fukumizu et al., 2004), the map $\xi \mapsto \mathfrak{m}_k(\xi)$ is a one-to-one mapping. This implies that the MMD defined by a characteristic kernel is a metric over the space of probability measures. Examples of characteristic kernels are the identity, Gaussian, and Laplace kernels.

A.2. Dynamical Schrödinger bridge problem

The classical result of Föllmer (1988) shows that by the disintegration property of the KL divergence, the optimal path measure $\mathbb{P}^\star$ that solves eq. (14) can be decoupled as

$$\mathbb{P}^\star_{\{0,T\}} = \pi^\star \; ; \quad \mathbb{P}^\star_{(0,T)|X_0,X_T} = \mathbb{P}^{\mathsf{ref}}_{(0,T)|X_0,X_T}$$

where $\pi^\star$ is the solution to the (primal) eOT problem in the static Schrödinger bridge form (eq. (3)) with $\pi^{\mathsf{ref}} = \mathbb{P}^{\mathsf{ref}}_{0,T}$, and $\mathbb{P}_{(0,T)|X_0,X_T}$ is the terminal-conditioned measure (also termed the bridge). An example of a reversible Markov process $\mathbb{P}^{\mathsf{ref}}$ is the reversible Brownian motion on $[0, T]$. In this case, the density of $\mathbb{P}^{\mathsf{ref}}_{\{0,T\}}$ satisfies $\mathbb{P}^{\mathsf{ref}}_{\{0,T\}}(x, y) \propto \exp\left(-\frac{\|x-y\|^2}{2T}\right)$. Finding $\pi^\star$ is equivalent to solving the eOT problem (eq. (2)) with cost $c(x, y) = \frac{\|x-y\|^2}{2}$ and $\varepsilon = T$, and the bridge in this setting corresponds to the classical Brownian bridge which draws a more substantive connection between eOT and the dynamic SB problem. However, in general, it is not feasible to directly sample from the bridge, and this motivates the design of generic methods for solving eq. (14).

Due to the above disintegration property and the form of the solution of the static Schrödinger bridge problem in eq. (4), under certain regularity conditions (see Léonard (2014, Thms. 2.8 and 2.9)), there exists measurable functions $\boldsymbol{\psi}_0$ and $\boldsymbol{\phi}_T$ such that the relative density of $\mathbb{P}^\star$ w.r.t. $\mathbb{P}^{\mathsf{ref}}$ satisfies

$$\frac{\mathrm{d}\mathbb{P}^\star_{[0,T]}}{\mathrm{d}\mathbb{P}^{\mathsf{ref}}_{[0,T]}}((X_t)_{t\in[0,T]}) = \exp\left(\boldsymbol{\phi}_T(X_T) - \boldsymbol{\psi}_0(X_0)\right) \; . \tag{17}$$

Moreover, if $\mathbb{P}^{\mathsf{ref}}$ is Markovian i.e., if the underlying stochastic process is a Markov process, then the factorisation in eq. (17) is necessary and sufficient for the characterisation of $\mathbb{P}^\star$ and is unique. This parallels the unique form of the optimal coupling form in eq. (4).

Let $\mathbb{P}^{\text{ref}}$ be the strong solution of an SDE (hence Markovian) as

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

By Girsanov's theorem (Pavliotis, 2014, Chap. 3), the dynamical SB problem can be reduced to a minimisation problem over path measures $\mathbb{P}$ induced by SDEs with a different drift vector field $(v_t)_{t\geq 0}$

$$\mathrm{d}X_t = v_t(X_t)\mathrm{d}t + \sqrt{2}\cdot \mathrm{d}B_t\,, \quad X_0 \sim \mu\,. \tag{19}$$

For the IPF algorithm described in Section 5.2, Rüschendorf (1995) shows that the initial and final marginals of the sequence of iterates $\{\mathbb{P}^n\}_{n\geq 1}$ generated by the above iteration with $\mathbb{P}^0 = \mathbb{P}^{\text{ref}}$ converges to μ and ν respectively as $n\to\infty$, and Bernton et al. (2019) gives a non-asymptotic rate of convergence that scales as $\frac{1}{N}$ in the number of iterations N building off the result by Rüschendorf (1995).

Appendix B. Details about the steepest ascent methods in Section 5.1

B.1. Signed semi-dual gradient ascent

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

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

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

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

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

Theorem 10 (Formal version of Theorem 6) *Let $\{\phi^n\}_{n\geq 1}$ be the sequence of potentials generated according to the following recursion for $n\geq 0$:*

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

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

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

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

B.2. Projected semi-dual gradient ascent

Here we consider proj-SGA and its accelerated variant proj-SGA++. As mentioned briefly previously, when the cost function is bounded in a certain manner, it is possible to show that the semi-dual satisfies a different notion of smoothness, but non-uniformly depending on the "size" of the domain considered. While Lemma 2 is a general statement, regularity condition Lemma 5 is parameterised by a "size" parameter B, and hence it is useful to understand what a reasonable choice of B is for the purposes of solving the eOT problem. If B is too small, then it is likely that the Schrödinger potential $\phi^\star$ would not satisfy $\|\phi^\star\|_{L^\infty(\mathcal{Y})} \leq B$. Interestingly however, the Schrödinger potentials $\phi^\star$ and $\psi^\star = (\phi^\star)^+$ inherit properties from the cost function $c(\cdot,\cdot)$, which allow us to determine a reasonable choice of B based on the cost function. This is formalised in the following proposition.

Proposition 11 ((Di Marino and Gerolin, 2020, Lem. 2.7)) *Consider the dual eOT problem defined in Equation (5). There exists Schrödinger potentials $\phi^\star, \psi^\star$ such that*

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

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

While proj-SGA may appear fortuitous, this is actually a natural recommendation based on Lemma 5. This is because it can be obtained as the solution to a truncated local quadratic approximation of the semi-dual given below (truncated due to the restriction to $\mathcal{S}_B$), which is inspired by ISTA (Beck and Teboulle, 2009):

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

From Lemma 5, when the cost $c(\cdot,\cdot)$ is non-negative and the step size satisfies $\eta \leq \lambda(B)^{-1}$, $J(\mathsf{M}^{\mathsf{proj-SGA}}_{\mathcal{S}_B}(\phi;\eta)) \geq J(\phi)$ for any $\phi \in \mathcal{S}_B$ as

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

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

Theorem 12 (Formal version of Theorem 7) *Suppose $c(\cdot,\cdot)$ is a non-negative cost function such that $\lambda(B) < \infty$. Let $\{\phi^n\}_{n\geq 1}$ be the sequence of potentials generated according to the following recursion for $n \geq 0$:*

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

where $\lambda(B)$ is defined in Lemma 5. Then, for all $N \geq 1$, $\phi^N \in \mathcal{S}_B$ and for $\widetilde{\phi}^\star = \operatorname{argmax}_{\phi\in\mathcal{S}_B} J(\phi)$

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

Theorem 13 (Formal version of Theorem 8) *Suppose $c(\cdot,\cdot)$ is a non-negative cost function such that $\lambda(B) < \infty$. Consider the sequences $\{\phi^n\}_{n\geq 2}, \{\overline{\phi}^n\}_{n\geq 1}$ generated according to proj-SGA++. Then, for any $N \geq 1$,*

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

From Proposition 11, we know that the maximum value of the semi-dual J over $\mathcal{S}_B$ for $B = \frac{3\|c\|_{L^\infty(\mathcal{X}\times\mathcal{Y})}}{2}$ is $J(\phi^\star)$ for the Schrödinger potential $\phi^\star$. This implies that the sequences of semi-dual values generated by proj-SGA and proj-SGA++ with the appropriate step sizes converge to $J(\phi^\star)$ at a $\frac{1}{N}$ and $\frac{1}{N^2}$ rate respectively. A crude bound on $\lambda(B)$ in this setting is given by $\lambda(B) \leq \exp\big(B\cdot(\varepsilon^{-1}+2)\big)$.

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

Appendix C. Further connections between Φ-match and the Schrödinger bridge problem

The connection to Φ-match can also be seen in the following manner. Suppose $\mathbb{P}^n$ satisfies the factorisation form in eq. (17) with some $\boldsymbol{\phi}^n_T$ and $\boldsymbol{\psi}^n_0$, and note that $\mathbb{P}^n_0 = \mu$. Then, one can show analogous to lemma 16 that $\mathbb{P}^{n+1}$ admits the factorisation form in eq. (17) where

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

and $\boldsymbol{\psi}_0^{n+1}$ set to satisfy $\mathbb{P}_0^{n+1} = \mu$ is equivalent to the output of path-Φ-match. From this, it can be seen that for each $n \geq 0$, $\mathbb{P}_{\{0,T\}}^{n+1}$ is the solution for the static Schrödinger bridge problem with marginals $(\mu, \mathbb{P}_T^n)$, and $\mathbb{P}_{(0,T)|X_0,X_T}^{n+1} = \mathbb{P}_{(0,T)|X_0,X_T}^{\mathsf{ref}}$. As a consequence, if $\mathbb{P}_T^n \to \nu$ then $\mathbb{P}^n \to \mathbb{P}^\star$. We expand on the value of this factorisation in the following discussion about a potential-space implementation of path-Φ-match.

A "dual" version of path-Φ-match We investigate here a "dual" potential-space implementation which is based on the factorisation form in eq. (17) in contrast to the discussion above which was based on the path measures. This connects to updates from Φ-match for eOT, and is similar to the continuous flow of SDE viewpoint in Reza Karimi et al. (2024, Sec. 4.4). Let $\phi^\star$ and $\psi^\star$ be the Schrödinger potentials associated with the static problem (eq. (3)) with marginals μ, ν, and suppose $(f_t^\star)_{t\geq 0}$ and $(g_t^\star)_{t\geq 0}$ are solutions to the Kolmogorov forward and backward equations under eq. (18):

$$\begin{aligned} \partial_t f_t^\star + \nabla \cdot (u^{\mathsf{ref}} f_t^\star) - \Delta f_t^\star = 0\,, \quad f_0^\star(x) = e^{\psi^\star(x)}\,, \\ \partial_t g_t^\star + u^{\mathsf{ref}\top} \nabla g_t^\star + \Delta g_t^\star = 0\,, \quad g_T^\star(y) = e^{\phi^\star(y)}\,. \end{aligned}$$

Then, according to Léonard (2014, Thm. 3.4) and the Markov property of $\mathbb{P}^{\mathsf{ref}}$, we have

$$\frac{\mathrm{d}\mathbb{P}^\star_{[k,l]}}{\mathrm{d}\mathbb{P}^{\mathsf{ref}}_{[k,l]}}((X_t)_{t\in[k,l]}) = f_k^\star(X_k) g_l^\star(X_l) \quad \forall\, k < l \in [0,T]\,.$$

Therefore with their dynamics fixed, knowing the two boundary functions $\boldsymbol{\phi}_T^\star, \boldsymbol{\psi}_0^\star$ (eq. (17)) 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 $\boldsymbol{\phi}_T^n, \boldsymbol{\psi}_0^n$ alone:

$$\frac{\mathrm{d}\mathbb{P}^n_{[0,T]}}{\mathrm{d}\mathbb{P}^{\mathsf{ref}}_{[0,T]}}((X_t^n)_{t\in[0,T]}) = \frac{\mathbb{P}^n_{0,T}}{\mathbb{P}^{\mathsf{ref}}_{0,T}}(X_0^n, X_T^n) = \exp(\boldsymbol{\phi}_T^n(X_T^n) + \boldsymbol{\psi}_0^n(X_0^n))\,. \tag{20}$$

This implies that they always solve the SB problem for their own marginals, and consequently remain within the reciprocal and Markovian class. If moreover, $\boldsymbol{\psi}_0^n = (\boldsymbol{\phi}_T^n)^+$ then we always have $\mathbb{P}_0^n = \mu$. In the lemma below, we derive alternative updates on the SDE drift by leverging the corresponding dual Schrödinger potentials from the static setting, whose proof can be found in section D.5.2.

Lemma 14 *Let $\{\phi^n\}_{n\geq 1}$ be the sequence of potentials obtained from Φ-match. For any $n \geq 1$, let $(g_t^n)_{t\in[0,T]}$ be the solution to the Kolmogorov backward equation*

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

The path measure $\mathbb{P}^n$ associated with the following SDE

$$\mathrm{d}X_t^n = \left\{ u^{\mathsf{ref}}(X_t^n) + \nabla \log g_t^n(X_t^n) \right\} \mathrm{d}t + \sqrt{2} \cdot \mathrm{d}B_t\,, \quad X_0^n \sim \mu \tag{21}$$

factorises as eq. (20) 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. (21) implies that v_t in eq. (19) is necessarily a gradient vector field. Through a Cole-Hopf transform, the Kolmogorov backward equation can be translated to a PDE in $\phi_t^n = \log g_t^n$ instead. This is also evident from the Feynman-Kac formula, which prescribes that $(\phi_t^n)_{t\in[0,T]}$ can be expressed as $\phi_t^n(y) = \log \mathbb{E}^{\mathsf{ref}}[e^{\phi^n(x_T)}|x_t = y]$ with respect to the the reference process eq. (18).

Appendix D. Proofs

D.1. Proofs for the properties of the semi-dual J

In this subsection, we give the proofs of Lemmas 1, 2 and 5. These lemmas are corollaries of the following lemma whose proof is given in Section D.2. For a distribution ξ and a suitably integrable function f, we use $\mathbb{V}_\xi[f]$ to denote the variance of $f(Z)$ where $Z \sim \xi$.

Lemma 15 *Let $\phi, \overline{\phi} \in L^1(\nu)$. For $t \in [0,1]$, define $\tilde{\phi}_t := \phi + t \cdot (\overline{\phi} - \phi)$ and the conditional distribution $\rho_t(.;x)$ whose density is*

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

Then,

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

Proof of Lemma 1 From Lemma 15, and the non-negativity of the variance, we have for any $\phi, \overline{\phi} \in L^1(\nu)$ that

$$J(\overline{\phi}) - J(\phi) - \langle \delta J(\phi), \overline{\phi} - \phi\rangle = -\frac{1}{2}\int_0^1 \mathbb{E}_{x\sim\mu}\left[\mathbb{V}_{\rho_t(.;x)}[\overline{\phi} - \phi]\right] \mathrm{d}t \leq 0 .$$

■

Proof of Lemma 2 By the definition of the variance, we have for $\phi, \overline{\phi} \in L^1(\nu)$ that

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

where the final inequality is due to Hölder's inequality. Instantiating Lemma 15, we get

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

■

Proof of Lemma 5 From the convexity of $\mathcal{S}_B$, note that $\tilde{\phi}_t = \phi + t \cdot (\overline{\phi} - \phi) \in \mathcal{S}_B$. Since $\mathcal{S}_B \subset L^1(\nu)$, we have by Lemma 15 that

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

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

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

Step (a) uses Jensen's inequality, and step (b) uses the fact that for y, y', $\tilde{\phi}_t(y) - \tilde{\phi}_t(y') \leq 2B$ for y, y' almost everywhere. Substituting this in the result of Lemma 15, we have

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

■

D.2. Proof of Lemma 15

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

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

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

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

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

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

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

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

Consequently,

$$\begin{aligned}
\langle j_x(\phi) - j_x(\overline{\phi}), \phi - \overline{\phi}\rangle &= \int_{\mathcal{Y}} (\phi(y) - \overline{\phi}(y)) \cdot (\rho_1(y;x) - \rho_0(y;x)) \, \mathrm{d}y \\
&= \int_{\mathcal{Y}} \int_0^1 (\phi(y) - \overline{\phi}(y)) \cdot \dot{\rho}_t(y;x) \, \mathrm{d}t\mathrm{d}y \\
&= -\int_0^1 \int_{\mathcal{Y}} (\overline{\phi}(y) - \phi(y))^2 \, \rho_t(y;x) \, \mathrm{d}y\mathrm{d}t \\
&\quad + \int_0^1 \left\{\int_{\mathcal{Y}} (\overline{\phi}(y) - \phi(y)) \cdot \rho_t(y;x)\mathrm{d}y\right\}^2 \mathrm{d}t \\
&= -\int_0^1 \mathbb{V}_{\rho_t(.;x)}[\overline{\phi} - \phi] \, \mathrm{d}t \, . \qquad (22)
\end{aligned}$$

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

$$\langle \delta J(\phi) - \delta J(\overline{\phi}), \phi - \overline{\phi}\rangle = -\int_0^1 \mathbb{E}_{x\sim\mu}\left[\mathbb{V}_{\rho_t(.;x)}[\overline{\phi} - \phi]\right] \, \mathrm{d}t \, . \qquad (23)$$

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

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

By the fundamental theorem of calculus,

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

■

D.3. Proofs for the statements in Section 4

Here, we give the proofs of Theorems 3 and 4.

D.3.1. PROOF OF THEOREM 3

Theorem 3 is a direct corollary of the following two lemmas pertaining to the iterative projection operations (Equations (11a) and (11b)) and the local greedy method (Equation (12)).

Lemma 16 *Consider $\phi^0 \in L^1(\nu)$, and let $\{\phi^n\}_{n\geq 1}$ be the sequence of potentials obtained as $\phi^{n+1} = \mathsf{M}^{\Phi\text{-match}}(\phi^n; \eta)$ for $\eta \in [0,1]$. Then, the sequence of distributions $\{\pi^n\}_{n\geq 0}$ where $\pi^n = \pi(\phi^n, (\phi^n)^+)$ satisfy for every $n \geq 0$*

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

Lemma 17 *Let $\pi \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ be such that $\pi_{\mathcal{X}} = \mu$. Then for $\eta \in [0,1]$,*

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

Lemmas 16 and 17 extend Lemmas 2 and 1 from Reza Karimi et al. (2024) respectively for Φ-match. Specifically, the results in Reza Karimi et al. (2024) were established for Φ-match with $\Phi : f \mapsto f$. The proofs of these lemmas are stated next.

Proof of Lemma 16 Consider some $n \geq 0$. By the decomposition of KL divergence,

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

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

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

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

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

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

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

where $\phi^{n+1/2}(y) = \phi^n(y) + (\log\Phi(\nu)(y) - \log\Phi(\pi^n_{\mathcal{Y}})(y))$ and $\psi^{n+1/2}(x) = \psi^n(x) + \log Z$. Moreover, by direct calcuation

$$\pi^{n+1/2}_{\mathcal{Y}|\mathcal{X}}(y|x) = \exp\left(\phi^{n+1/2}(y) - \frac{c(x,y)}{\varepsilon}\right)\nu(y)\cdot\left(\int_{\mathcal{Y}}\exp\left(\phi^{n+1/2}(y) - \frac{c(x,y)}{\varepsilon}\right)\nu(y)\right)^{-1}$$

From (Reza Karimi et al., 2024, 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

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

where

$$\begin{aligned}\bar\phi(y) &= \eta\cdot\phi^{n+1/2}(y) + (1-\eta)\cdot\phi^n(y)\\ &= \phi^n(y) + \eta\cdot(\log\Phi(\nu)(y) - \log\Phi(\pi^n_{\mathcal{Y}})(y))\,.\end{aligned}\tag{24}$$

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

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

Hence, comparing Equation (24) with Φ-match we have $\mathsf{project}_{\mathcal{X},\mu}(\pi^{n+1/2},\pi^n;\eta) = \pi(\phi^{n+1};(\phi^{n+1})^+)$ which completes the proof. ■

Proof of Lemma 17 For convenience, we use the shorthand notation $\tilde{\pi} = \mathsf{project}_{\mathcal{Y},\nu}(\pi; \Phi)$. As in the proof of Lemma 16, Equation (11a) ensures that

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

The objective in Equation (12) with $\mathcal{F} \leftarrow \mathcal{V}_\Phi$ can be simplified as

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

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

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

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

D.3.2. PROOF OF THEOREM 4

To prove Theorem 4, we use the following key lemma from Aubin-Frankowski et al. (2022) 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 18 ((Aubin-Frankowski et al., 2022, Prop. 14)) *Let $k : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ be a bounded, positive definite kernel where $c_k := \sup_{y \in \mathcal{Y}} k(y, y) < \infty$. Then for any $\xi, \overline{\xi} \in \mathcal{P}(\mathcal{Y})$,*

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

Consequently,

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

Proof of Theorem 4 The proof is be obtained in the manner of the proof of Aubin-Frankowski et al. (2022, Thm. 4) while catering to the squared MMD $\mathcal{L}_k$. We give the details here for completeness.

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

$$\begin{aligned} &\eta \cdot \langle \mathcal{V}_{\Phi_k}(\pi^n), \overline{\pi} - \pi^n \rangle + \mathrm{d}_{\mathrm{KL}}(\overline{\pi} \| \pi^n) \\ &\qquad \geq \eta \cdot \langle \mathcal{V}_{\Phi_k}(\pi^n), \pi^{n+1} - \pi^n \rangle + \mathrm{d}_{\mathrm{KL}}(\pi^{n+1} \| \pi^n) + \mathrm{d}_{\mathrm{KL}}(\overline{\pi} \| \pi^{n+1}) \ . \end{aligned} \tag{25}$$

This is obtained by the three-point identity (Aubin-Frankowski et al., 2022, Lem. 3) with

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

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

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

From Proposition 18, we know that that in this case

$$\begin{aligned} \mathcal{L}_k(\pi^{n+1}_{\mathcal{Y}}; \nu) &\leq \mathcal{L}_k(\pi^n_{\mathcal{Y}}; \nu) + \langle \delta \mathcal{L}_k(\pi^n_{\mathcal{Y}}; \nu), \pi^{n+1}_{\mathcal{Y}} - \pi^n_{\mathcal{Y}} \rangle + 2c_k \cdot \mathrm{d}_{\mathrm{KL}}(\pi^{n+1}_{\mathcal{Y}} \| \pi^n_{\mathcal{Y}}) \\ &= \mathcal{L}_k(\pi^n_{\mathcal{Y}}; \nu) + \langle \mathcal{V}_{\Phi}(\pi^n), \pi^{n+1} - \pi^n \rangle + 2c_k \cdot \mathrm{d}_{\mathrm{KL}}(\pi^{n+1}_{\mathcal{Y}} \| \pi^n_{\mathcal{Y}}) \\ &\overset{(a)}{\leq} \mathcal{L}_k(\pi^n_{\mathcal{Y}}; \nu) + \langle \mathcal{V}_{\Phi}(\pi^n), \pi^{n+1} - \pi^n \rangle + 2c_k \cdot \mathrm{d}_{\mathrm{KL}}(\pi^{n+1} \| \pi^n) \qquad (26) \\ &\overset{(b)}{\leq} \mathcal{L}_k(\pi^n_{\mathcal{Y}}; \nu) + \left(2c_k - \frac{1}{\eta}\right) \cdot \mathrm{d}_{\mathrm{KL}}(\pi^{n+1} \| \pi^n) - \frac{1}{\eta} \cdot \mathrm{d}_{\mathrm{KL}}(\pi^n \| \pi^{n+1}) \\ &\overset{(c)}{\leq} \mathcal{L}_k(\pi^n_{\mathcal{Y}}; \nu) \ . \qquad (27) \end{aligned}$$

Step (a) is a consequence of the KL decomposition, step (b) applies Equation (25) for $\overline{\pi} \leftarrow \pi^n$, and step (c) uses the fact that $\eta = \min\left\{\frac{1}{2c_k}, 1\right\} \leq \frac{1}{2c_k}$ and the non-negativity of the KL divergence. We also have by Proposition 18 that

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

Substituting the above and Equation (26) in Equation (25) with $\eta = \min\left\{\frac{1}{2c_k}, 1\right\}$, we obtain

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

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

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

We know that $\mathcal{L}_k(\pi_{\mathcal{Y}}^{n+1}; \nu) \leq \mathcal{L}_k(\pi_{\mathcal{Y}}^n; \nu)$ from Equation (27). Noting that $\pi^\star$ satisfies $\pi_{\mathcal{X}}^\star = \mu$ and $\pi_{\mathcal{Y}}^\star = \nu$, we substitute $\overline{\pi} \leftarrow \pi^\star$ above and this leads to

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

■

D.4. Proofs of the theorems in Appendix B

D.4.1. PROOF FOR THE NON-ASYMPTOTIC GUARANTEE FOR SIGN-SGA

Proof of Theorem 10 From Lemmas 1 and 2, we have for any $\phi, \overline{\phi} \in L^1(\nu) \cap L^\infty(\mathcal{Y})$ that

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

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

$$\begin{aligned} J(\phi^{n+1/2}) &\geq J(\phi^n) + \langle \delta J(\phi^n), \phi^{n+1/2} - \phi^n \rangle - \frac{\|\phi^{n+1/2} - \phi^n\|_\infty^2}{2} \\ &\overset{(a)}{=} J(\phi^n) + \|\delta J(\phi^n)\|_{L^1(\mathcal{Y})} \cdot \|\delta J(\phi^n)\|_{L^1(\mathcal{Y})} - \frac{1}{2} \cdot \|\delta J(\phi^n)\|_{L^1(\mathcal{Y})}^2 \\ &= J(\phi^n) + \frac{\|\delta J(\phi^n)\|_{L^1(\mathcal{Y})}^2}{2} \,. \end{aligned} \tag{28}$$

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

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

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

$$J(\widetilde{\phi}^\star) \leq J(\phi^n) + \langle \delta J(\phi^n), \widetilde{\phi}^\star - \phi^n \rangle \,. \tag{29}$$

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

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

Above, step (a) applies Equations (29) and (28), and step (b) applies the Hölder-Young inequality. Finally, step (c) uses the fact that $\phi^n, \widetilde{\phi^\star} \in \mathcal{T}_{\phi^0, y_{\text{anc}}}$ shown previously. Summing the above inequality from $n = 0$ to $n = N - 1$, we get

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

■

D.4.2. PROOFS FOR THE NON-ASYMPTOTIC GUARANTEES FOR PROJ-SGA AND PROJ-SGA++

Before we give the proofs, we lay out some preliminaries and intermediate results that will come in handy to prove Theorem 13 later.

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

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

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

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

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

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

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

Lemma 19 *Let $\phi \in \mathcal{S}_{\bar{B}}$. Then, for $\eta \leq \lambda(\max\{B, \bar{B}\})^{-1}$,*

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

Lemma 20 *Let $\phi \in \mathcal{S}_{\bar{B}}$. For any $\overline{\phi} \in L^2(\nu)$ and $\eta \leq \lambda(\max\{B, \bar{B}\})^{-1}$, we have that*

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

Proof of Theorem 12 For $\phi^0 \in \mathcal{S}_B$, each step according to proj-SGA ensures that $\phi^n \in \mathcal{S}_B$ for all $n \geq 1$. For $\eta \leq \frac{1}{\lambda(B)}$, we have from Lemma 20 applied to $\overline{\phi} \leftarrow \widetilde{\phi}^\star$ and $\phi \leftarrow \phi^n$ for an arbitrary $n \geq 0$ that

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

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

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

Additionally from Lemma 19, we have for the choice of η,

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

Hence,

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

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

The proof for Theorem 13 uses the following two lemmas. The second lemma is analogous to (Beck and Teboulle, 2009, Lem. 4.1).

Lemma 21 *Consider the recursion*

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

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

Lemma 22 *Let* $\{\overline{\phi}^n\}_{n \geq 1}$ *be obtained from proj-SGA++. Define* $v_n = \overline{J}(\phi^\star) - \overline{J}(\overline{\phi}^n)$ *and* $u_n = t_n \cdot \overline{\phi}^n - (t_n - 1) \cdot \overline{\phi}^{n-1} - \widetilde{\phi}^\star$. *Then,*

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

Proof of Theorem 13 Since $\lambda(3B) \geq \lambda(B)$ and $\phi^1, \overline{\phi}^1 \in \mathcal{S}_B$, Lemma 20 with $\phi \leftarrow \phi^1, \overline{\phi} \leftarrow \widetilde{\phi}^\star$ gives

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

In the notation of Lemma 22,

$$-v_1 \geq \frac{\lambda(3B)}{2} \cdot \|u_1\|^2_{L^2(\nu)} - \frac{\lambda(3B)}{2} \cdot \|\overline{\phi}^0 - \widetilde{\phi}^\star\|^2_{L^2(\nu)} \,. \tag{30}$$

Telescoping the identity from Lemma 22 for $n = 1$ to $N - 1$ gives

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

Rearranging the terms, we have

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

where the last step follows from Equation (30). Since $t_N \geq \frac{N+1}{2}$, we have

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

■

D.4.3. PROOF OF SUPPLEMENTARY RESULTS

Proof of Lemma 21 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 $\frac{t_k - 1}{t_{k+1}} \leq 1$. ■

Proof of Lemma 22 First, since $\overline{\phi}^n \in \mathcal{S}_B$ for all $n \geq 1$ and $\frac{t_n - 1}{t_{n+1}} \leq 1$ (Lemma 21), 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 20 to two settings. First, with $\phi \leftarrow \phi^{n+1}, \overline{\phi} \leftarrow \overline{\phi}^n, \bar{B} \leftarrow 3B$, we have

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

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

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

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 (Beck and Teboulle, 2009, Lem. 4.1). ■

D.5. Proofs for the statements pertaining to the dynamical Schrödinger bridge problem

D.5.1. PROOF OF PROPOSITION 9

Proof We begin by noting that $\mathbb{P}^{n+1/2}$ is the solution to the following SDE

$$\mathrm{d}Y_t = [-v^n_{T-t}(Y_t) + 2\nabla \log p^n_{T-t}(Y_t)]\mathrm{d}t + \sqrt{2}\cdot \mathrm{d}B_t, \quad Y_0 \sim \widetilde{\mathbb{P}}^n$$

which corresponds to the time-reversal of the SDE that defines $\mathbb{P}^n$ and with initial condition given by $\widetilde{\mathbb{P}}^n$ whose density is proportional to $\mathbb{P}^n_T \cdot \frac{\Phi(\nu)}{\Phi(\mathbb{P}^n_T)}$.

The second step follows from Reza Karimi et al. (2024, Thm. 4.2), which permits us to represent path-Φ-match(b) as

$$\begin{aligned}\mathrm{d}X_t &= \left(\eta\left[v^n_t(X_t) - 2\nabla \log p^n_t(X_t) + 2\nabla \log p^{n+1/2}_t(X_t)\right] + (1-\eta)v^n_t(X_t) - 2\nabla V_t(X_t)\right)\mathrm{d}t \\ &\quad + \sqrt{2}\cdot \mathrm{d}B_t\,, \\ &= [v^n_t(X_t) - 2\eta\nabla \log p^n_t(X_t) + 2\eta\nabla \log p^{n+1/2}_t(X_t) - 2\nabla V_t(X_t)]\mathrm{d}t + \sqrt{2}\cdot \mathrm{d}B_t \end{aligned} \tag{31}$$

where $X_0 \sim \mu$. The extra drift V_t is as defined in the statement of the proposition. ■

D.5.2. PROOF OF LEMMA 14

Proof The proof of this statement is based on two key equivalences. Recall that $\phi^n_t = \log g^n_t$ for all $n \geq 0$ and $t \in [0, T]$. These equivalences are:

1. between the dual potential ϕ^n from eOT and backward dynamics on $\{\phi^n_t\}_t$ determined by the reference transition: this follows from Caluya and Halder (2022); Léonard (2014) in the case of nonlinear drift (i.e., $u^{\mathsf{ref}} \neq 0$).

2. between the updates on the drifts of the SDE $v^n_t = u^{\mathsf{ref}} + \nabla\phi^n_t$ and the path measures $\mathbb{P}^n$ factorized as Equation (20): this follows from classical result on Doob's h-transform, which implies that the optimal additional drift for $\mathbb{P}^\star$ should be in the form of $(\nabla \log g^\star_t)_{t\in[0,T]}$ built from the optimal potential $g^\star_T = e^{\phi^\star}$. The fact that Equation (20) is the same as the law of the SDE Equation (21) is also a consequence of the same twisted kernel argument (Dai Pra, 1991).

The claim about convergence is primarily due to the factorisation of the path measure. Since the bridges for these path measures satisfy $\mathbb{P}^n(X_{t\in(0,T)}|X_0, X_T) = \mathbb{P}^{\mathsf{ref}}(X_{t\in(0,T)}|X_0, X_T)$, we have that $\mathrm{d}_{\mathrm{KL}}(\mathbb{P}^n\|\mathbb{P}^\star) = \mathrm{d}_{\mathrm{KL}}(\pi^n\|\pi^\star)$. Additionally, due to constraint that $\mathbb{P}^n_0 = \mu$ in Equation (21), this rate is determined by $\mathbb{P}^n_T \to \nu$ (or equivalently $\phi^n_T = \phi^n \to \phi^\star$), similar to the static two-marginal case for Φ-match. Notably, we also maintain the coupling π^n as $\pi(\phi^n, (\phi^n)^+)$ which ensures that they satisfy the form of the optimal coupling in Equation (4). ■