On Convergence of Approximate Schrödinger Bridge with Bounded Cost

Wei Deng [*1] **Yu Chen** [*1] **Nicole Tianjiao Yang** [*2] **Hengrong Du** [3] **Qi Feng** [4] **Ricky T. Q. Chen** [5]

Abstract

The Schrödinger bridge has demonstrated promising applications in generative models. It is an entropy-regularized optimal-transport (EOT) approach that employs the iterative proportional fitting (IPF) algorithm to solve an alternating projection problem. However, due to the complexity of finding precise solutions for the projections, approximations are often required. In our study, we study the convergence of the IPF algorithm using approximated projections and a bounded cost function. Our results demonstrate an approximate linear convergence with bounded perturbations. While the outcome is not unexpected, the rapid linear convergence towards smooth trajectories suggests the potential to examine the efficiency of the Schrödinger bridge compared to diffusion models.

1. Introduction

The Schrödinger bridge (SB) problem, originally stemming from quantum mechanics, offers a dynamic formulation of entropy-regularized optimal transport (EOT) (Peyré & Cuturi, 2019). However, the dynamic EOT map is often too computationally demanding to solve directly. Instead, solving the SB problem relies on iterative projections to every other marginal distribution, which motivates the iterative proportional fitting (IPF) algorithm (Kullback, 1968; Ruschendorf, 1995). The first training stage of IPF is equivalent to the exact training of diffusion models, allowing the empirical knowledge gained from diffusion models to be seamlessly inherited. The following training stages of IPF continue to optimize the transport efficiency, which significantly facilitates the estimation of score functions and enables non-linear transport (Liu et al., 2023; Deng et al., 2020; 2022). As such, the minimized transport cost not only *accelerates the inference speed* but also potentially *enhances the generation quality*. To accelerate inference, various active research (Albergo & Vanden-Eijnden, 2023; Liu, 2022; Lipman et al., 2023; Pooladian et al., 2023) has been studied to encourage straighter trajectories, while the *generation enhancement* is still not well understood *in theory* due in part to the Hutchinson estimator in the divergence evaluation (Hutchinson, 1989).

In addition, the understanding of the convergence of Schrödinger bridge in the perspective of optimal transport with a bounded cost is still not well studied in the literature. To tackle this issue, we show an approximately linear convergence of the approximate IPF algorithm based on a bounded transport cost function. In particular, the analysis allows extra perturbations in the projections and bridges the gap between theory and practice. The fast linear convergence of the optimal transport toward smooth trajectories implies a potential for studying the speed advantages of the Schrödinger bridge over diffusion models.

2. Preliminaries

2.1. Schrödinger Bridge Problem

Despite the noticeable success of diffusion models in generative tasks, they have limitations in efficiently transporting distributions (Lavenant & Santambrogio, 2022). Notably, the forward process requires a significantly long time T to approach the prior distribution and facilitate the score estimation, which inevitably leads to a slow inference and large numerical errors (De Bortoli et al., 2021; Lee et al., 2022). To solve that problem, the dynamical Schrödinger Bridge problem (SBP) proposes to minimize a Kullback–Leibler projection (Peyré & Cuturi, 2019)

$$\inf_{\mathbb{P}\in\mathcal{D}(\mu_\star,\nu_\star)} \mathrm{KL}(\mathbb{P}|\mathbb{Q}), \tag{1}$$

where the coupling $\mathbb{P}$ belongs to the set of *path measures* $\mathcal{D}(\mu_\star, \nu_\star)$ with marginal measures $\mu_\star$ at time $t = 0$ and $\nu_\star$ at $t = T$; $\mathbb{Q}$ is the prior path measure, such as the measure induced by the path of the Brownian motion (Wiener process) or Ornstein-Uhlenbeck process. From the perspective of stochastic optimal control (SOC) (see Section 4.4 of Chen et al. (2021)), the dynamical SBP aims to alleviate

[*]Equal contribution [1]Machine Learning Research, Morgan Stanley [2]Department of Mathematics, Emory University [3]Department of Mathematics, Vanderbilt University [4]Department of Mathematics, University of Michigan [5]Meta AI (FAIR). Correspondence to: Wei Deng <weideng056@gmail.com>.

Workshop on New Frontiers in Learning, Control, and Dynamical Systems at the International Conference on Machine Learning (ICML), Honolulu, Hawaii, USA, 2023.

the expensive transport cost by minimizing the control cost along the forward process

$$\inf_{\boldsymbol{u}\in\mathcal{U}} \mathbb{E}\bigg\{\int_0^T \frac{1}{2}\|\boldsymbol{u}(\mathbf{x}_t,t)\|_2^2 \mathrm{d}t\bigg\} \tag{2}$$

$$\text{s.t. } \mathrm{d}\mathbf{x}_t = [\boldsymbol{f}(\mathbf{x}_t,t)+g(t)\boldsymbol{u}(\mathbf{x}_t,t)]\,\mathrm{d}t + \sqrt{2\varepsilon}g(t)\mathrm{d}\mathbf{w}_t \tag{3}$$

$$\mathbf{x}_0 \sim \mu_\star(\cdot),\ \ \mathbf{x}_T \sim \nu_\star(\cdot),$$

where $\boldsymbol{u}(\cdot)$ is a deterministic control function belongs to an admissible control set $\mathcal{U}$; ε is a scalar and is also the regularizer for the underlying entropic optimal transport (EOT).

3. From Schrödinger Bridge to Entropic Optimal Transport

Diffusion models suffer from sub-optimal transport efficiency and are slow in the inference. By contrast, Schrödinger bridge leverages optimal transport (De Bortoli et al., 2021; Vargas et al., 2021; Chen et al., 2022b), which accelerates the inference period and also facilitates score estimations. The connections of Schrödinger bridge and entropic optimal transport are detailed hereinafter.

By the disintegration of measures (Léonard, 2014), the dynamical SBP (1) yields a chain rule (De Bortoli et al., 2021) as follows

$$\mathrm{KL}(\mathbb{P}|\mathbb{Q}) = \mathrm{KL}(\pi|\mathcal{G}) + \iint_{\mathrm{X}\times\mathrm{Y}} \mathrm{KL}(\mathbb{P}_\pi|\mathbb{Q}_\mathcal{G})\mathrm{d}\pi(\mathbf{x}_0,\mathbf{x}_T). \tag{4}$$

where $\pi\in\Pi(\mu_\star,\nu_\star)$ and Π is the space of couplings with marginals $\mu_\star$ and $\nu_\star$; $\mathcal{G}$ denotes a Gibbs measure that follows $\mathrm{d}\mathcal{G}\propto e^{-c_\varepsilon}\mathrm{d}(\mu_\star\otimes\nu_\star)$; the product measure is denoted by $\otimes$ and $c_\varepsilon(\mathbf{x},\mathbf{y})$ is a loss function that models the transport cost between particles $\mathbf{x}$ and $\mathbf{y}$; the conditional probability of $\mathbb{P}$ (or $\mathbb{Q}$) given conditional information π (or $\mathcal{G}$) is denoted by $\mathbb{P}_\pi$ (or $\mathbb{Q}_\mathcal{G}$) (De Bortoli, 2022a). Forcing $\mathbb{P}_\pi=\mathbb{Q}_\mathcal{G}$ yields the *static* SBP with the optimal coupling $\pi_\star$:

$$\pi_\star = \underset{\pi\in\Pi(\mu_\star,\nu_\star)}{\arg\min}\ \mathrm{KL}(\pi|\mathcal{G}), \tag{5}$$

where $\pi_\star$ is known as the Schrödinger bridge from $\mu_\star$ to $\nu_\star$ (if exists). The static SBP formulation yields a structure property (Peyré & Cuturi, 2019; Nutz, 2022; Chen et al., 2023) and allows us to represent the Schrödinger bridge $\pi_\star$ using Schrödinger potentials $\varphi_\star$ and $\psi_\star$:

$$\mathrm{d}\pi_\star(\mathbf{x},\mathbf{y}) = e^{\varphi_\star(\mathbf{x})+\psi_\star(\mathbf{y})-c_\varepsilon(\mathbf{x},\mathbf{y})}\mathrm{d}(\mu_\star\otimes\nu_\star). \tag{6}$$

This crucial representation enables us to efficiently compute the optimal coupling and thereby accelerate the inference process. Moreover, it establishes a connection between the static SBP and entropic optimal transport.

$$\begin{aligned}\mathrm{KL}(\pi|\mathcal{G}) &= \iint_{\mathrm{X}\times\mathrm{Y}} \log\bigg(\frac{\mathrm{d}\pi}{d(\mu_\star\otimes\nu_\star)}\frac{d(\mu_\star\otimes\nu_\star)}{d\mathcal{G}}\bigg)\,\mathrm{d}\pi \\ &= \iint_{\mathrm{X}\times\mathrm{Y}} c_\varepsilon \mathrm{d}\pi + \mathrm{KL}(\pi|\mu_\star\otimes\nu_\star)+\mathcal{C}.\end{aligned} \tag{7}$$

where $\mathrm{X}\in\mathbb{S}^d$ and $\mathrm{Y}\in\mathbb{S}^d$ are the parameter spaces of interest, $\mathcal{C}$ is the normalizing constant for $\mathcal{G}$. For the rescaled cost of the form $c_\varepsilon=c/\varepsilon$, the static SBP in the problem (5) is equivalent to the standard EOT with a ε-regularizer:

$$\inf_{\pi\in\Pi(\mu_\star,\nu_\star)} \iint_{\mathrm{X}\times\mathrm{Y}} c(\mathbf{x},\mathbf{y})\pi(\mathrm{d}\mathbf{x},\mathrm{d}\mathbf{y}) + \varepsilon\cdot\mathrm{KL}(\pi|\mu_\star\otimes\nu_\star). \tag{8}$$

3.1. Duality for Static Schrödinger bridges

The static Schrödinger bridge is a constrained optimization problem and naturally yields a dual formulation.

Lemma 1 (Duality. Theorem 3.2 (Nutz, 2022))**.** *Assume $\pi\in\Pi(\mu_\star,\nu_\star)$ is the Schrödinger bridge with potentials $(\varphi,\psi)\in L^1(\mu_\star)\times L^1(\nu_\star)$. We have that*

$$\begin{aligned}&\min_{\pi\in\Pi(\mu_\star,\nu_\star)} \mathrm{KL}(\pi|\mathcal{G}) = \max_{\varphi,\psi} G(\varphi,\psi),\\ &G(\varphi,\psi) := \mu_\star(\varphi)+\nu_\star(\psi) - \iint_{\mathrm{X}\times\mathrm{Y}} e^{\varphi\oplus\psi}\mathrm{d}\mathcal{G} + 1,\end{aligned} \tag{9}$$

where $\mu_\star(\varphi)=\int_\mathrm{X}\varphi\mathrm{d}\mu_\star$, $\nu_\star(\psi)=\int_\mathrm{Y}\psi\mathrm{d}\nu_\star$.

We notice that there is no duality gap, which motivates us to maximize the concave dual problem G in Eq.(9) instead. An effective solver for this problem is an alternating maximization scheme. In this scheme, we first optimize $\varphi_{k+1}=\arg\max_\varphi G(\cdot,\psi_k)$, and then optimize $\psi_{k+1}=\arg\max_\psi G(\varphi_{k+1},\cdot)$.

From a geometric perspective, alternating maximization corresponds to alternating projections

$$\begin{aligned}&\varphi_{k+1} = \arg\max_\varphi G(\cdot,\psi_k) \Longrightarrow \\ &\qquad \text{the first marginal of } \pi(\varphi_{k+1},\psi_k) \text{ is } \mu_\star,\end{aligned} \tag{10a}$$

$$\begin{aligned}&\psi_{k+1} = \arg\max_\psi G(\varphi_{k+1},\cdot) \Longrightarrow \\ &\qquad \text{the second marginal of } \pi(\varphi_{k+1},\psi_{k+1}) \text{ is } \nu_\star.\end{aligned} \tag{10b}$$

To see why (10b) holds. We first denote the second marginal of $\mathrm{d}\pi(\varphi_k,\psi_k) := e^{\varphi_k\oplus\psi_k}\mathrm{d}\mathcal{G}$ by ν_k and then proceed to show $\nu_k=\nu_\star$ (Nutz, 2022). Recall that G is concave and $\psi_k=\arg\max_\psi G(\varphi_k,\cdot)$, it suffices to show that given fixed $\varphi_k\in L^1(\mu_\star)$, $\psi_k\in L^1(\nu_\star)$, a constant η and bounded measurable function $\delta_\psi:\mathbb{R}^d\to\mathbb{R}$, the maximality of

$G(\varphi_k, \psi_k)$ implies that

$$\begin{aligned}0 =& \frac{\mathrm{d}}{\mathrm{d}\eta}\Big|_{\eta=0} G(\varphi_k, \psi_k + \eta\delta_\psi) \\ =& \nu_\star(\delta_\psi) - \iint_{\mathrm{X}\times\mathrm{Y}} \delta_\psi e^{\varphi_k \oplus \psi_k} \mathrm{d}\mathcal{G} \\ =& \nu_\star(\delta_\psi) - \int_{\mathrm{Y}} \delta_\psi \mathrm{d}\nu_k = \nu_\star(\delta_\psi) - \nu_k(\delta_\psi),\end{aligned} \tag{11}$$

where the second equality follows by Taylor expansion. Similarly, we can show (10a).

3.2. Approximated iterative proportional fitting (aIPF)

Schrödinger bridge proposes to solve the Schrödinger equation (Nutz & Wiesel, 2022)

$$\begin{aligned}\psi_\star(\mathbf{y}) &= -\log \int_{\mathrm{X}} e^{\varphi_\star(\mathbf{x}) - c_\varepsilon(\mathbf{x},\mathbf{y})} \mu_\star(\mathrm{d}\mathbf{x}), \\ \varphi_\star(\mathbf{x}) &= -\log \int_{\mathrm{Y}} e^{\psi_\star(\mathbf{y}) - c_\varepsilon(\mathbf{x},\mathbf{y})} \nu_\star(\mathrm{d}\mathbf{y}).\end{aligned} \tag{12}$$

Since the Schrödinger potential functions $(\psi_\star, \varphi_\star)$ are not known *a priori*, the iterative proportional fitting (IPF) algorithm (Kullback, 1968; Ruschendorf, 1995), also known as Sinkhorn algorithm, was proposed to solve the alternating projections in Eq.(10) as follows

$$\begin{aligned}\text{IPF}: \quad \psi_k(\mathbf{y}) &= -\log \int_{\mathrm{X}} e^{\varphi_k(\mathbf{x}) - c_\varepsilon(\mathbf{x},\mathbf{y})} \mu_\star(\mathrm{d}\mathbf{x}), \\ \varphi_{k+1}(\mathbf{x}) &= -\log \int_{\mathrm{Y}} e^{\psi_k(\mathbf{y}) - c_\varepsilon(\mathbf{x},\mathbf{y})} \nu_\star(\mathrm{d}\mathbf{y}).\end{aligned} \tag{13}$$

However, given a finite computational budget, projecting to the ideal measure $\mu_\star$ (or $\nu_\star$) in Eq.(10) at each iteration may not be practical. Instead, some close approximation $\mu_{\star,k+1}$ (or $\nu_{\star,k}$) is used at iteration $2k+1$ (or $2k$) via Gaussian processes (Vargas et al., 2021) or neural networks (De Bortoli et al., 2021; Chen et al., 2022b). Therefore, one may resort to an approximate solution that still achieves a reasonable accuracy within a finite budget:

$$\mu_{2k+1} = \mu_{\star,k+1} \approx \mu_\star, \quad \nu_{2k} = \nu_{\star,k} \approx \nu_\star, \tag{14}$$

We refer to the IPF algorithm with approximate marginals as approximate IPF (aIPF) and present aIPF in Algorithm 1. The difference between IPF and aIPF is detailed in Figure 1.

The structure representation (6) can be naturally extended based on approximate marginals as follows

$$\begin{aligned}\mathrm{d}\pi_{2k} &= e^{\varphi_k \oplus \psi_k - c_\varepsilon} \mathrm{d}(\mu_{\star,k} \otimes \nu_{\star,k}), \\ \mathrm{d}\pi_{2k-1} &= e^{\varphi_k \oplus \psi_{k-1} - c_\varepsilon} \mathrm{d}(\mu_{\star,k} \otimes \nu_{\star,k-1}),\end{aligned} \tag{16}$$

where π_k is the approximate coupling at iteration k.

Algorithm 1 One iteration of approximate IPF (aIPF). IPF serves as the theoretical solver for the EOT formulation of SBP. In practice, the forward-backward SDE (Chen et al., 2022b; 2023) can be utilized to solve the integrals (15) and obtain the approximate measures $\mu_{\star,k}$ and $\nu_{\star,k}$.

$$\begin{aligned}\psi_k(\mathbf{y}) &= -\log \int_{\mathrm{X}} e^{\varphi_k(\mathbf{x}) - c_\varepsilon(\mathbf{x},\mathbf{y})} \mu_{\star,k}(\mathrm{d}\mathbf{x}), \\ \varphi_{k+1}(\mathbf{x}) &= -\log \int_{\mathrm{Y}} e^{\psi_k(\mathbf{y}) - c_\varepsilon(\mathbf{x},\mathbf{y})} \nu_{\star,k}(\mathrm{d}\mathbf{y}).\end{aligned} \tag{15}$$

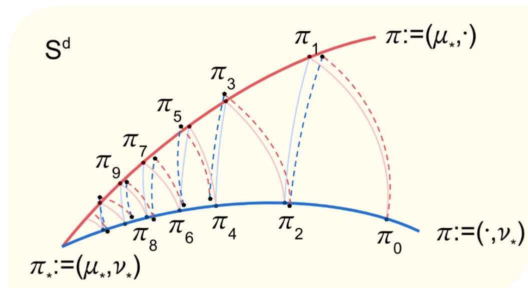

Figure 1. Comparison between IPF and aIPF. The exact (or approximate) projections of IPF (or aIPF) are highlighted through the solid lines (or dotted lines). IPF projects to the exact marginals, but aIPF may not.

4. Convergence of Entropic Optimal Transport with Bounded Cost

The Schrödinger bridge problem (SBP) has made significant progress both theoretically and empirically (Nutz & Wiesel, 2022; Ghosal & Nutz, 2022; Eckstein & Nutz, 2022; De Bortoli et al., 2021; Chen et al., 2022b; Vargas et al., 2021). In particular, SBP provides a classical linear convergence result based on bounded cost functions. However, the convergence of aIPF based on approximate marginals has not been studied yet. To fill this gap, we extend the analysis by introducing controllable perturbations. The key to our proof is the strong convexity of the dual formulation. For any $\varphi \in L^1(\mu_\star)$ and $\psi \in L^1(\nu_\star)$, we consider the objective function of the dual EOT problem based on the ideal measures $\mu_\star$ and $\nu_\star$, which can be expressed as follows:

$$\begin{aligned}G(\varphi, \psi) :=& \mu_\star(\varphi) + \nu_\star(\psi) \\ &- \iint_{\mathrm{X}\times\mathrm{Y}} e^{\varphi \oplus \psi - c_\varepsilon} \mathrm{d}(\mu_\star \otimes \nu_\star) + 1.\end{aligned} \tag{17}$$

It is worth noting that the marginals $\mu_\star$ and $\nu_\star$ at iteration k are not directly accessible and we opt for approximates measures $\mu_{\star,k}$ and $\nu_{\star,k}$. To quantify the convergence of the dual objective, we first introduce the following assumptions

Assumption A1 (Bounded Cost)**.** *The cost function $c_\varepsilon(x, y)$ defined on $\mathrm{X} \times \mathrm{Y}$ is continuous and bounded.*

This assumption has been widely used in Nutz & Wiesel (2022); Carlier & Laborde (2020); Deligiannidis et al. (2021); Eckstein & Nutz (2022). Similarly, one can also prove the linear convergence for continuous marginals when the parameter space $\mathrm{X} \times \mathrm{Y}$ is compact (Chen et al., 2016). Interested readers are encouraged to explore the study of diffusion models on constrained domains (Fishman et al., 2023).

Assumption A2 (Lipschitz smoothness)**.** *The energy functions of $\mu_\star$ and $\nu_\star$ are L-Lipschitz smooth.*

$$\Big\|\nabla \log \frac{\mathrm{d}\mu_\star}{\mathrm{d}\mathbf{x}}(\mathbf{x}_1) - \nabla \log \frac{\mathrm{d}\mu_\star}{\mathrm{d}\mathbf{x}}(\mathbf{x}_2)\Big\|_2 \leq L\|\mathbf{x}_1 - \mathbf{x}_2\|_2,$$
$$\Big\|\nabla \log \frac{\mathrm{d}\nu_\star}{\mathrm{d}\mathbf{y}}(\mathbf{y}_1) - \nabla \log \frac{\mathrm{d}\nu_\star}{\mathrm{d}\mathbf{y}}(\mathbf{y}_2)\Big\|_2 \leq L\|\mathbf{y}_1 - \mathbf{y}_2\|_2.$$

The assumption is standard and has been used in Lee et al. (2022); Chen et al. (2022a; 2023).

Assumption A3 (Score approximation)**.** *For any $k \in \mathbb{N}$, $\mathbf{x} \in \mathrm{X}$ and $\mathbf{y} \in \mathrm{Y}$, $\nabla \log \frac{\mathrm{d}\mu_{\star,k}(\mathbf{x})}{\mathrm{d}\mathbf{x}}$ and $\nabla \log \frac{\mathrm{d}\nu_{\star,k}(\mathbf{y})}{\mathrm{d}\mathbf{y}}$ are the ϵ approximation of score functions $\nabla \log \frac{\mathrm{d}\mu_\star(\mathbf{x})}{\mathrm{d}\mathbf{x}}$ and $\nabla \log \frac{\mathrm{d}\nu_\star(\mathbf{y})}{\mathrm{d}\mathbf{y}}$ at the k-th iteration, respectively*

$$\Big\|\nabla \log \frac{\mathrm{d}\mu_{\star,k}(\mathbf{x})}{\mathrm{d}\mathbf{x}} - \nabla \log \frac{\mathrm{d}\mu_\star(\mathbf{x})}{\mathrm{d}\mathbf{x}}\Big\|_2 \leq \epsilon(1 + \|\mathbf{x}\|_2),$$
$$\Big\|\nabla \log \frac{\mathrm{d}\nu_{\star,k}(\mathbf{y})}{\mathrm{d}\mathbf{y}} - \nabla \log \frac{\mathrm{d}\nu_\star(\mathbf{y})}{\mathrm{d}\mathbf{y}}\Big\|_2 \leq \epsilon(1 + \|\mathbf{y}\|_2).$$

The assumption is a standard one in the field (De Bortoli, 2022b), and stronger assumptions have been utilized in related works such as De Bortoli et al. (2021); Lee et al. (2022); Chen et al. (2022a; 2023). The errors in the two marginals don't have to be the same, and we use a unified ϵ mainly for analytical convenience.

Assumption A4 (Fourth Moment)**.** *The probability densities for $\mu_\star$ and $\nu_\star$ have bounded fourth moment.*

The assumption is standard and is slightly stronger than the bounded second-moment assumption used in Chen et al. (2022a).

Approximately linear convergence and proof sketches We first follow Carlier (2022) and Marino & Gerolin (2020) in building a *centered* aIPF algorithm in Algorithm 2 with scaled potential functions $\bar{\varphi}_k$ and $\bar{\psi}_k$ such that $\mu_\star(\bar{\varphi}_k) = 0$. The *centering* operation doesn't change the dual objective but ensures that the aIPF iterates are uniformly bounded in Lemma 3. Moreover, the centering operation allows perturbations in the marginals for practical analysis.

We next exploit the *strong convexity* of the exponential function e^x associated with the dual in a supporting Lemma 4 to show a key result $G(\bar{\varphi}_\star, \bar{\psi}_\star) - G(\bar{\varphi}_k, \bar{\psi}_k) \lesssim O\big(\|\bar{\varphi}_{k+1} - \bar{\varphi}_k\|^2_{L^2(\mu_\star)} + \epsilon\big)$ in Lemma 6. Together with the upper bound $\|\bar{\varphi}_{k+1} - \bar{\varphi}_k\|^2_{L^2(\mu_\star)} \lesssim O\big(G(\bar{\varphi}_{k+1}, \bar{\psi}_{k+1}) - G(\bar{\varphi}_k, \bar{\psi}_k) + \epsilon\big)$ in Lemma 5, we can derive the desired contraction properties for the dual objective:

$$G(\bar{\varphi}_\star, \bar{\psi}_\star) - G(\bar{\varphi}_k, \bar{\psi}_{k+1}) \lesssim \beta_\varepsilon \bigg(G(\bar{\varphi}_\star, \bar{\psi}_\star) - G(\bar{\varphi}_k, \bar{\psi}_k) \bigg),$$

where $\beta_\varepsilon := 1 - e^{-24\|c_\varepsilon\|_\infty} \in (0, 1)$. Since the centering operation doesn't change the dual objective, we can freely obtain our first main theorem regarding the convergence of Algorithm 1:

Theorem 1. *Given assumptions A1, A2, A3, and A4 hold. Let $(\varphi_k, \psi_k)_{k\geq 0}$ be the iterates of Algorithm 1. Then*

$$\begin{aligned} &G(\varphi_\star, \psi_\star) - G(\varphi_k, \psi_k) \\ &\leq \beta_\varepsilon^k \big(G(\varphi_\star, \psi_\star) - G(\varphi_0, \psi_0)\big) + O\big(\epsilon\big), \end{aligned} \tag{18}$$

where $\beta_\varepsilon := 1 - e^{\frac{-24\|c\|_\infty}{\varepsilon}} \in (0, 1)$; ϵ is the score approximation error at the k-th IPF iteration.

As we increase the entropic regularizer ε, we observe a faster linear convergence. However, this increase also leads to a larger bias towards the solution of the generalized Kantorovich problem $\iint_{\mathrm{X}\times\mathrm{Y}} c(\mathbf{x}, \mathbf{y})\pi(\mathrm{d}\mathbf{x}, \mathrm{d}\mathbf{y})$ (Peyré & Cuturi, 2019; Nutz, 2022). Therefore, in practical applications, it becomes necessary to strike a balance and carefully consider the trade-off involved. One can also consider the Hilbert-Birkhoff projective metric to derive similar iterative contraction properties as in Chen et al. (2016); Franklin & Lorenz (1989), where the approximate convergence can be left as a future work.

5. Conclusions and Limitation

Diffusion models have emerged as the backbone of deep generative models and have made significant advancements. However, despite their statistical and dimension-free potential, diffusion models suffer from sub-optimal transport efficiency and slow inference. The Schrödinger bridge algorithm overcomes these issues by utilizing entropic optimal transport to optimize transport efficiency. We demonstrate the convergence analysis of the approximate IPF algorithm based on a bounded cost function. The fast linear convergence towards *smoother trajectories* also implies that the Schrödinger bridge algorithm may be more efficient in generating samples of higher quality.

However, we acknowledge that bounded cost functions are limited to specific applications and may not offer sufficient generality. Future work may extend this assumption to general cost functions by leveraging the contraction properties of the marginals (Conforti et al., 2023). Moreover, the adaptable formulation shows the promise in effectively utilizing non-linear priors, as indicated by (Deng et al., 2020; 2022), to expedite generative tasks that deal with complex multimodal distributions.

Acknowledgements

We thank the reviewers for their valuable suggestions.

References

Albergo, M. S. and Vanden-Eijnden, E. Building Normalizing Flows with Stochastic Interpolants. In *Proc. of the International Conference on Learning Representation (ICLR)*, 2023.

Carlier, G. On the Linear Convergence of the Multi-Marginal Sinkhorn Algorithm. *SIAM Journal on Optimization*, 32:2:10.1137, 2022.

Carlier, G. and Laborde, M. A Differential Approach to the Multi-marginal Schrödinger System. *SIAM J. Math. Anal.*, 52:1:709–717, 2020.

Chen, S., Chewi, S., Li, J., Li, Y., Salim, A., and Zhang, A. R. Sampling is as Easy as Learning the Score: Theory for Diffusion Models with Minimal Data Assumptions. *arXiv preprint arXiv:2209.11215v2*, 2022a.

Chen, T., Liu, G.-H., and Theodorou, E. A. Likelihood Training of Schrödinger Bridge using Forward-Backward SDEs Theory. In *Proc. of the International Conference on Learning Representation (ICLR)*, 2022b.

Chen, Y., Georgiou, T. T., and Pavon, M. Entropic and Displacement Interpolation: a Computational Approach using the Hilbert Metric. *SIAM Journal on Applied Mathematics*, 76:2375–2396, 2016.

Chen, Y., Georgiou, T. T., and Pavon, M. Stochastic Control Liaisons: Richard Sinkhorn Meets Gaspard Monge on a Schrödinger Bridge. *SIAM Review*, 63(2):249–313, 2021.

Chen, Y., Deng, W., Fang, S., Li, F., Yang, N., Zhang, Y., Rasul, K., Zhe, S., Schneider, A., and Nevmyvaka, Y. Provably Convergent Schrödinger Bridge with Applications to Probabilistic Time Series Imputation. In *Proc. of the International Conference on Machine Learning (ICML)*, 2023.

Conforti, G., Durmus, A., and Greco, G. Quantitative Contraction Rates for Sinkhorn Algorithm: Beyond Bounded Costs and Compact Marginals. *arXiv:2304.04451*, 2023.

De Bortoli, V. Generative Modeling via Schrödinger bridge (basics on Schrödinger bridge). *Slides*, 2022a.

De Bortoli, V. Convergence of Denoising Diffusion Models under the Manifold Hypothesis. *Transactions on Machine Learning Research (TMLR)*, 2022b.

De Bortoli, V., Thornton, J., Heng, J., and Doucet, A. Diffusion Schrödinger Bridge with Applications to Score-Based Generative Modeling. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Deligiannidis, G., Bortoli, V. D., and Doucet, A. A differential approach to the multi-marginal schrödinger system. *Preprint arXiv:2108.08129v1*, 2021.

Deng, W., Lin, G., and Liang, F. A Contour Stochastic Gradient Langevin Dynamics Algorithm for Simulations of Multi-modal Distributions. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Deng, W., Lin, G., and Liang, F. An Adaptively Weighted Stochastic Gradient MCMC Algorithm for Monte Carlo Simulation and Global Optimization. *Statistics and Computing*, pp. 32–58, 2022.

Eckstein, S. and Nutz, M. Quantitative Stability of Regularized Optimal Transport and Convergence of Sinkhorn's Algorithm. *SIAM Journal on Mathematical Analysis*, 54: 5922–5948, 2022.

Fishman, N., Klarner, L., Bortoli, V. D., Mathieu, E., and Hutchinson, M. Diffusion Models for Constrained Domains. In *arXiv:2304.05364v1*, 2023.

Franklin, J. and Lorenz, J. On the Scaling of Multidimensional Matrices. *Linear Algebra and its Applications*, 114/115:717–735, 1989.

Ghosal, P. and Nutz, M. On the Convergence Rate of Sinkhorn's Algorithm. *arXiv:2212.06000*, 2022.

Hutchinson, M. F. A Stochastic Estimator of the Trace of the Influence Matrix for Laplacian Smoothing Splines. *Communications in Statistics-Simulation and Computation*, 18(3):1059–1076, 1989.

Kullback, S. Probability Densities with Given Marginals. *Ann. Math. Statist.*, 1968.

Lavenant, H. and Santambrogio, F. The Flow Map of the Fokker–Planck Equation Does Not Provide Optimal Transport. *Applied Mathematics Letters*, 133, 2022.

Lee, H., Lu, J., and Tan, Y. Convergence for Score-based Generative Modeling with Polynomial Complexity. *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Léonard, C. Some Properties of Path Measures. *Séminaire de Probabilités XLVI*, pp. 207–230, 2014.

Lipman, Y., Chen, R. T. Q., Ben-Hamu, H., Nickel, M., and Le, M. Flow Matching for Generative Modeling. In *Proc. of the International Conference on Learning Representation (ICLR)*, 2023.

Liu, G.-H., Vahdat, A., Huang, D.-A., Theodorou, E. A., Nie, W., and Anandkumar, A. I²SB: Image-to-Image Schrödinger Bridge. In *Proc. of the International Conference on Machine Learning (ICML)*, 2023.

Liu, Q. Rectified Flow: A Marginal Preserving Approach to Optimal Transport. *arXiv:2209.14577*, 2022.

Marino, S. D. and Gerolin, A. An Optimal Transport Approach for the Schrödinger Bridge Problem and Convergence of Sinkhorn Algorithm. *Journal of Scientific Computing*, 85:2:27–28, 2020.

Nutz, M. Introduction to Entropic Optimal Transport. *Lecture Notes*, 2022.

Nutz, M. and Wiesel, J. Stability of Schrödinger Potentials and Convergence of Sinkhorn's Algorithm. *Annals of Probability*, 2022.

Peyré, G. and Cuturi, M. *Computational Optimal Transport: With Applications to Data Science*. Foundations and Trends in Machine Learning, 2019.

Pooladian, A.-A., Ben-Hamu, H., Domingo-Enrich, C., Amos, B., Lipman, Y., and Chen, R. T. Q. Multisample Flow Matching: Straightening Flows with Minibatch Couplings. In *Proc. of the International Conference on Machine Learning (ICML)*, 2023.

Ruschendorf, L. Convergence of the Iterative Proportional Fitting Procedure. *Annals of Statistics*, 1995.

Vargas, F., Thodoroff, P., Lamacraft, A., and Lawrence, N. Solving Schrödinger Bridges via Maximum Likelihood. *Entropy*, 23(9):1134, 2021.

A. Proof of Convergence with Bounded Cost

Next, we modify Algorithm 1 following the centering method developed in (Carlier, 2022),

Algorithm 2 Centered Sinkhorn. Set $\bar{\varphi}_0 := 0$. For $k \geq 0$, the iterate follows

$$\bar{\psi}_k(\mathbf{y}) := -\log \int_{\mathrm{X}} e^{\bar{\varphi}_k(\mathbf{x}) - c_\varepsilon(\mathbf{x},\mathbf{y})} \mu_{\star,k}(\mathrm{d}\mathbf{x}) \tag{19}$$

$$\bar{\varphi}_{k+1}(\mathbf{x}) := -\log \int_{\mathrm{Y}} e^{\bar{\psi}_k(\mathbf{y}) - c_\varepsilon(\mathbf{x},\mathbf{y})} \nu_{\star,k}(\mathrm{d}\mathbf{y}) + \lambda_k, \quad \text{where} \tag{20}$$

$$\lambda_k := \int_{\mathrm{X}} \log \left(\int_{\mathrm{Y}} e^{\bar{\psi}_k(\mathbf{y}) - c_\varepsilon(\mathbf{x},\mathbf{y})} \nu_{\star,k}(\mathrm{d}\mathbf{y}) \right) \mu_\star(\mathrm{d}\mathbf{x}).$$

The algorithm differs from Algorithm 1 in that an additional centering operation is included in the updates of $\bar{\varphi}_{k+1}$ to ensure $\mu_\star(\bar{\varphi}_{k+1}) = 0$. Notably, $\mu_\star$ is required for the centering operation to upper bound the divergence, although it is not directly accessible and no implementation is needed. The main contribution of the centering operation is that the two coordinates $(\bar{\varphi}, \bar{\psi})$ become separable in L^2

$$\|\bar{\varphi} \oplus \bar{\psi}\|^2_{L^2(\mu_\star \otimes \nu_\star)} = \|\bar{\varphi}\|^2_{L^2(\mu_\star)} + \|\bar{\psi}\|^2_{L^2(\nu_\star)} \quad \text{if} \quad \mu_\star(\bar{\varphi}) = 0. \tag{21}$$

The coordinate ascent is equivalent to the following updates

$$\bar{\psi}_k(y) = \operatorname*{arg\,max}_{\bar{\psi} \in L^1(\nu_\star)} G(\bar{\varphi}_k, \bar{\psi}), \quad \bar{\varphi}_k(y) = \operatorname*{arg\,max}_{\bar{\varphi} \in L^1(\mu_\star): \mu_\star(\bar{\varphi}) = 0} G(\bar{\varphi}, \bar{\psi}_k),$$

The relation between the Schrödinger potentials (φ_k, ψ_k) and centered Schrödinger potentials $(\bar{\varphi}_k, \bar{\psi}_k)$ is characterized as follows

Lemma 2. *Denote by (φ_k, ψ_k) the Sinkhorn iterates in Algorithm 1. For all $k \geq 0$, $\mu_\star(\varphi_k) = -(\lambda_0 + \cdots + \lambda_{k-1})$. Moreover, we have*

$$\bar{\varphi}_k = \varphi_k - \mu_\star(\varphi_k), \quad \bar{\psi}_k = \psi_k + \mu_\star(\psi_k). \tag{22}$$

In particular, $\bar{\varphi}_k \oplus \bar{\psi}_k = \varphi_k \oplus \psi_k$ and $G(\bar{\varphi}_k, \bar{\psi}_k) = G(\varphi_k, \psi_k)$.

Proof Applying the induction method completes the proof directly. □

Recall how $\bar{\psi}_k$ is defined through the Schrödinger equation

$$\text{The second marginal of } \pi_{2k}(\bar{\varphi}_k, \bar{\psi}_k) = e^{\bar{\varphi}_k \oplus \bar{\psi}_k - c_\varepsilon} \mathrm{d}(\mu_{\star,k} \otimes \nu_{\star,k}) \text{ is } \nu_{\star,k}, \tag{23}$$

as in Eq.(14). However, $\mathrm{d}\pi_{2k+1}(\bar{\varphi}_{k+1}, \bar{\psi}_k) = e^{\bar{\varphi}_{k+1} \oplus \bar{\psi}_k - c_\varepsilon} \mathrm{d}(\mu_{\star,k} \otimes \nu_{\star,k})$ fails to yield the first marginal $\mu_{\star,k}$ due to the centering constraint.

Next, we show the modified iterates are still bounded given the bounded cost function c.

Lemma 3. *For every $k \geq 0$, we have*

$$\|\bar{\varphi}_k\|_\infty \leq 2\|c\|_\infty, \quad \|\bar{\psi}_k\|_\infty \leq 3\|c_\varepsilon\|_\infty. \tag{24}$$

Proof Recall the definition of $\bar{\varphi}_{k+1}$ in Algorithm 2, we have $\forall \mathbf{x}_1, \mathbf{x}_2 \in \mathrm{X} \subset \mathbb{S}^d$,

$$\begin{aligned}
&\bar{\varphi}_{k+1}(\mathbf{x}_1) - \bar{\varphi}_{k+1}(\mathbf{x}_2) \\
&= \log \int_{\mathrm{Y}} e^{\bar{\psi}_k(\mathbf{y}) - c_\varepsilon(\mathbf{x}_2,\mathbf{y})} \nu_{\star,k}(\mathrm{d}\mathbf{y}) - \log \int_{\mathrm{X}} e^{\bar{\psi}_k(\mathbf{y}) - c_\varepsilon(\mathbf{x}_1,\mathbf{y})} \nu_{\star,k}(\mathrm{d}\mathbf{y}) \\
&\leq \log \left[e^{\sup_{\mathbf{y} \in \mathrm{Y}} |c_\varepsilon(\mathbf{x}_1,\mathbf{y}) - c_\varepsilon(\mathbf{x}_2,\mathbf{y})|} \int_{\mathrm{Y}} e^{\bar{\psi}_k(\mathbf{y}) - c_\varepsilon(\mathbf{x}_1,\mathbf{y})} \nu_{\star,k}(\mathrm{d}\mathbf{y}) \right] - \log \int_{\mathrm{Y}} e^{\bar{\psi}_k(\mathbf{y}) - c_\varepsilon(\mathbf{x}_1,\mathbf{y})} \nu_{\star,k}(\mathrm{d}\mathbf{y}) \\
&= \sup_{\mathbf{y} \in \mathrm{Y}} |c_\varepsilon(\mathbf{x}_1,\mathbf{y}) - c_\varepsilon(\mathbf{x}_2,\mathbf{y})| \leq 2\|c_\varepsilon\|_\infty.
\end{aligned} \tag{25}$$

As $\mu_\star(\bar{\varphi}_k) = 0$, we must have $\sup_x \bar{\varphi}_k(\mathbf{x}) \geq 0$ and $\inf_{\mathbf{x}} \bar{\varphi}_k(\mathbf{x}) \leq 0$, hence the above implies that $\|\bar{\varphi}_k\|_\infty \leq 2\|c\|_\infty$. The definition of $\bar{\psi}_k$ in Eq.(19) yields $\|\bar{\psi}_k\|_\infty \leq \|\bar{\varphi}_k\|_\infty + \|c_\varepsilon\|_\infty \leq 3\|c_\varepsilon\|_\infty$. □

The key to the proof is to adopt the strong convexity of the function e^x for $x \in [-\alpha, \infty)$ and some constant $\alpha \in \mathbb{R}$,

$$e^b - e^a \geq (b-a)e^a + \frac{e^{-\alpha}}{2}|b-a|^2 \quad \text{for } a, b \in [-\alpha, \infty). \tag{26}$$

We also present two supporting lemmas in order to complete the proof

Lemma 4. *Given $\varphi, \varphi' \in L^2(\mu_\star)$ and $\psi, \psi' \in L^2(\nu_\star)$, and define*

$$\begin{aligned} \partial_1 G(\varphi, \psi)(\mathbf{x}) &= 1 - \int_{\mathrm{Y}} e^{\varphi(\mathbf{x})+\psi(\mathbf{y})-c_\varepsilon(\mathbf{x},\mathbf{y})}\nu_\star(\mathrm{d}\mathbf{y}) \\ \partial_2 G(\varphi, \psi)(\mathbf{y}) &= 1 - \int_{\mathrm{X}} e^{\varphi(\mathbf{x})+\psi(\mathbf{y})-c_\varepsilon(\mathbf{x},\mathbf{y})}\mu_\star(\mathrm{d}\mathbf{x}). \end{aligned} \tag{27}$$

If both $\varphi \otimes \psi - c_\varepsilon \geq -\alpha$ and $\varphi' \oplus \psi' - c_\varepsilon \geq -\alpha$ for some $\alpha \in \mathbb{R}$, we have

$$\begin{aligned} G(\varphi', \psi') - G(\varphi, \psi) \geq & \int_{\mathrm{X}} \partial_1 G(\varphi', \psi')(\mathbf{x})[\varphi'(\mathbf{x}) - \varphi(\mathbf{x})]\mu_\star(\mathrm{d}\mathbf{x}) \\ & + \int_{\mathrm{Y}} \partial_2 G(\varphi', \psi')(\mathbf{y})[\psi'(\mathbf{y}) - \psi(\mathbf{y})]\nu_\star(\mathrm{d}\mathbf{y}) \\ & + \frac{e^{-\alpha}}{2}\|(\varphi - \varphi') \oplus (\psi - \psi')\|_{L^2(\mu_\star \otimes \nu_\star)}. \end{aligned}$$

Proof By Eq.(26), we have

$$\begin{aligned} & G(\varphi', \psi') - G(\varphi, \psi) \\ &= \mu_\star(\varphi' - \varphi) + \nu_\star(\psi' - \psi) + \iint_{\mathrm{X}\times\mathrm{Y}} (e^{\varphi\oplus\psi - c_\varepsilon} - e^{\varphi'\oplus\psi' - c_\varepsilon})\mathrm{d}(\mu_\star \otimes \nu_\star) \\ &\geq \mu_\star(\varphi' - \varphi) + \nu_\star(\psi' - \psi) + \iint_{\mathrm{X}\times\mathrm{Y}} (\varphi \oplus \psi - \varphi' \oplus \psi')e^{\varphi'\oplus\psi' - c_\varepsilon}\mathrm{d}(\mu_\star \otimes \nu_\star) \\ &\qquad + \frac{e^{-\alpha}}{2}\iint_{\mathrm{X}\times\mathrm{Y}} \|\varphi \oplus \psi - \varphi' \oplus \psi'\|_2^2 \mathrm{d}(\mu_\star \otimes \nu_\star) \\ &= \int_{\mathrm{X}} \partial_1 G(\varphi', \psi')(\mathbf{x})[\varphi'(x) - \varphi(\mathbf{x})]\mu_\star(\mathrm{d}\mathbf{x}) + \int_{\mathrm{Y}} \partial_2 G(\varphi', \psi')(\mathbf{y})[\psi'(\mathbf{y}) - \psi(\mathbf{y})]\nu_\star(\mathrm{d}\mathbf{y}) \\ &\quad + \frac{e^{-\alpha}}{2}\|(\varphi - \varphi') \oplus (\psi - \psi')\|_{L^2(\mu_\star \otimes \nu_\star)}. \end{aligned}$$

□

Lemma 5. $G(\bar{\varphi}_{k+1}, \bar{\psi}_{k+1}) - G(\bar{\varphi}_k, \bar{\psi}_k) \geq \frac{\sigma}{2}\left(\|\bar{\varphi}_{k+1} - \bar{\varphi}_k\|^2_{L^2(\mu_\star)} + \|\bar{\psi}_{k+1} - \bar{\psi}_k\|^2_{L^2(\nu_\star)}\right) - O(\epsilon)$, *where $\sigma := e^{-6\|c_\varepsilon\|_\infty}$; the big-O notation mainly depends on the smoothness A2 and tail properties A4.*

Proof We first decompose the LHS as follows

$$G(\bar{\varphi}_{k+1}, \bar{\psi}_{k+1}) - G(\bar{\varphi}_k, \bar{\psi}_k) = \underbrace{G(\bar{\varphi}_{k+1}, \bar{\psi}_{k+1}) - G(\bar{\varphi}_{k+1}, \bar{\psi}_k)}_{\mathrm{I}} + \underbrace{G(\bar{\varphi}_{k+1}, \bar{\psi}_k) - G(\bar{\varphi}_k, \bar{\psi}_k)}_{\mathrm{II}}.$$

For the estimate of I, by Lemma 4 with $\sigma = e^{-6\|c_\varepsilon\|_\infty}$, we have

$$\mathrm{I} \geq \int_{\mathrm{Y}} \partial_2 G(\bar{\varphi}_{k+1}, \bar{\psi}_{k+1})(\mathbf{y})[\bar{\psi}_{k+1}(\mathbf{y}) - \bar{\psi}_k(\mathbf{y})]\nu_\star(\mathrm{d}\mathbf{y}) + \frac{\sigma}{2}\|\bar{\psi}_k - \bar{\psi}_{k+1}\|_{L^2(\nu_\star)}.$$

For the integral above, by the definition of $\partial_2 G$ in Eq.(27), we have

$$\begin{aligned}
&\partial_2 G(\bar{\varphi}_{k+1}, \bar{\psi}_{k+1})(\mathbf{y})\nu_\star(\mathrm{d}\mathbf{y}) \\
&= \nu_\star(\mathrm{d}\mathbf{y}) - \int_{\mathrm{X}} e^{\bar{\varphi}_{k+1}(\mathbf{x}) + \bar{\psi}_{k+1}(\mathbf{y}) - c_\varepsilon(\mathbf{x},\mathbf{y})} \mu_\star(\mathrm{d}\mathbf{x})\nu_\star(\mathrm{d}\mathbf{y}) \\
&= \nu_\star(\mathrm{d}\mathbf{y}) - \int_{\mathrm{X}} \pi_{2k+2}(\mathrm{d}\mathbf{x}, \mathrm{d}\mathbf{y}) \frac{\mathrm{d}\mu_\star \otimes \mathrm{d}\nu_\star}{\mathrm{d}\mu_{\star,k+1} \otimes \mathrm{d}\nu_{\star,k+1}},
\end{aligned} \tag{28}$$

where the last equality follows by the LHS of Eq.(16).

Apply Lemma 7 with respect to $\frac{\mathrm{d}\mu_\star}{\mathrm{d}\mu_{\star,k+1}}(\mathbf{x})$

$$\begin{aligned}
\int_{\mathrm{X}} \pi_{2k+2}(\mathrm{d}\mathbf{x}, \mathrm{d}\mathbf{y}) \frac{\mathrm{d}\mu_\star \otimes \mathrm{d}\nu_\star}{\mathrm{d}\mu_{\star,k+1} \otimes \mathrm{d}\nu_{\star,k+1}} &\leq \int_{\mathrm{X}} \big(1 + O(\epsilon\|\mathbf{x}\|_2^2 + \epsilon)\big) \pi_{2k+2}(\mathrm{d}\mathbf{x}, \mathrm{d}\mathbf{y}) \frac{\nu_\star(\mathrm{d}\mathbf{y})}{\nu_{\star,k+1}(\mathrm{d}\mathbf{y})} \\
&\leq \big(1 + O(\epsilon)\big) \nu_{\star,k+1}(\mathrm{d}\mathbf{y}) \frac{\nu_\star(\mathrm{d}\mathbf{y})}{\nu_{\star,k+1}(\mathrm{d}\mathbf{y})} \\
&= \big(1 + O(\epsilon)\big) \nu_\star(\mathrm{d}\mathbf{y}),
\end{aligned} \tag{29}$$

where the second inequality holds by Lemma 8 and the fact that the second marginal of π_{2k+2} is $\nu_{\star,k+1}$ in Eq.(14). Similarly, we can show $\int_{\mathrm{X}} \pi_{2k+2}(\mathrm{d}\mathbf{x}, \mathrm{d}\mathbf{y}) \frac{\mathrm{d}\mu_\star \otimes \mathrm{d}\nu_\star}{\mathrm{d}\mu_{\star,k+1} \otimes \mathrm{d}\nu_{\star,k+1}} \gtrsim (1 - O(\epsilon))\nu_\star(\mathrm{d}\mathbf{y})$.

Combining Eq.(28) and (29), we have

$$|\partial_2 G(\bar{\varphi}_{k+1}, \bar{\psi}_{k+1})(\mathbf{y})\nu_\star(\mathrm{d}\mathbf{y})| \lesssim \epsilon \nu_\star(\mathrm{d}\mathbf{y}). \tag{30}$$

We now build the lower bound of the integral as follows

$$\begin{aligned}
&\int_{\mathrm{Y}} \partial_2 G(\bar{\varphi}_{k+1}, \bar{\psi}_{k+1})(\mathbf{y})[\bar{\psi}_{k+1}(\mathbf{y}) - \bar{\psi}_k(\mathbf{y})]\nu_\star(\mathrm{d}\mathbf{y}) \\
&\gtrsim -\epsilon \int_{\mathrm{Y}} \big|\bar{\psi}_{k+1}(\mathbf{y}) - \bar{\psi}_k(\mathbf{y})\big| \nu_\star(\mathrm{d}\mathbf{y}) \\
&\gtrsim -\epsilon,
\end{aligned} \tag{31}$$

where the first inequality follows by Eq.(28) and the second inequality follows by the boundedness of the potential function in Lemma 3. The above means that $\mathrm{I} \geq \frac{\sigma}{2}\|\bar{\psi}_k - \bar{\psi}_{k+1}\|_{L^2(\mu_\star\otimes\nu_\star)} - O(\epsilon)$. For the estimate of II, Lemma 4 yields

$$\mathrm{II} \geq \int_{\mathrm{X}} \partial_1 G(\bar{\varphi}_{k+1}, \bar{\psi}_k)(\mathbf{x})[\bar{\varphi}_{k+1}(\mathbf{x}) - \bar{\varphi}_k(\mathbf{x})]\mu_\star(\mathrm{d}\mathbf{x}) + \frac{\sigma}{2}\|\bar{\varphi}_k - \bar{\varphi}_{k+1}\|_{L^2(\mu_\star)}.$$

Recall the definition of $\bar{\varphi}_{k+1}$ in Eq.(20) states that $\int_{\mathrm{Y}} e^{\bar{\psi}_k(\mathbf{y}) - c_\varepsilon(\mathbf{x},\mathbf{y})}\nu_{\star,k}(\mathrm{d}\mathbf{y}) = e^{-\bar{\varphi}_{k+1}(\mathbf{x}) + \lambda_k}$. Apply Lemma 7 with respect to $\frac{\mathrm{d}\nu_\star}{\mathrm{d}\nu_{\star,k}}(\mathbf{x})$

$$\begin{aligned}
\partial_1 G(\bar{\varphi}_{k+1}, \bar{\psi}_k)(\mathbf{x}) &= 1 - e^{\bar{\varphi}_{k+1}(\mathbf{x})} \int_{\mathrm{Y}} e^{\bar{\psi}_k(\mathbf{y}) - c_\varepsilon(\mathbf{x},\mathbf{y})}\nu_{\star,k}(\mathrm{d}\mathbf{y}) \frac{\nu_\star(\mathrm{d}\mathbf{y})}{\nu_{\star,k}(\mathrm{d}\mathbf{y})} \\
&\geq 1 - e^{\bar{\varphi}_{k+1}(\mathbf{x})} \int_{\mathrm{Y}} \big(1 + O(\epsilon\|\mathbf{y}\|_2^2 + \epsilon)\big) e^{\bar{\psi}_k(\mathbf{y}) - c_\varepsilon(\mathbf{x},\mathbf{y})}\nu_{\star,k}(\mathrm{d}\mathbf{y}) \\
&\geq 1 - (1 + O(\epsilon))e^{\lambda_k} - O(\epsilon) \underbrace{\int_{\mathrm{Y}} \|\mathbf{y}\|_2^2 e^{\bar{\varphi}_{k+1}(\mathbf{x}) + \bar{\psi}_k(\mathbf{y}) - c_\varepsilon(\mathbf{x},\mathbf{y})}\nu_{\star,k}(\mathrm{d}\mathbf{y})}_{\text{a bounded non-negative function } R(\mathbf{x}) \text{ in Eq.(45)}} \\
&= 1 - (1 + O(\epsilon))e^{\lambda_k} - O(\epsilon)R(\mathbf{x}) \\
\partial_1 G(\bar{\varphi}_{k+1}, \bar{\psi}_k)(\mathbf{x}) &\leq 1 + (1 + O(\epsilon))e^{\lambda_k} + O(\epsilon)R(\mathbf{x}),
\end{aligned}$$

which includes a deterministic scalar (independent of $\mathbf{x}$) and a small perturbation (dependent of $\mathbf{x}$ and ϵ) and the last two inequalities follow by Lemma 9. Combining the centering operation with $\mu_\star(\bar{\varphi}_{k+1}) = \mu_\star(\bar{\varphi}_k) = 0$

$$
\begin{aligned}
&\int_{\mathrm{X}} \partial_1 G(\bar{\varphi}_{k+1}, \bar{\psi}_k)(\mathbf{x})[\bar{\varphi}_{k+1}(\mathbf{x}) - \bar{\varphi}_k(\mathbf{x})]\mu_\star(\mathrm{d}\mathbf{x}) \\
&= \text{deterministic scalar} \cdot \underbrace{\int_{\mathrm{X}} [\bar{\varphi}_{k+1}(\mathbf{x}) - \bar{\varphi}_k(\mathbf{x})]\mu_\star(\mathrm{d}\mathbf{x})}_{:=0 \text{ by the centering operation}} + \epsilon \underbrace{\int_{\mathrm{X}} R(\mathbf{x})[\bar{\varphi}_{k+1}(\mathbf{x}) - \bar{\varphi}_k(\mathbf{x})]\mu_\star(\mathrm{d}\mathbf{x})}_{\text{integrable by the boundedness of } R, \bar{\varphi}_{k+1}, \psi_k} \\
&= O(\epsilon).
\end{aligned} \tag{32}
$$

Combining the estimates of I and II completes the proof. $\square$

Now, we are ready to present an important result

Lemma 6. *Denoted by $(\bar{\varphi}_\star, \bar{\psi}_\star)$ the unique Schrödinger potentials with $\mu_\star(\bar{\varphi}_\star) = 0$. The iterates $(\bar{\varphi}_k, \bar{\psi}_k)_{k\geq 0}$ of Algorithm 2 given a bounded cost function c_ε satisfy*

$$
G(\bar{\varphi}_\star, \bar{\psi}_\star) - G(\bar{\varphi}_k, \bar{\psi}_k) \leq \beta_\varepsilon^k \big(G(\bar{\varphi}_\star, \bar{\psi}_\star) - G(\bar{\varphi}_0, \bar{\psi}_0)\big) + O(\epsilon), \tag{33}
$$

where $\beta_\varepsilon := 1 - e^{-24\|c_\varepsilon\|_\infty} \in (0, 1)$.

Proof

By Lemma 4 with $\alpha = 6\|c\|_\infty$ and the decomposition in Eq.(21), we have

$$
\begin{aligned}
G(\bar{\varphi}_k, \bar{\psi}_k) - G(\bar{\varphi}_\star, \bar{\psi}_\star) \geq & \int_{\mathrm{X}} \partial_1 G(\bar{\varphi}_k, \bar{\psi}_k)(\mathbf{x})[\bar{\varphi}_k(\mathbf{x}) - \bar{\varphi}_\star(\mathbf{x})]\mu_\star(\mathrm{d}\mathbf{x}) \\
& + \int_{\mathrm{Y}} \partial_2 G(\bar{\varphi}_k, \bar{\psi}_k)(\mathbf{y})[\bar{\psi}_k(\mathbf{y}) - \bar{\psi}_\star(\mathbf{y})]\nu_\star(\mathrm{d}\mathbf{y}) \\
& + \frac{\sigma}{2}\left(\|\bar{\varphi}_k - \bar{\varphi}_\star\|^2_{L^2(\mu_\star)} + \|\bar{\psi}_k - \bar{\psi}_\star\|^2_{L^2(\nu_\star)}\right) \\
\geq & \int_{\mathrm{X}} \partial_1 G(\bar{\varphi}_k, \bar{\psi}_k)(\mathbf{x})[\bar{\varphi}_k(\mathbf{x}) - \bar{\varphi}_\star(\mathbf{x})]\mu_\star(\mathrm{d}\mathbf{x}) + \frac{\sigma}{2}\|\bar{\varphi}_k - \bar{\varphi}_\star\|^2_{L^2(\mu_\star)} - O(\epsilon),
\end{aligned} \tag{34}
$$

where $\sigma := e^{-6\|c_\varepsilon\|_\infty}$, and the last inequality follows by Eq.(30) and boundedness of $\bar{\psi}_k$ and $\bar{\psi}_\star$ in Lemma 3. For the first integral, we note that as in Eq.(32), $\int_{\mathrm{X}} \partial_1 G(\bar{\varphi}_{k+1}, \bar{\psi}_k)(\mathbf{x})[\bar{\varphi}_k(\mathbf{x}) - \bar{\varphi}_\star(\mathbf{x})]\mu_\star(\mathrm{d}\mathbf{x}) = 0$ because $\partial_1 G(\bar{\varphi}_{k+1}, \bar{\psi}_k)(\mathbf{x})$ is deterministic and $\mu_\star(\bar{\varphi}_k(\mathbf{x}) = \mu_\star(\bar{\varphi}_\star(\mathbf{x})) = 0$.

Hence

$$
\begin{aligned}
&\int_{\mathrm{X}} \partial_1 G(\bar{\varphi}_k, \bar{\psi}_k)(\mathbf{x})[\bar{\varphi}_k(\mathbf{x}) - \bar{\varphi}_\star(\mathbf{x})]\mu_\star(\mathrm{d}\mathbf{x}) \\
&= \int_{\mathrm{X}} [\partial_1 G(\bar{\varphi}_k, \bar{\psi}_k)(\mathbf{x}) - \partial_1 G(\bar{\varphi}_{k+1}, \bar{\psi}_k)(\mathbf{x})][\bar{\varphi}_k(\mathbf{x}) - \bar{\varphi}_\star(\mathbf{x})]\mu_\star(\mathrm{d}\mathbf{x}) \\
&\geq -\frac{1}{2\sigma}\|\partial_1 G(\bar{\varphi}_k, \bar{\psi}_k) - \partial_1 G(\bar{\varphi}_{k+1}, \bar{\psi}_k)\|^2_{L^2(\mu_\star)} - \frac{\sigma}{2}\|\bar{\varphi}_k(\mathbf{x}) - \bar{\varphi}_\star(\mathbf{x})\|^2_{L^2(\mu_\star)},
\end{aligned} \tag{35}
$$

where the inequality follows from Hölder's inequality and Young's inequality.

Plugging Eq.(35) into Eq.(34), we have

$$
G(\bar{\varphi}_\star, \bar{\psi}_\star) - G(\bar{\varphi}_k, \bar{\psi}_k) \leq \frac{1}{2\sigma}\|\partial_1 G(\bar{\varphi}_k, \bar{\psi}_k) - \partial_1 G(\bar{\varphi}_{k+1}, \bar{\psi}_k)\|^2_{L^2(\mu_\star)} + O(\epsilon). \tag{36}
$$

Note that

$$
\begin{aligned}
|\partial_1 G(\bar{\varphi}_k, \bar{\psi}_k)(\mathbf{x}) - \partial_1 G(\bar{\varphi}_{k+1}, \bar{\psi}_k)(\mathbf{x})| &\leq \int_{\mathrm{Y}} \left| e^{\bar{\varphi}_{k+1} \oplus \bar{\psi}_k - c_\varepsilon} - e^{\bar{\varphi}_k \oplus \bar{\psi}_k - c_\varepsilon} \right| \nu_\star(\mathrm{d}\mathbf{y}) \\
&\leq e^{6\|c_\varepsilon\|_\infty} \int_{\mathrm{Y}} |\bar{\varphi}_{k+1} \oplus \bar{\psi}_k - \bar{\varphi}_k \oplus \bar{\psi}_k|\nu_\star(\mathrm{d}\mathbf{y}) \\
&= \frac{1}{\sigma}|\bar{\varphi}_{k+1}(\mathbf{x}) - \bar{\varphi}_k(\mathbf{x})|,
\end{aligned} \tag{37}
$$

where the second inequality follows by Lemma 3 and the exponential function follows a Lipschitz continuity such that: $e^a - e^b \le e^M |b-a|$ for $a, b \le M$; $\sigma := e^{-6\|c_\varepsilon\|_\infty}$.

First combining Eq.(36) and (37) and then including Lemma 5, we conclude that

$$\begin{aligned} G(\bar{\varphi}_\star, \bar{\psi}_\star) - G(\bar{\varphi}_k, \bar{\psi}_k) &\le \frac{1}{2\sigma^3} \|\bar{\varphi}_{k+1} - \bar{\varphi}_k\|^2_{L^2(\mu_\star)} + O(\epsilon) \\ &\le \frac{1}{\sigma^4} \left(G(\bar{\varphi}_{k+1}, \bar{\psi}_{k+1}) - G(\bar{\varphi}_k, \bar{\psi}_k) \right) + \frac{O(\epsilon)}{\sigma} \end{aligned}$$

where the last inequality follows by $\sigma \le 1$. Further writing $\Delta_k = G(\bar{\varphi}_\star, \bar{\psi}_\star) - G(\bar{\varphi}_k, \bar{\psi}_k)$, we have

$$\Delta_k \le \frac{1}{\sigma^4} \left(\Delta_k - \Delta_{k+1} \right) + \frac{O(\epsilon)}{\sigma}.$$

In other words, we can derive the contraction property as follows

$$\Delta_{k+1} \le (1-\sigma^4)\Delta_k + \sigma^3 \cdot O(\epsilon) \le \cdots \le (1-\sigma^4)^{k+1}\Delta_0 + \sum_{i=0}^{k} (1-\sigma^4)^i \sigma^3 \cdot O(\epsilon).$$

Denote $\beta_\varepsilon := 1 - e^{-24\|c_\varepsilon\|_\infty} \in (0,1)$. We hereby complete the first claim of the theorem for any $k \ge 1$

$$\Delta_k \le \beta_\varepsilon^k \Delta_0 + \sum_{i=0}^{k-1} \beta_\varepsilon^i \sigma^3 \cdot O(\epsilon) \le \beta_\varepsilon^k \Delta_0 + \underbrace{\frac{1+\beta_\varepsilon}{\beta_\varepsilon} \sum_{i=0}^{k} \beta_\varepsilon^i \sigma^3 O(\epsilon)}_{\text{denoted by } C \sum_{i=0}^{k} \beta_\varepsilon^i \epsilon}, \tag{38}$$

where $C = \frac{1+\beta_\varepsilon}{\beta_\varepsilon} \sigma^3$.

□

Since we are still interested in the convergence of the original (un-centered) Sinkhorn algorithm, now we extend the result to Theorem 1 and provide the proof as follows

Proof [Proof of Theorem 1]

As $G(\bar{\varphi}_k, \bar{\psi}_k) = G(\varphi_k, \psi_k)$ by Lemma 2, the convergence of (18) follows directly from (33).

□

A.1. Auxiliary Results

Lemma 7. *Given probability densities $\rho(\mathbf{x}) = e^{-U(\mathbf{x})}/C$ and $\widetilde{\rho}(\mathbf{x}) = e^{-\widetilde{U}(\mathbf{x})}/\widetilde{C}$ defined on $\mathbb{S}^d$, where C and $\widetilde{C}$ are the normalizing constants, the energy functions U and $\widetilde{U}$ are differentiable and satisfy*

$$\|\nabla \widetilde{U}(\mathbf{x}) - \nabla U(\mathbf{x})\|_2 \le \epsilon(1 + \|\mathbf{x}\|_2). \tag{39}$$

Moreover, U satisfies the smoothness assumption A2, then for small enough ϵ, we have

$$1 - O(\epsilon\|\mathbf{x}\|_2^2 + \epsilon) \le \frac{\rho(\mathbf{x})}{\widetilde{\rho}(\mathbf{x})} \le 1 + O(\epsilon\|\mathbf{x}\|_2^2 + \epsilon), \quad 1 - O(\epsilon\|\mathbf{x}\|_2^2 + \epsilon) \le \frac{\widetilde{\rho}(\mathbf{x})}{\rho(\mathbf{x})} \le 1 + O(\epsilon\|\mathbf{x}\|_2^2 + \epsilon). \tag{40}$$

Proof Assumption A2 implies $\|\nabla U(\mathbf{x})\|_2 \le \|\nabla U(\mathbf{x}) - \nabla U(\mathbf{0})\|_2 + \|\nabla U(\mathbf{0})\|_2 \le L\|\mathbf{x}\|_2 + \|\nabla U(\mathbf{0})\|_2$.

Note that for any $\mathbf{x}, \mathbf{y} \in \mathbb{S}^d$

$$U(\mathbf{x}) - U(\mathbf{y}) = \int_0^1 \frac{\mathrm{d}}{\mathrm{d}t} U(t\mathbf{x} + (1-t)\mathbf{y}) = \int_0^1 \langle \mathbf{x} - \mathbf{y}, \nabla U(t\mathbf{x} + (1-t)\mathbf{y}) \rangle \mathrm{d}t. \tag{41}$$

Moreover, there exist $\mathbf{x}_0$ such that $U(\mathbf{x}_0) = \widetilde{U}(\mathbf{x}_0)$ since ρ and $\widetilde{\rho}$ are probability densities. Applying Eq.(39) and Eq.(41), we have

$$
\begin{aligned}
|\widetilde{U}(\mathbf{x}) - U(\mathbf{x})| &= \left| \int_0^1 \langle \mathbf{x} - \mathbf{x}_0, \nabla \widetilde{U}(\cdot) - \nabla U(\cdot) \mathrm{d}t \right| \\
&\leq \int_0^1 \|\mathbf{x} - \mathbf{x}_0\|_2 \cdot \left\| \nabla \widetilde{U}(\cdot) - \nabla U(\cdot) \right\|_2 \mathrm{d}t \\
&\lesssim \epsilon(1 + \|\mathbf{x}\|_2)(\|\mathbf{x}\|_2 + \|\mathbf{x}_0\|_2) \lesssim \epsilon(\|\mathbf{x}\|_2^2 + 1),
\end{aligned}
\tag{42}
$$

where the second inequality holds by Eq.(39). As such, we have

$$
|\widetilde{C} - C| \leq \int_{\mathbb{S}^d} e^{-U(\mathbf{x})} \left| e^{-\widetilde{U}(\mathbf{x}) + U(\mathbf{x})} - 1 \right| \mathrm{d}\mathbf{x} \lesssim \epsilon \underbrace{\int_{\mathbb{S}^d} e^{-U(\mathbf{x})} (\|\mathbf{x}\|_2^2 + 1) \mathrm{d}\mathbf{x}}_{\text{integrable}},
\tag{43}
$$

where the last inequality follows by Eq.(42) and $e^a \leq 1 + 2a$ for $a \leq 1$. Eq.(40) follows by combining Eq.(42) and (43). $\square$

Lemma 8. *Suppose π_{2k} is a coupling in Eq.(16) with marginals $\mu_{\star,k}$ and $\nu_{\star,k}$. Given bounded cost function A1, the Lipschitz smoothness assumption A2, and the bounded fourth moment assumption A4, we have*

$$
\frac{\int_{\mathrm{X}} \|\mathbf{x}\|_2^2 \pi_{2k}(\mathrm{d}\mathbf{x}, \mathrm{d}\mathbf{y})}{\nu_{\star,k}(\mathrm{d}\mathbf{y})} < \infty.
\tag{44}
$$

Proof

By Eq.(16), we have

$$
\begin{aligned}
\frac{\int_{\mathrm{X}} \|\mathbf{x}\|_2^2 \pi_{2k}(\mathrm{d}\mathbf{x}, \mathrm{d}\mathbf{y})}{\nu_{\star,k}(\mathrm{d}\mathbf{y})} &= \int_{\mathrm{X}} \|\mathbf{x}\|_2^2 e^{\varphi_k \oplus \psi_k - c_\varepsilon} \mu_{\star,k}(\mathrm{d}\mathbf{x}) \\
&= \int_{\mathrm{X}} \|\mathbf{x}\|_2^2 e^{\bar{\varphi}_k \oplus \bar{\psi}_k - c_\varepsilon} \mu_{\star,k}(\mathrm{d}\mathbf{x}) \\
&\lesssim \int_{\mathrm{X}} \|\mathbf{x}\|_2^2 \mu_{\star,k}(\mathrm{d}\mathbf{x}) \\
&\lesssim \int_{\mathrm{X}} \|\mathbf{x}\|_2^2 \big(1 + O(\epsilon \|\mathbf{x}\|_2^2 + \epsilon)\big) \mu_\star(\mathrm{d}\mathbf{x}) \\
&< \infty,
\end{aligned}
$$

where the first inequality follows by the bounded cost function in Assumption A1 and the bounded potential function in Lemma 3; the second inequality follows by Lemma 7; the last inequality follows by assumption A4. $\square$.

Lemma 9. *Suppose π_{2k+1} is a coupling in Eq.(16) with marginals $\mu_{\star,k+1}$ and $\nu_{\star,k}$. Given bounded cost function A1, the Lipschitz smoothness assumption A2, and the bounded fourth moment assumption A4, we have*

$$
R(\mathbf{x}) := \int_{\mathrm{Y}} \|\mathbf{y}\|_2^2 e^{\bar{\varphi}_{k+1} \oplus \bar{\psi}_k - c_\varepsilon(\mathbf{x}, \mathbf{y})} \nu_{\star,k}(\mathrm{d}\mathbf{y}) < \infty,
\tag{45}
$$

where $R(\mathbf{x})$ is a bounded function.

Proof By a similar proof in Lemma 8 except the coupling is π_{2k+1} and the integral is w.r.t. Y, we have

$$
\frac{\int_{\mathrm{Y}} \|\mathbf{y}\|_2^2 \pi_{2k+1}(\mathrm{d}\mathbf{x}, \mathrm{d}\mathbf{y})}{\mu_{\star,k+1}(\mathrm{d}\mathbf{y})} < \infty.
$$

By Eq.(16), we have

$$
\int_{\mathrm{Y}} \|\mathbf{y}\|_2^2 e^{\bar{\varphi}_{k+1} \oplus \bar{\psi}_k - c_\varepsilon(\mathbf{x}, \mathbf{y})} \nu_{\star,k}(\mathrm{d}\mathbf{y}) = \frac{\int_{\mathrm{Y}} \|\mathbf{y}\|_2^2 \pi_{2k+1}(\mathrm{d}\mathbf{x}, \mathrm{d}\mathbf{y})}{\mu_{\star,k+1}(\mathrm{d}\mathbf{x})} < \infty.
$$

$\square$