

Plug-in estimation of Schrödinger bridges

Aram-Alexandre Pooladian* Jonathan Niles-Weed†

August 22, 2024

Abstract

We propose a procedure for estimating the Schrödinger bridge between two probability distributions. Unlike existing approaches, our method does not require iteratively simulating forward and backward diffusions or training neural networks to fit unknown drifts. Instead, we show that the potentials obtained from solving the static entropic optimal transport problem between the source and target samples can be modified to yield a natural plug-in estimator of the time-dependent drift that defines the bridge between two measures. Under minimal assumptions, we show that our proposal, which we call the *Sinkhorn bridge*, provably estimates the Schrödinger bridge with a rate of convergence that depends on the intrinsic dimensionality of the target measure. Our approach combines results from the areas of sampling, and theoretical and statistical entropic optimal transport.

1 Introduction

Modern statistical learning tasks often involve not merely the comparison of two unknown probability distributions but also the estimation of *transformations* from one distribution to another. Estimating such transformations is necessary when we want to generate new samples, infer trajectories, or track the evolution of particles in a dynamical system. In these applications, we want to know not only how “close” two distributions are, but also how to “go” between them.

Optimal transport theory defines objects that are well suited for both of these tasks (Villani, 2009; Santambrogio, 2015; Chewi et al., 2024). The 2-Wasserstein distance is a popular tool for comparing probability distributions for data analysis in statistics (Carlier et al., 2016; Chernozhukov et al., 2017; Ghosal and Sen, 2022), machine learning (Salimans et al., 2018), and the applied sciences (Bunne et al., 2023b;

*Center for Data Science, New York University. aram-alexandre.pooladian@nyu.edu

†Center for Data Science and Courant Institute for Mathematical Science, New York University. jnw@cims.nyu.edu

Manole et al., 2022). Under suitable conditions, the two probability measures that we want to compare (say, μ and ν) induce an optimal transport map: the uniquely defined vector-valued function which acts as a transport map¹ between μ and ν such that the distance traveled is minimal in the L^2 sense (Brenier, 1991). Despite being a central object in many applications, the optimal transport map is difficult to compute and suffers from poor statistical estimation guarantees in high dimensions; see Hütter and Rigollet (2021); Manole et al. (2021); Divil et al. (2022).

These drawbacks of the optimal transport map suggest that other approaches for defining a transport between two measures may often be more appropriate. For example, *flow based* or iterative approaches have recently begun to dominate in computational applications—these methods sacrifice the L^2 -optimality of the optimal transport map to place greater emphasis on the tractability of the resulting transport. The work of Chen et al. (2018) proposed continuous normalizing flows (CNFs), which use neural networks to model the vector field in an ordinary differential equation (ODE). This machinery was exploited by several groups simultaneously (Albergo and Vanden-Eijnden, 2022; Lipman et al., 2022; Liu et al., 2022b) for the purpose of developing tractable constructions of vector fields that satisfy the *continuity equation* (see Section 2.1.2 for a definition), and whose flow maps therefore yield valid transports between source and target measures.

An increasingly popular alternative method for iterative transport is based on the Fokker–Planck equation (see Section 2.1.3 for a definition). This formulation incorporates a diffusion term, and the resulting dynamics follow a *stochastic* differential equation (SDE). Though there exist many stochastic dynamics that give rise to valid transports, a canonical role is played by the *Schrödinger bridge* (SB). Just as the optimal transport map minimizes the L^2 distance in transporting between two distributions, the SB minimizes the *relative entropy* of the diffusion process, and therefore has an interpretation as the “simplest” stochastic process bridging the two distributions—indeed, the SB originates as a *Gedankenexperiment* (or “thought experiment”) of Erwin Schrödinger in modeling the large deviations of diffusing gasses (Schrödinger, 1932). There are many equivalent formulations of the SB problem (see Section 2), though for the purposes of transport, its most important property is that it gives rise to a pair of SDEs that interpolate between two measures μ and ν :

$$dX_t = b_t^*(X_t) dt + \sqrt{\varepsilon} dB_t, \quad X_0 \sim \mu, X_1 \sim \nu, \quad (1)$$

$$dY_t = d_t^*(Y_t) dt + \sqrt{\varepsilon} dB_t, \quad Y_0 \sim \nu, Y_1 \sim \mu, \quad (2)$$

where $\varepsilon > 0$ plays the role of thermal noise.² Concretely, (1) indicates that samples from ν can be obtained by drawing samples from μ and simulating an SDE with drift b_t^* , and (2) shows how this process can be performed in reverse. Though these

¹ T is a transport map between μ and ν if given a sample $X \sim \mu$, its image under T satisfies $T(X) \sim \nu$.

²We assume throughout our work that the reference process is Brownian motion with volatility ε ; see Section 2.2.

dynamics are of obvious use in generating samples, the difficulty lies in obtaining estimators for the drifts.

Nearly a century later, Schrödinger’s thought experiment has been brought to reality, having found applications in the generation of new images, protein structures, and more (Kawakita et al., 2022; Liu et al., 2022a; Nusken et al., 2022; Thornton et al., 2022; Shi et al., 2022; Lee et al., 2024). The foundation for these advances is the work of De Bortoli et al. (2021), who propose to train two neural networks to act as the forward and backward drifts, which are iteratively updated to ensure that each diffusion yields samples from the appropriate distribution. This is reminiscent of the iterative proportion fitting procedure of Fortet (1940), and can be interpreted as a version of Sinkhorn’s matrix-scaling algorithm (Sinkhorn, 1964; Cuturi, 2013) on path space.

While the framework of De Bortoli et al. (2021) is popular from a computational perspective, it is worth emphasizing that this method is relatively costly, as it necessitates the undesirable task of simulating an SDE at each training iteration. Moreover, despite the recent surge in applications, current methods do not come with statistical guarantees to quantify their performance. In short, existing work leaves open the problem of developing tractable, statistically rigorous estimators for the Schrödinger bridge.

Contributions

We propose and analyze a computationally efficient estimator of the Schrödinger bridge which we call the *Sinkhorn Bridge*. Our main insight is that it is possible to estimate the *time-dependent* drifts in (1) and (2) by solving a *single, static* entropic optimal transport problem between samples from the source and target measures. Our approach is to compute the potentials $(\hat{f}, \hat{g})$ obtained by running Sinkhorn’s algorithm on the data $X_1, \dots, X_m \sim \mu$ and $Y_1, \dots, Y_n \sim \nu$ and plug these estimates into a simple formula for the drifts. For example, in the forward case, our estimator reads

$$\hat{b}_t(z) := (1 - t)^{-1} \left(-z + \frac{\sum_{j=1}^n Y_j \exp((\hat{g}_j - \frac{1}{2(1-t)} \|z - Y_j\|^2)/\varepsilon)}{\sum_{j=1}^n \exp((\hat{g}_j - \frac{1}{2(1-t)} \|z - Y_j\|^2)/\varepsilon)} \right).$$

See Section 3.1 for a detailed motivation for the choice of $\hat{b}_t$. Once the estimated potential $\hat{g}$ is obtained from a single use of Sinkhorn’s algorithm on the source and target data at the beginning of the procedure, computing $\hat{b}_t(z)$ for any $z \in \mathbb{R}^d$ and any $t \in (0, 1)$ is trivial.

We show that the solution to a discretized SDE implemented with the estimated drift $\hat{b}_t$ closely tracks the law of the solution to (1) on the whole interval $[0, \tau]$, for any $\tau \in [0, 1)$. Indeed, writing $\mathbf{P}_{[0, \tau]}^*$ for the law of the process solving (1) on $[0, \tau]$ and $\hat{\mathbf{P}}_{[0, \tau]}$ for the law of the process obtained by initializing from a fresh sample $X_0 \sim \mu$

and solving a discrete-time SDE with drift $\hat{b}_t$, we prove bounds on the risk

$$\mathbb{E}[\text{TV}^2(\hat{\mathbf{P}}_{[0,\tau]}, \mathbf{P}_{[0,\tau]}^*)]$$

that imply that, for fixed $\varepsilon > 0$ and $\tau \in [0, 1)$, the Schrödinger bridge can be estimated at the *parametric* rate. Moreover, though it is well known that such bounds must diverge as $\varepsilon \rightarrow 0$ or $\tau \rightarrow 1$, we demonstrate that the rate of growth depends on the *intrinsic* dimension k of the target measure rather than the ambient dimension d . When $k \ll d$, this gives strong justification for the use of the Sinkhorn Bridge estimator in high-dimensional problems.

To give a particular example in a special case, our results provide novel estimation rates for the *Föllmer bridge*, an object which has also garnered interest in the machine learning community (Vargas et al., 2023; Chen et al., 2024b; Huang, 2024). In this setting, the source measure is a Dirac mass, and we suppose the target measure ν is supported on a ball of radius R contained within a k -dimensional smooth submanifold of $\mathbb{R}^d$. Taking the volatility level to be unity, we show that the Föllmer bridge up to time $\tau \in [0, 1)$ can be estimated in total variation with precision ϵ_{TV} using n samples and N SDE-discretization steps, where

$$n \asymp R^2(1 - \tau)^{-k-2}\epsilon_{\text{TV}}^{-2}, \quad N \lesssim dR^4(1 - \tau)^{-4}\epsilon_{\text{TV}}^{-2}.$$

As advertised, for fixed $\tau \in [0, 1)$, these bounds imply parametric scaling on the number of samples—which matches similar findings in the entropic optimal transport literature, see, e.g., Stromme (2023b)—and exhibit a “curse of dimensionality” only with respect to the *intrinsic* dimension of the target, k . As our main theorem shows, these phenomena are not unique to the Föllmer bridge, and hold for arbitrary volatility levels and general source measures. Moreover, by tuning τ appropriately, we show how these estimation results yield guarantees for sampling from the target measure ν , see Section 4.3. These guarantees also suffer only from a “curse of intrinsic dimensionality.” Since the drifts arising from the Föllmer bridge can be viewed as the score of a kernel density estimator of ν with a Gaussian kernel (see (27)), this benign dependence on the ambient dimension is a significant improvement over guarantees recently obtained for such estimators in the context of denoising diffusion probabilistic models (Wibisono et al., 2024). Our improved rates are due to the intimate connection between the SB problem and entropic optimal transport in which intrinsic dimensionality plays a crucial role (Groppe and Hundrieser, 2023; Stromme, 2023b). We expound on this connection in the main text.

We are not the first to notice the simple connection between the static entropic potentials and the SB drift. Finlay et al. (2020) first proposed to exploit this connection to simulate the SB by learning static potentials via a neural network-based implementation of Sinkhorn’s algorithm; however, due to some notational inaccuracies and implementation errors, the resulting procedure was not scalable. This work shows the theoretical soundness of their approach, with a much simpler, tractable algorithm and with rigorous statistical guarantees.

Outline. Section 2 contains the background information on both entropic optimal transport and the Schrödinger bridge problem, and unifies the notation between these two problems. Our proposed estimator, the Sinkhorn bridge, is described in Section 3, and Section 4 contains our main results and proof sketches, with the technical details deferred to the appendix. Simulations are performed in Section 5.

Notation

We denote the space of probability measures over $\mathbb{R}^d$ with finite second moment by $\mathcal{P}_2(\mathbb{R}^d)$. We write $B(x, R) \subseteq \mathbb{R}^d$ to indicate the (Euclidean) ball of radius $R > 0$ centered at $x \in \mathbb{R}^d$. We denote the maximum of a and b by $a \vee b$. We write $a \lesssim b$ (resp. $a \asymp b$) if there exists constants $C > 0$ (resp. $c, C > 0$) such that $a \leq Cb$ (resp. $cb \leq a \leq Cb$). We let $\mathbf{path} := \mathcal{C}([0, 1], \mathbb{R}^d)$ be the space of paths with $X_t : \mathbf{path} \rightarrow \mathbb{R}^d$ given by the canonical mapping $X_t(h) = h_t$ for any $h \in \mathbf{path}$ and any $t \in [0, 1]$. For a path measure $\mathbf{P} \in \mathcal{P}(\mathbf{path})$ and any $t \in [0, 1]$, we write $\mathbf{P}_t := (X_t)_\# \mathbf{P} \in \mathcal{P}(\mathbb{R}^d)$ for the t^{th} marginal of $\mathbf{P}$. Similarly, for $s, t \in [0, 1]$, we can define the joint probability measure $\mathbf{P}_{st} := (X_s, X_t)_\# \mathbf{P}$. We write $\mathbf{P}_{[0, t]}$ for the restriction of the $\mathbf{P}$ to $\mathcal{C}([0, t], \mathbb{R}^d)$. Since $\mathbf{path}$ is a Polish space, we can define regular conditional probabilities for the law of a path given its value at time t , which we denote $\mathbf{P}_{|t}$. For any $s > 0$, we write $\Lambda_s := (2\pi s)^{-d/2}$ for the normalizing constant of the density of the Gaussian distribution $\mathcal{N}(0, sI)$.

1.1 Related work

On Schrödinger bridges. The connection between entropic optimal transport and the Schrödinger bridge (SB) problem is well studied; see the comprehensive survey by Léonard (2013). We were also inspired by the works of Ripani (2019), Gentil et al. (2020), as well as Chen et al. (2016, 2021b) (which cover these topics from the perspective of optimal control), and the more recent article by Kato (2024) (which revisits the large-deviation perspective of this problem). The special case of the Föllmer bridge and its variants has been a topic of recent study in theoretical communities (Eldan et al., 2020; Mikulincer and Shenfeld, 2024).

Interest in computational methods for SBs has been explosive in over the last few years, see De Bortoli et al. (2021); Shi et al. (2022); Bunne et al. (2023a); Tong et al. (2023); Vargas et al. (2023); Yim et al. (2023); Chen et al. (2024b); Shi et al. (2024) for recent developments in deep learning. The works by Bernton et al. (2019); Pavon et al. (2021); Vargas et al. (2021) use more traditional statistical methods to estimate the SB, with various goals in mind. For example Bernton et al. (2019) propose a sampling scheme based on trajectory refinements using an approximate dynamic programming approach. Pavon et al. (2021) and Vargas et al. (2021) propose methods to compute the (intermediate) density directly based on maximum likelihood-type estimators: Pavon et al. (2021) directly model the densities of interest and devise a scheme to update the weights; Vargas et al. (2021) use Gaussian processes to model the forward and backward drifts, and update them via a maximum-likelihood type loss.

On entropic optimal transport. Our work is closely related to the growing literature on statistical entropic optimal transport, specifically on the developments surrounding the *entropic transport map*. This object was introduced by [Pooladian and Niles-Weed \(2021\)](#) as a computationally friendly estimator for optimal transport maps in the regime $\varepsilon \rightarrow 0$; see also [Pooladian et al. \(2023\)](#) for minimax estimation rates in the semi-discrete regime. When ε is fixed, the theoretical properties of the entropic maps have been analyzed ([Chiarini et al., 2022](#); [Conforti, 2022](#); [Chewi and Pooladian, 2023](#); [Conforti et al., 2023](#); [Divol et al., 2024](#)) as well as their statistical properties ([del Barrio et al., 2022](#); [Goldfeld et al., 2022b,a](#); [Gonzalez-Sanz et al., 2022](#); [Rigollet and Stromme, 2022](#); [Werenski et al., 2023](#)). [Nutz and Wiesel \(2021\)](#); [Ghosal et al. \(2022\)](#) study the stability of entropic potentials and plans in a qualitative sense under minimal regularity assumptions. Most recently, [Stromme \(2023b\)](#) and [Groppe and Hundrieser \(2023\)](#) established the connections between statistical entropic optimal transport and intrinsic dimensionality (for both maps and costs). [Daniels et al. \(2021\)](#) investigates sampling using entropic optimal transport couplings combined with neural networks. Closely related are the works by [Chizat et al. \(2022\)](#) and [Lavenant et al. \(2021\)](#), which highlight the use of entropic optimal transport for trajectory inference. A more flexible alternative to the entropic transport map was recently developed by [Kassraie et al. \(2024\)](#), who proposed a transport that progressively displaces the source measure to the target measure by computing a new entropic transport map at each step to approximate the *McCann interpolation* ([McCann, 1997](#)).

2 Background

2.1 Entropic optimal transport

2.1.1 Static formulation

Let $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$ and fix $\varepsilon > 0$. The *entropic optimal transport* problem between μ and ν is written as

$$\text{OT}_\varepsilon(\mu, \nu) := \inf_{\pi \in \Pi(\mu, \nu)} \iint \frac{1}{2} \|x - y\|^2 d\pi(x, y) + \varepsilon \text{KL}(\pi \| \mu \otimes \nu), \quad (3)$$

where $\Pi(\mu, \nu) \subseteq \mathcal{P}_2(\mathbb{R}^d \times \mathbb{R}^d)$ is the set of joint measures with left-marginal μ and right-marginal ν , called the set of *plans* or couplings, and where we define the Kullback–Leibler divergence as

$$\text{KL}(\pi \| \mu \otimes \nu) := \int \log \left(\frac{d\pi(x, y)}{d\mu(x) d\nu(y)} \right) d\pi(x, y),$$

whenever π admits a density with respect to $\mu \otimes \nu$, and $+\infty$ otherwise. Note that when $\varepsilon = 0$, (3) reduces to the 2-Wasserstein distance between μ and ν (see, e.g., [Villani \(2009\)](#); [Santambrogio \(2015\)](#)). The entropic optimal transport problem was

introduced to the machine learning community by [Cuturi \(2013\)](#) as a numerical scheme for approximating the 2-Wasserstein distance on the basis of samples.

Equation (3) is a strictly convex problem, and thus admits a unique minimizer, called the *optimal entropic plan*, written $\pi^* \in \Pi(\mu, \nu)$.³ Moreover, a dual formulation also exists (see [Genevay \(2019\)](#))

$$\text{OT}_\varepsilon(\mu, \nu) = \sup_{(f,g) \in \mathcal{F}} \Phi^{\mu\nu}(f, g) \quad (4)$$

where $\mathcal{F} = L^1(\mu) \times L^1(\nu)$ and

$$\Phi^{\mu\nu}(f, g) := \int f \, d\mu + \int g \, d\nu - \varepsilon \iint \left(\Lambda_\varepsilon e^{(f(x)+g(y)-\frac{1}{2}\|x-y\|^2)/\varepsilon} - 1 \right) d\mu(x) d\nu(y), \quad (5)$$

where we recall $\Lambda_\varepsilon = (2\pi\varepsilon)^{-d/2}$. Solutions are guaranteed to exist when $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$, and we call the dual optimizers the *optimal entropic (Kantorovich) potentials*, written (f^*, g^*) . [Csiszár \(1975\)](#) showed that the primal and dual optima are intimately connected through the following relationship.⁴

$$d\pi^*(x, y) = \Lambda_\varepsilon \exp\left(\frac{f^*(x) + g^*(y) - \frac{1}{2}\|x - y\|^2}{\varepsilon}\right) d\mu(x) d\nu(y). \quad (6)$$

Though f^* and g^* are *a priori* defined almost everywhere on the support of μ and ν , they can be extended to all of $\mathbb{R}^d$ (see [Mena and Niles-Weed \(2019\)](#); [Nutz and Wiesel \(2021\)](#)) via the optimality conditions

$$\begin{aligned} f^*(x) &= -\varepsilon \log \left(\Lambda_\varepsilon \int e^{(g^*(y) - \|x-y\|^2/2)/\varepsilon} d\nu(y) \right), \\ g^*(y) &= -\varepsilon \log \left(\Lambda_\varepsilon \int e^{(f^*(x) - \|x-y\|^2/2)/\varepsilon} d\mu(x) \right). \end{aligned}$$

At times, it will be convenient to work with *entropic Brenier potentials*, defined as

$$(\varphi^*, \psi^*) := \left(\frac{1}{2}\|\cdot\|^2 - f^*, \frac{1}{2}\|\cdot\|^2 - g^* \right).$$

Note that the gradients of the entropic Brenier potentials⁵ are related to barycentric projections of the optimal entropic coupling

$$\nabla\varphi^*(x) = \mathbb{E}_{\pi^*}[Y|X=x], \quad \nabla\psi^*(y) = \mathbb{E}_{\pi^*}[X|Y=y].$$

³Though π^* and the other objects discussed in this section depend on ε , we will omit this dependence for the sake of readability, though we will track the dependence on ε in our bounds.

⁴The normalization factor Λ_ε is not typically used in the computational optimal transport literature, but it simplifies some formulas in what follows. Since the procedure we propose is invariant under translation of the optimal entropic potentials, this normalization factor does not affect either our algorithm or its analysis.

⁵Passing the gradient under the integral is permitted via dominated convergence under suitable tail conditions on μ and ν .

See [Pooladian and Niles-Weed \(2021, Proposition 2\)](#) for a proof of this fact. By analogy with the unregularized optimal transport problem, these are called *entropic Brenier maps*. The following relationships can also be readily verified (see [Chewi and Pooladian \(2023, Lemma 1\)](#)):

$$\nabla^2 \varphi^*(x) = \varepsilon^{-1} \text{Cov}_{\pi^*}[Y|X=x], \quad \nabla^2 \psi^*(y) = \varepsilon^{-1} \text{Cov}_{\pi^*}[X|Y=y]. \quad (8)$$

2.1.2 A dynamic formulation via the continuity equation

Entropic optimal transport can also be understood from a dynamical perspective. Let $(\mathbf{p}_t)_{t \in [0,1]}$ be a family of measures in $\mathcal{P}_2(\mathbb{R}^d)$, and let $(v_t)_{t \in [0,1]}$ be a family of vector fields. We say that the pair satisfies the *continuity equation*, written $(\mathbf{p}_t, v_t) \in \mathfrak{C}$, if $\mathbf{p}_0 = \mu$, $\mathbf{p}_1 = \nu$, and, for $t \in [0, 1]$,

$$\partial_t \mathbf{p}_t + \nabla \cdot (v_t \mathbf{p}_t) = 0. \quad (9)$$

Solutions to (9) are understood to hold in the weak sense (that is, with respect to suitably smooth test functions).

The continuity equation can be viewed as the analogue of the marginal constraints being satisfied (i.e., the set $\Pi(\mu, \nu)$ above): it represents both the conservation of mass and the requisite end-point constraints for the path $(\mathbf{p}_t)_{t \in [0,1]}$. With this, we can cite a clean expression of the dynamic formulation of (3) (see [Conforti and Tamanini \(2021\)](#) or [Chizat et al. \(2020\)](#)) if μ and ν are absolutely continuous and have finite entropy:

$$\text{OT}_\varepsilon(\mu, \nu) + C_0(\varepsilon, \mu, \nu) = \inf_{(\mathbf{p}_t, v_t) \in \mathfrak{C}} \int_0^1 \int \left(\frac{1}{2} \|v_t(x)\|^2 + \frac{\varepsilon^2}{8} \|\nabla \log \mathbf{p}_t(x)\|^2 \right) d\mathbf{p}_t(x) dt, \quad (10)$$

where $C_0(\varepsilon, \mu, \nu) := \varepsilon \log(\Lambda_\varepsilon) + \frac{\varepsilon}{2}(H(\mu) + H(\nu))$ is an additive constant, with $H(\mu) := \int \log(d\mu) d\mu$, similarly for $H(\nu)$.

The case $\varepsilon = 0$ reduces to the celebrated Benamou–Brenier formulation of optimal transport ([Benamou and Brenier, 2000](#)).

2.1.3 A stochastic formulation via the Fokker–Planck equation

Yet another formulation of the dynamic problem exists, this time based on the *Fokker–Planck equation*, which is said to be satisfied by a pair $(\mathbf{p}_t, b_t) \in \mathfrak{F}$ if $\mathbf{p}_0 = \mu$, $\mathbf{p}_1 = \nu$, and, for $t \in [0, 1]$,

$$\partial_t \mathbf{p}_t + \nabla \cdot (b_t \mathbf{p}_t) = \frac{\varepsilon}{2} \Delta \mathbf{p}_t.$$

Then, under the same conditions as above,

$$\text{OT}_\varepsilon(\mu, \nu) + C_1(\varepsilon, \mu) = \inf_{(\mathbf{p}_t, b_t)} \int_0^1 \int \frac{1}{2} \|b_t(x)\|^2 d\mathbf{p}_t(x) dt, \quad (11)$$

where $C_1(\varepsilon, \mu) = \varepsilon \log(\Lambda_\varepsilon) + \varepsilon H(\mu)$. The equivalence between the objective functions (10) and (11), as well as the continuity equation and Fokker–Planck equations, is classical. For completeness, we provide details of these computations in Appendix A. A key property of this equivalence is the following relationship which relates the optimizers of (10), written $(\mathbf{p}_t^*, v_t^*)$ and (11), written $(\mathbf{p}_t^*, b_t^*)$:

$$b_t^* = v_t^* + \frac{\varepsilon}{2} \nabla \log \mathbf{p}_t^*.$$

We stress that the minimizer $\mathbf{p}_t^*$ is the same for both (10) and (11).

2.2 The Schrödinger Bridge problem

We will now briefly develop the required machinery to understand the Schrödinger bridge problem. We will largely following the expositions of Léonard (2012, 2013); Ripani (2019); Gentil et al. (2020).

For $\varepsilon > 0$, we let $\mathbf{R} \in \mathcal{P}(\text{path})$ denote the law of the reversible Brownian motion on $\mathbb{R}^d$ with volatility ε , with the Lebesgue measure as the initial distribution.⁶ We write the joint distribution of the initial and final positions under $\mathbf{R}$ by $\mathbf{R}_{01}(\mathrm{d}x, \mathrm{d}y) = \Lambda_\varepsilon \exp(-\frac{1}{2}\|x - y\|^2/\varepsilon) \mathrm{d}x \mathrm{d}y$.

With the above, we arrive at Schrödinger’s bridge problem over path measures:

$$\min_{\mathbf{P} \in \mathcal{P}(\text{path})} \varepsilon \text{KL}(\mathbf{P} \parallel \mathbf{R}) \quad \text{s.t.} \quad \mathbf{P}_0 = \mu, \mathbf{P}_1 = \nu, \quad (12)$$

where $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$ and are absolutely continuous with finite entropy. Let $\mathbf{P}^*$ be the unique solution to (12), which exists as the problem is strictly convex. Léonard (2013) shows that there exist two non-negative functions $\mathbf{f}^*, \mathbf{g}^* : \mathbb{R}^d \rightarrow \mathbb{R}_+$ such that

$$\mathbf{P}^* = \mathbf{f}^*(X_0) \mathbf{g}^*(X_1) \mathbf{R}, \quad (13)$$

where $\text{Law}(X_0) = \mu$ and $\text{Law}(X_1) = \nu$.

A further connection can be made: if we apply the chain-rule for the KL divergence by conditioning on times $t = 0, 1$, the objective function (12) decomposes into

$$\varepsilon \text{KL}(\mathbf{P} \parallel \mathbf{R}) = \varepsilon \text{KL}(\mathbf{P}_{01} \parallel \mathbf{R}_{01}) + \varepsilon \mathbb{E}_{\mathbf{P}} \text{KL}(\mathbf{P}_{|01} \parallel \mathbf{R}_{|01}).$$

Under the assumption that μ and ν have finite entropy, it can be shown that the first term on the right-hand side is equivalent to the objective for the entropic optimal transport problem in (4). Moreover, the second term vanishes if we choose the measure $\mathbf{P}$ so that the conditional measure $\mathbf{P}_{|01}$ is the same as $\mathbf{R}_{|01}$, i.e., a Brownian bridge. Therefore, the objective function in (12) is minimized when $\mathbf{P}_{01}^* = \pi^*$ and when $\mathbf{P}$

⁶The problem below remains well posed even though $\mathbf{R}$ is not a probability measure; see Léonard (2013) for complete discussions.

writes as a mixture of Brownian bridges with the distribution of initial and final points given by π^* :

$$\mathbf{P}^* = \iint \mathbf{R}(\cdot | X_0 = x_0, X_1 = x_1) \pi^*(dx_0, dx_1). \quad (14)$$

Much of the discussion above assumed that μ and ν are absolutely continuous with finite entropy; indeed, the manipulations in this section as well as in Sections 2.1.2 and 2.1.3 are not justified if this condition fails. Though the finite entropy conditioned is adopted liberally in the literature on Schrödinger bridges, in this work we will have to consider bridges between measures that may not be absolutely continuous (for example, empirical measures). Noting that the entropic optimal transport problem (3) has a unique solution for *any* $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$, we leverage this fact to use (14) as the *definition* of the Schrödinger bridge between two probability measures: for any pair of probability distributions $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$, their *Schrödinger bridge* is the mixture of Brownian bridges given by (14), where π^* is the solution to the entropic optimal transport problem (3).

3 Proposed estimator: The Sinkhorn bridge

Our goal is to efficiently estimate the Schrödinger bridge (SB) on the basis of samples. Let $\mathbf{P}^*$ denote the SB between μ and ν , and define the the time-marginal flow of the bridge by

$$\mathbf{p}_t^* := \mathbf{P}_t^*, \quad t \in [0, 1]. \quad (15)$$

This choice of notation is deliberate: when μ and ν have finite entropy, the t -marginals of $\mathbf{P}^*$ for $t \in [0, 1]$ solve the dynamic formulations (10) and (11) (Léonard, 2013, Proposition 4.1). In the existing literature, $\mathbf{p}_t^*$ is sometimes called the the *entropic interpolation* between μ and ν . See Léonard (2012, 2013); Ripani (2019); Gentil et al. (2020) for interesting properties of entropic interpolations (for example, their relation to functional inequalities). Our goal is to provide an estimator $\hat{\mathbf{P}}$ such that $\mathbb{E}[\text{TV}^2(\hat{\mathbf{P}}_{[0, \tau]}, \mathbf{P}_{[0, \tau]}^*)]$ is small for all $\tau < 1$. In particular, this marginals of our estimator $\hat{\mathbf{P}}$ are estimators $\hat{\mathbf{p}}_t$ of $\mathbf{p}_t^*$ for all $t \in [0, 1)$.⁷

We call our estimator the *Sinkhorn bridge*, and we outline its construction below. Our main observation involves revisiting some finer properties of entropic interpolations as a function of the static entropic potentials. Once everything is concretely expressed, a natural plug-in estimator will arise which is amenable to both computational and statistical considerations.

⁷For reasons that will be apparent in the next section, time $\tau = 1$ must be excluded from the analysis.

3.1 From Schrödinger to Sinkhorn and back

We outline two crucial observations from which our estimator naturally arises. First, we note that $\mathbf{p}_t^*$ can be explicitly expressed as the following density (Léonard, 2013, Theorem 3.4)

$$\mathbf{p}_t^*(dz) := \mathcal{H}_{(1-t)\varepsilon}[\exp(g^*/\varepsilon)\nu](z)\mathcal{H}_{t\varepsilon}[\exp(f^*/\varepsilon)\mu](z) dz, \quad (16)$$

where $\mathcal{H}_s$ is the *heat semigroup*, which acts on a measure Q via

$$Q \mapsto \mathcal{H}_s[Q](z) := \Lambda_s \int e^{-\frac{1}{2s}\|x-z\|^2} Q(dx).$$

This expression for the marginal of distribution $\mathbf{p}_t^*$ follows directly from (14):

$$\begin{aligned} \mathbf{p}_t^*(z) &:= \iint \mathbf{R}_t(z|X_0 = x_0, X_1 = x_1)\pi^*(dx_0, dx_1) \\ &= \iint \mathcal{N}(z|ty + (1-t)x, t(1-t)\varepsilon)\pi^*(dx, dy) \\ &= \Lambda_\varepsilon \iint e^{((f^*(x)+g^*(y)-\frac{1}{2}\|x-y\|^2)/\varepsilon)} \mathcal{N}(z|ty + (1-t)x, t(1-t)\varepsilon)\mu(dx)\nu(dy) \\ &= \int e^{g^*(y)/\varepsilon} \mathcal{N}(z|y, (1-t)\varepsilon)\nu(dy) \int e^{f^*(x)/\varepsilon} \mathcal{N}(z|x, t\varepsilon)\mu(dx) \\ &= \mathcal{H}_{1-t}[\exp(g^*/\varepsilon)\nu](z)\mathcal{H}_t[\exp(f^*/\varepsilon)\mu](z) \end{aligned}$$

where throughout we use $\mathcal{N}(z|m, \sigma^2)$ to denote the Gaussian density with mean m and covariance $\sigma^2 I$, and the fourth equality follows from computing the explicit density of the product of two Gaussians.

Also, Léonard (2013, Proposition 4.1) shows that when μ and ν have finite entropy, the optimal drift in (11) is given by

$$b_t^*(z) = \varepsilon \nabla \log \mathcal{H}_{(1-t)\varepsilon}[\exp(g^*/\varepsilon)\nu](z),$$

whence the pair $(\mathbf{p}_t^*, b_t^*)$ satisfies the Fokker–Planck equation. This fact implies that if X_t solves

$$dX_t = b_t^*(X_t) dt + \sqrt{\varepsilon} dB_t, \quad X_0 \sim \mu, \quad (17)$$

then $\mathbf{p}_t^* = \text{Law}(X_t)$. In fact, more is true: the SDE (17) give rise to a path measure, which exactly agrees with the Schrödinger bridge. Though Léonard (2013) derives these facts for μ and ν with finite entropy, we show in Proposition 3.1, below, that they hold in more generality.

Further developing the expression for b_t^* , we obtain

$$\begin{aligned} b_t^*(z) &= (1-t)^{-1} \left(-z + \frac{\int y e^{(g^*(y)-\frac{1}{2(1-t)}\|z-y\|^2)/\varepsilon} d\nu(y)}{\int e^{(g^*(y)-\frac{1}{2(1-t)}\|z-y\|^2)/\varepsilon} d\nu(y)} \right) \\ &=: (1-t)^{-1} (-z + \nabla \varphi_{1-t}^*(z)). \end{aligned} \quad (18)$$

Thus, our final expression for the SDE that yields the Schrödinger bridge is

$$dX_t = (-(1-t)^{-1}X_t + (1-t)^{-1}\nabla\varphi_{1-t}^*(X_t))dt + \sqrt{\varepsilon}dB_t. \quad (19)$$

Once again, we emphasize that our choice of notation here is deliberate: the drift is expressed as a function of a particular entropic Brenier map, namely, the entropic Brenier map between $\mathbf{p}_t^*$ and ν with regularization parameter $(1-t)\varepsilon$.

We summarize this collection of crucial properties in the following proposition; see Appendix A.2 for proofs. We note that this result avoids the finite entropy requirements of analogous results in the literature (Léonard, 2013; Shi et al., 2024).

Proposition 3.1. *Let π be a probability measure of the form*

$$\pi(dx_0, dx_1) = \Lambda_\varepsilon \exp((f(x_0) + g(x_1) - \frac{1}{2}\|x_0 - x_1\|^2)/\varepsilon) \mu_0(dx_0) \mu_1(dx_1), \quad (20)$$

for any measurable f and g and any probability measures $\mu_0, \mu_1 \in \mathcal{P}_2(\mathbb{R}^d)$. Let $\mathbf{M}$ the path measure given by a mixture of Brownian bridges with respect to (20) as in (14), with t -marginals $\mathbf{m}_t$ for $t \in [0, 1]$. The following hold:

1. The path measure $\mathbf{M}$ is Markov;
2. The marginal $\mathbf{m}_t$ is given by

$$\mathbf{m}_t(dz) = \mathcal{H}_{(1-t)\varepsilon}[\exp(g/\varepsilon)\mu_1](z) \mathcal{H}_{t\varepsilon}[\exp(f/\varepsilon)\mu_0](z) dz;$$

3. $\mathbf{M}$ is the law of the solution to the SDE

$$dX_t = \varepsilon \nabla \log \mathcal{H}_{(1-t)\varepsilon}[\exp(g/\varepsilon)\mu_1](X_t) dt + \sqrt{\varepsilon} dB_t, \quad X_0 \sim \mu_0;$$

4. The drift above can be expressed as $b_t(z) = (1-t)^{-1}(z - \nabla\varphi_{1-t}(z))$, where $\nabla\varphi_{1-t}$ is the entropic Brenier map between $\mathbf{m}_t$ and ρ with regularization strength $(1-t)\varepsilon$, where

$$\rho(dx_1) = \mu_1(dx_1) \exp(g(x_1)/\varepsilon + \log \mathcal{H}_\varepsilon[e^{f/\varepsilon}\mu_0](x_1)).$$

If (20) is the optimal entropic coupling between μ_0 and μ_1 , then $\rho \equiv \mu_1$.

3.2 Defining the estimator

In light of (18), it is easy to define an estimator on the basis of samples. Let $X_1, \dots, X_m \sim \mu$ and $Y_1, \dots, Y_n \sim \nu$, and let $\mu_m := m^{-1} \sum_{i=1}^m \delta_{X_i}$, and similarly $\nu_n := n^{-1} \sum_{j=1}^n \delta_{Y_j}$. Let $(\hat{f}, \hat{g}) \in \mathbb{R}^m \times \mathbb{R}^n$ be the optimal entropic potentials associated with $\text{OT}_\varepsilon(\mu_m, \nu_n)$, which can be computed efficiently via Sinkhorn's algorithm (Cuturi,

2013; Peyré and Cuturi, 2019) with a runtime of $O(mn/\varepsilon)$ (Altschuler et al., 2017). A natural plug-in estimator for the optimal drift is thus

$$\begin{aligned}\hat{b}_t(z) &:= \varepsilon \nabla \log \mathcal{H}_{(1-t)\varepsilon}[\exp(\hat{g}/\varepsilon)\nu_n] \\ &= (1-t)^{-1} \left(-z + \frac{\sum_{j=1}^n Y_j \exp((\hat{g}_j - \frac{1}{2(1-t)}\|z - Y_j\|^2)/\varepsilon)}{\sum_{j=1}^n \exp((\hat{g}_j - \frac{1}{2(1-t)}\|z - Y_j\|^2)/\varepsilon)} \right) \\ &=: (1-t)^{-1}(-z + \nabla \hat{\varphi}_{1-t}(z))\end{aligned}\tag{21}$$

Further discussions on the numerical aspects of our estimator are deferred to Section 5. Since we want to estimate the path given by $\mathbf{P}^*$, our estimator is given by the solution to the following SDE:

$$d\hat{X}_t = (-(1-k\eta)^{-1}\hat{X}_{k\eta} + (1-k\eta)^{-1}\nabla \hat{\varphi}_{1-k\eta}(\hat{X}_{k\eta})) dt + \sqrt{\varepsilon} dB_t, \tag{22}$$

for $t \in [k\eta, (k+1)\eta]$, where $\eta \in (0, 1)$ is some step-size, and k is the iteration number. Though it is convenient to write the drift in terms of a time-varying entropic Brenier map, (21) shows that for all $t \in (0, 1)$, our estimator is a simple function of the potential $\hat{g}$ obtained from a single call to Sinkhorn’s algorithm.

Remark 3.2. To the best of our knowledge, the idea of using static potentials to estimate the SB drift was first explored by Finlay et al. (2020). However, their proposal had some inconsistencies. For example, they assume a finite entropy condition on the source and target measures, and perform a standard Gaussian convolution on $\mathbb{R}^d$ instead of our proposed convolution $\mathcal{H}_{(1-t)\varepsilon}[\exp(\hat{g}/\varepsilon)\nu_n]$. The former leads to a computationally intractable estimator, whereas, as we have shown above, the former has a simple form that is trivial to compute.

Remark 3.3. An alternative approach to computing the Schrödinger bridge is due to Stromme (2023a): Given n samples from the source and target measure, one can efficiently compute the in-sample entropic optimal coupling $\hat{\pi}$ on the basis of samples via Sinkhorn’s algorithm. Resampling a pair $(X', Y') \sim \hat{\pi}$ and computing the Brownian bridge between X' and Y' yields an approximate sample from the Schrödinger bridge. We remark that the computational complexity of our approach is significantly lower than that of Stromme (2023a). While both methods use Sinkhorn’s algorithm to compute an entropic optimal coupling between the source and target measures, Stromme’s estimator necessitates n *fresh* samples from μ and ν to obtain a single approximate sample from the SB. By contrast, having used our method to estimate the drifts, fresh samples from μ can be used to generate unlimited approximate samples from the SB.

4 Main results and proof sketch

We now present the proof sketches to our main result. We first present a sketch focusing purely on the statistical error incurred by our estimator, and later, using

standard tools (Chen et al., 2022; Lee et al., 2023), we incorporate the additional time-discretization error. All omitted proofs in this section are deferred to Appendix B.

4.1 Statistical analysis

We restrict our analysis to the one-sample estimation task, as it is the closest to real-world applications where the source measure is typically known (e.g., the standard Gaussian) and the practitioner is given finitely many samples from a distribution of interest (e.g., images). Thus, we assume full access to μ and access to ν through i.i.d. data, and let $(\hat{f}, \hat{g})$ correspond to the optimal entropic potentials solving $\text{OT}_\varepsilon(\mu, \nu_n)$, which give rise to an optimal entropic plan π_n . Formally, this corresponds to the $m \rightarrow \infty$ limit of the setting described in Section 3.2; the estimator for the drift (21) is unchanged.

Let $\tilde{\mathbf{P}}$ be the Markov measure associated with the mixture of Brownian bridges defined with respect to π_n . By Proposition 3.1, the t -marginals are given by

$$\tilde{\mathbf{p}}_t(z) = \mathcal{H}_{(1-t)\varepsilon}[\exp(\hat{g}/\varepsilon)\nu_n](z)\mathcal{H}_{t\varepsilon}[\exp(\hat{f}/\varepsilon)\mu](z), \quad (23)$$

and the one-sample empirical drift is equal to

$$\hat{b}_t(z) = \varepsilon \nabla \log \mathcal{H}_{(1-t)\varepsilon}[\exp(\hat{g}/\varepsilon)\nu_n](z).$$

Thus, $\tilde{\mathbf{P}}$ is the law of the following process with $\tilde{X}_0 \sim \mu$

$$d\tilde{X}_t = \hat{b}_t(\tilde{X}_t) dt + \sqrt{\varepsilon} dB_t. \quad (24)$$

Note that this agrees with our estimator in (22), but without discretization. This process is not technically implementable, but forms an important theoretical tool in our analysis.

Our main result of this section is the following theorem.

Theorem 4.1 (One-sample estimation; no discretization). *Suppose both $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$, and ν is supported on a k -dimensional smooth submanifold of $\mathbb{R}^d$ whose support is contained in a ball of radius $R > 0$. Let $\tilde{\mathbf{P}}$ (resp. $\mathbf{P}$) be the path measure corresponding to (24) (resp. (18)). Then it holds that, for any $\tau \in [0, 1)$,*

$$\mathbb{E}[\text{TV}^2(\tilde{\mathbf{P}}_{[0,\tau]}, \mathbf{P}_{[0,\tau]}^*)] \lesssim \left(\frac{\varepsilon^{-k/2-1}}{\sqrt{n}} + \frac{R^2 \varepsilon^{-k}}{(1-\tau)^{k+2}n} \right).$$

As mentioned in the introduction, the parametric rates will not be surprising given the proof sketch below, which incorporates ideas from entropic optimal transport. The rates diverge exponentially in k as $\tau \rightarrow 1$; this is a consequence of the fact that the estimated drift $\hat{b}_t$ enforces that the samples exactly collapse onto the training data at terminal time, which is far from the true target measure.

The proof of Theorem 4.1 uses key ideas from Stromme (2023b): We introduce the following entropic plan

$$\bar{\pi}_n(x, y) := \Lambda_\varepsilon \exp((\bar{f}(x) + g^*(y) - \tfrac{1}{2}\|x - y\|^2)/\varepsilon) \mu(dx) \nu_n(dy), \quad (25)$$

where g^* is the optimal entropic potential for the population measures (μ, ν) , and where we call $\bar{f} : \mathbb{R}^d \rightarrow \mathbb{R}$ a *rounded* potential, defined as

$$\bar{f}(x) := -\varepsilon \log \left(\Lambda_\varepsilon \cdot n^{-1} \sum_{j=1}^n \exp((g^*(Y_j) - \tfrac{1}{2}\|x - Y_j\|^2)/\varepsilon) \right).$$

Note that $\bar{f}$ can be viewed as the Sinkhorn update involving the potential g^* and measure ν_n , and that $\bar{\pi}_n \in \Gamma(\mu, \bar{\nu}_n)$, where $\bar{\nu}_n$ is a rescaled version of ν_n . We again exploit Proposition 3.1. Consider the path measure associated to the mixture of Brownian bridges with respect to $\bar{\pi}_n$, denoted $\bar{\mathbf{P}}$ (with t -marginals $\bar{\mathbf{p}}_t$), which corresponds to an SDE with drift

$$\begin{aligned} \bar{b}_t(z) &= \varepsilon \nabla \log \mathcal{H}_{1-t}[\exp(g^*/\varepsilon) \nu_n](z) \\ &= (1-t)^{-1} \left(-z + \frac{\sum_{j=1}^N Y_j \exp((g^*(Y_j) + \frac{1}{2(1-t)}\|z - Y_j\|^2)/\varepsilon)}{\sum_{j=1}^N \exp((g^*(Y_j) + \frac{1}{2(1-t)}\|z - Y_j\|^2)/\varepsilon)} \right). \end{aligned} \quad (26)$$

Introducing the path measure $\bar{\mathbf{P}}_{[0,\tau]}$ into the bound via triangle inequality and then applying Pinsker's inequality, we arrive at

$$\begin{aligned} \mathbb{E}[\text{TV}^2(\tilde{\mathbf{P}}_{[0,\tau]}, \mathbf{P}_{[0,\tau]}^*)] &\lesssim \mathbb{E}[\text{TV}^2(\tilde{\mathbf{P}}_{[0,\tau]}, \bar{\mathbf{P}}_{[0,\tau]})] + \mathbb{E}[\text{TV}^2(\bar{\mathbf{P}}_{[0,\tau]}, \mathbf{P}_{[0,\tau]}^*)] \\ &\lesssim \mathbb{E}[\text{KL}(\tilde{\mathbf{P}}_{[0,\tau]} \parallel \bar{\mathbf{P}}_{[0,\tau]})] + \mathbb{E}[\text{KL}(\mathbf{P}_{[0,\tau]}^* \parallel \bar{\mathbf{P}}_{[0,\tau]})], \end{aligned}$$

We analyse the two terms separately, each term involving proof techniques developed by Stromme (2023b). We summarize the results in the following propositions, which yield the proof of Theorem 4.1.

Proposition 4.2. *Assume the conditions of Theorem 4.1, then for any $\tau \in [0, 1)$*

$$\mathbb{E}[\text{KL}(\tilde{\mathbf{P}}_{[0,\tau]} \parallel \bar{\mathbf{P}}_{[0,\tau]})] \leq \frac{1}{\varepsilon} \mathbb{E}[\text{OT}_\varepsilon(\mu, \nu_n) - \text{OT}_\varepsilon(\mu, \nu)] \leq \varepsilon^{-(k/2+1)} n^{-1/2}.$$

Proposition 4.3. *Assume the conditions of Theorem 4.1, then*

$$\mathbb{E}[\text{KL}(\mathbf{P}_{[0,\tau]}^* \parallel \bar{\mathbf{P}}_{[0,\tau]})] \leq \frac{R^2 \varepsilon^{-k}}{n} (1 - \tau)^{-k-2}.$$

4.2 Completing the results

We now incorporate the discretization error. Letting $\hat{\mathbf{P}}$ denote the path measure induced by the dynamics of (22), we use the triangle inequality to introduce the path measure $\tilde{\mathbf{P}}$:

$$\mathbb{E}[\text{TV}^2(\hat{\mathbf{P}}_{[0,\tau]}, \mathbf{P}_{[0,\tau]}^*)] \lesssim \mathbb{E}[\text{TV}^2(\hat{\mathbf{P}}_{[0,\tau]}, \tilde{\mathbf{P}}_{[0,\tau]})] + \mathbb{E}[\text{TV}^2(\tilde{\mathbf{P}}_{[0,\tau]}, \mathbf{P}_{[0,\tau]}^*)].$$

The second term is precisely the statistical error, controlled by Theorem 4.1. For the first term, we employ a now-standard discretization argument (see e.g., [Chen et al. \(2022\)](#)) which bounds the total variation error as a function of the step-size parameter η and the Lipschitz constant of the empirical drift, which can be easily bounded in our setting.

Proposition 4.4. *Suppose $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$. Denoting L_τ for the Lipschitz constant of $\hat{b}_\tau$ (recall Equation (21)) for $t \in [0, 1)$ and η the step-size of the SDE discretization, it holds that*

$$\mathbb{E}[\text{TV}^2(\hat{\mathbf{P}}_{[0,\tau]}, \tilde{\mathbf{P}}_{[0,\tau]})] \lesssim (\varepsilon + 1)L_\tau^2 d\eta.$$

In particular, if $\text{supp}(\nu) \subseteq B(0; R)$, then

$$\mathbb{E}[\text{TV}^2(\hat{\mathbf{P}}_{[0,\tau]}, \tilde{\mathbf{P}}_{[0,\tau]})] \lesssim (\varepsilon + 1)(1 - \tau)^{-2} d\eta (1 \vee R^4 (1 - \tau)^{-2} \varepsilon^{-2}).$$

We now aggregate the statistical and approximation error into one final result.

Theorem 4.5. *Suppose $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$ with $\text{supp}(\nu) \subseteq B(0, R) \subseteq \mathcal{M}$, where $\mathcal{M}$ is a k -dimensional submanifold of $\mathbb{R}^d$. Given n i.i.d. samples from ν , the one-sample Sinkhorn bridge $\hat{\mathbf{P}}$ estimates the Schrödinger bridge $\mathbf{P}^*$ with the following error*

$$\begin{aligned} \mathbb{E}[\text{TV}^2(\hat{\mathbf{P}}_{[0,\tau]}, \mathbf{P}_{[0,\tau]}^*)] &\lesssim \left(\frac{\varepsilon^{-k/2-1}}{\sqrt{n}} + \frac{R^2 \varepsilon^{-k}}{(1 - \tau)^{k+2} n} \right) \\ &\quad + (\varepsilon + 1)(1 - \tau)^{-2} d\eta (1 \vee R^4 (1 - \tau)^{-2} \varepsilon^{-2}). \end{aligned}$$

Assuming $R \geq 1$ and $\varepsilon = 1$, the Schrödinger bridge can be estimated in total variation distance to accuracy ϵ_{TV} with n samples and N Euler–Maruyama steps, where

$$n \asymp \frac{R^2}{(1 - \tau)^{k+2} \epsilon_{\text{TV}}^2} \vee \epsilon_{\text{TV}}^{-4}, \quad N \lesssim \frac{dR^4}{(1 - \tau)^4 \epsilon_{\text{TV}}^2}.$$

Note that our error rates improve as $\varepsilon \rightarrow \infty$; since this is also the regime in which Sinkhorn’s algorithm terminates rapidly, it is natural to suppose that ε should be large in practice. This is misleading, however: as ε grows, the Schrödinger bridge becomes less and less informative,⁸ and the marginal $\mathbf{p}_\tau^*$ only resembles ν when τ becomes very close to 1. We elaborate on the use of the SB for sampling in the following section.

⁸In other words, the transport path is more and more volatile.

4.3 Application: Sampling with the Föllmer bridge

Theorem 4.5 does not immediately imply guarantees for sampling from the target distribution ν . Obtaining such guarantees requires arguing that simulating the Sinkhorn bridge on a suitable interval $[0, \tau]$ for τ close to 1 yields samples close to the true density (without completely collapsing onto the training data). We provide such a guarantee in this section, for the special case of the Föllmer bridge. We adopt this setting only for concreteness; similar arguments apply more broadly.

The Föllmer bridge is a special case of the Schrödinger bridge due to Hans Föllmer (Föllmer, 1985). In this setting, $\mu = \delta_a$ for any $a \in \mathbb{R}^d$, and our estimator takes a particularly simple form:

$$\hat{b}_t^F(z) = (1-t)^{-1} \left(-z + \frac{\sum_{j=1}^n Y_j \exp\left(\left(\frac{1}{2}\|Y_j\|^2 - \frac{1}{2(1-t)}\|z - Y_j\|^2\right)/\varepsilon\right)}{\sum_{j=1}^n \exp\left(\left(\frac{1}{2}\|Y_j\|^2 - \frac{1}{2(1-t)}\|z - Y_j\|^2\right)/\varepsilon\right)} \right), \quad (27)$$

Note that in this special case, calculating the drift does not require the use of Sinkhorn’s algorithm, and the drift, in fact, corresponds to the score of a kernel density estimator applied to ν_n . We provide a calculation of these facts in Appendix B.3 for completeness.

We then have the following guarantee.

Corollary 4.6. *Consider the assumptions of Theorem 4.5, further suppose that $\mu = \delta_0$ and $\varepsilon = 1$ and that the second moment of ν is bounded by d . Suppose we use n samples from ν to estimate the Föllmer drift, and simulate the resulting SDE using N Euler–Maruyama iterations until time $\tau = 1 - \epsilon_{W_2}^2/d$, with*

$$n \asymp \frac{R^2 d^{k+2}}{\epsilon_{W_2}^{2k+4} \epsilon_{TV}^2} \vee \epsilon_{TV}^{-4} \quad N \lesssim \frac{R^4 d^5}{\epsilon_{W_2}^8 \epsilon_{TV}^2}.$$

Then the density given by the Sinkhorn bridge at time τ iterations will be ϵ_{TV} -close in total variation to a measure which is ϵ_{W_2} -close to ν in the 2-Wasserstein distance.

Note that the choice $\varepsilon = 1$ was merely out of convenience. If instead the practitioner was willing to pay the computational price of solving Sinkhorn’s algorithm for small ε and large n , then the number of requisite iterations N would decrease. Finally, notice that the number of samples scales exponentially in the intrinsic dimension $k \ll d$ instead of the ambient dimension d . This is, of course, unavoidable, but improves upon recent work that uses kernel density estimators to prove a similar result for denoising diffusion probabilistic models (Wibisono et al., 2024).

Remark 4.7. Recently, Huang (2024) also proposed (27) to estimate the Föllmer drift. They provide no statistical estimation guarantees of the drift, nor any sampling guarantees; their contributions are largely empirical, demonstrating that the proposed estimator is tractable for high-dimensional tasks. The work of Huang et al. (2021) also proposes an estimator for the Föllmer bridge based on having partial access to the log-density ratio of the target distribution (without the normalizing constant).

Algorithm 1 Sinkhorn bridges

Input: Data $\{X_i\}_{i=1}^m \sim \mu$, $\{Y_j\}_{j=1}^n \sim \nu$, parameters $\varepsilon > 0$, $\tau \in (0, 1)$, and $N \geq 1$
Compute: Sinkhorn potentials $(\hat{f}, \hat{g}) \in \mathbb{R}^m \times \mathbb{R}^n$ $\triangleright$ Using [POT](#) or [OTT](#)
Initialize: $x^{(0)} = x \sim \mu$, $k = 0$, stepsize $\eta = \tau/N$
while $k \leq N - 1$ **do**
 $x^{(k+1)} = x^{(k)} + \eta \hat{b}_{k\eta}(x^{(k)}) + \sqrt{\eta\varepsilon}\xi$ $\triangleright \xi \sim \mathcal{N}(0, I)$
 $k \leftarrow k + 1$
end while
Return: $x^{(N)}$

5 Numerical performance

Our approach is summarized in Algorithm 1, and open-source code for replicating our experiments is available at <https://github.com/APooladian/SinkhornBridge>.⁹

For a fixed regularization parameter $\varepsilon > 0$, the runtime of computing $(\hat{f}, \hat{g})$ on the basis of samples has complexity $\mathcal{O}(mn/(\varepsilon\delta_{\text{tol}}))$, where δ_{tol} is a required tolerance parameter that measures how closely the the marginal constraints are satisfied ([Cuturi, 2013](#); [Peyré and Cuturi, 2019](#); [Altschuler et al., 2022](#)). Once these are computed, the evaluation of $\hat{b}_{k\eta}$ is $\mathcal{O}(n)$, with the remaining runtime being the number of iteration steps, denoted by N . In all our experiments, we take $m = n$, thus the total runtime complexity of the algorithm is a fixed cost of $\mathcal{O}(n^2/(\varepsilon\delta_{\text{tol}}))$, followed by $\mathcal{O}(nN)$ for each new sample to be generated (which can be parallelized).

5.1 Qualitative illustration

As a first illustration, we consider standard two-dimensional datasets from the machine learning literature. For all examples, we use $n = 2000$ training points from both the source and target measure, and run Sinkhorn’s algorithm with $\varepsilon = 0.1$. For generation, we set $\tau = 0.9$, and consider $N = 50$ Euler–Maruyama steps. Figure 1 contains the resulting simulations, starting from out-of-sample points. We see reasonable performance in each case.

5.2 Quantitative illustrations

We quantitatively assess the performance of our estimator using synthetic examples from the deep learning literature ([Bunne et al., 2023a](#); [Gushchin et al., 2023](#)).

5.2.1 The Gaussian case

We first demonstrate that we are indeed learning the drift and that the claimed rates are empirically justified. As a first step, we consider the simple case where $\mu = \mathcal{N}(a, A)$

⁹Our estimator is implemented in both the [POT](#) and [OTT-JAX](#) frameworks.

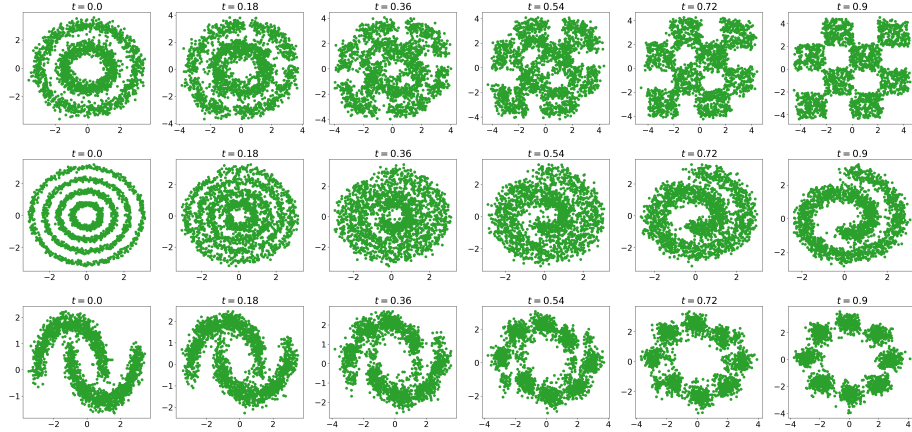


Figure 1: Schrödinger bridges on the basis of samples from toy datasets.

and $\nu = \mathcal{N}(b, B)$ for two positive-definite $d \times d$ matrices A and B and arbitrary vectors $a, b \in \mathbb{R}^d$. In this regime, the optimal drift b_τ^* and $\mathbf{p}_\tau^*$ has been computed in closed-form by Bunne et al. (2023a); see equations (25)-(29) in their work.

To verify that we are indeed learning the drift, we first draw n samples from μ and ν , and compute our estimator, $\hat{b}_\tau$ for any $\tau \in [0, 1)$. We then evaluate the mean-squared error

$$\text{MSE}(n, \tau) = \|\hat{b}_\tau - b_\tau^*\|_{L^2(\mathbf{p}_\tau^*)}^2,$$

by a Monte Carlo approximation, with $n_{\text{MC}} = 10000$. For simplicity, with $d = 3$, we choose $A = I$ and randomly generate a positive-definite matrix B , and center the Gaussians. We fix $\varepsilon = 1$ and vary n used to define our estimator, and perform the simulation ten times to generate error bars across various choices of $\tau \in [0, 1)$; see Figure 2.

It is clear from the plot that the *constant* associated to the rate of estimation gets worse as $\tau \rightarrow 1$, but the overall rate of convergence appears unchanged, which hovers around n^{-1} for all choices of τ shown in the plot, as expected from e.g., Proposition 4.2.

5.2.2 Multimodal measures with closed-form drift

The next setting is due to Gushchin et al. (2023); they devised a drift that defines the Schrödinger bridge between a Gaussian and a more complicated measure with multiple modes. This explicit drift allowed them to benchmark multiple neural network based methods for estimating the Schrödinger bridge for non-trivial couplings (e.g., beyond the Gaussian to Gaussian setting). We briefly remark that the approaches discussed in their work fall under the “continuous estimation” paradigm, where researchers assume they can endlessly sample from the distributions when training (using new samples per training iteration).

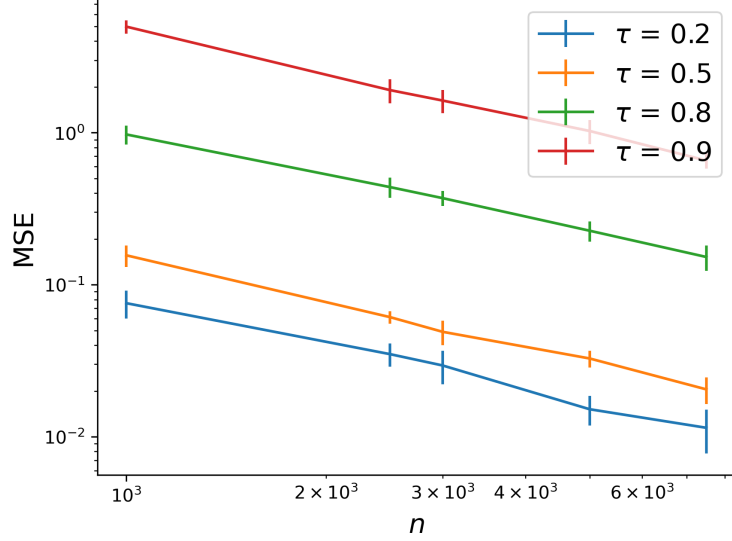


Figure 2: MSE for estimating the Gaussian drift as (n, τ) vary, averaged over 10 trials.

We consider the same pre-fixed drift as found in their publicly available code, which transports the standard Gaussian to a distribution with four modes. We consider the case $d = 64$ and $\varepsilon = 1$, as these hyperparameters are most extensively studied in their work, where they provide the most details on the other models. We use $n = 4096$ training samples from the source and target data they construct (which is significantly less than the total number of samples required for any of the neural network based models) and perform our estimation procedure, and we take $N = 100$ discretization steps (which is half as many as most of the works they consider) to simulate to time $\tau = 0.99$. To best illustrate the four mixture components, Figure 3 contains a scatter plot of the first and fifteenth dimension, containing fresh target samples and our generated samples.

We compare to the ground-truth samples using the unexplained variance percentage (UVP) based on the Bures–Wasserstein distance (Bures, 1969):

$$\mu \mapsto \text{BW-UVP}_\nu(\mu) := 100 \frac{\text{BW}^2(\mathcal{N}_\mu, \mathcal{N}_\nu)}{0.5 \cdot \text{Var}(\nu)},$$

where $\mathcal{N}_\mu = \mathcal{N}(\mathbb{E}_\mu[X], \text{Cov}_\mu(X))$, and same for $\mathcal{N}_\nu$.¹⁰ While seemingly ad hoc, the BW-UVP is widely used in the machine learning literature as a means of quantifying the quality of the generated samples (see e.g., Daniels et al. (2021)). We compute the BW-UVP with 10^4 generated samples from the target and our approach, averaged over 5 trials, and used the results of Gushchin et al. (2023) for the remaining methods (MLE-SB is by Vargas et al. (2021), EgNOT is by Mokrov et al. (2023), and FB-SDE-A is by Chen et al. (2021a)). We see that the Sinkhorn bridge has significantly lower

¹⁰For us, these quantities are computed on the basis of samples.

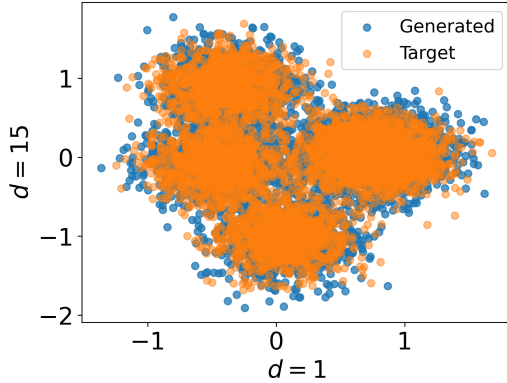


Figure 3: Plotting generated and re-sampled target data in $d = 64$.

Method	BW-UVF
Ours	0.41 ± 0.03
MLE-SB	0.56
EgNOT	0.85
FB-SDE-A	0.65

Table 1: Comparison to neural network approaches in BW-UVF for $d = 64$.

BW-UVF compared to the other approaches while requiring less compute resources and training data.

6 Conclusion

This work makes a connection between the static entropic optimal transport problem, the Schrödinger bridge problem, and Sinkhorn’s algorithm, which appeared to be lacking in the literature. We proposed and analyzed a plug-in estimator of the Schrödinger bridge, which we call the Sinkhorn bridge. Due to a Markov property enjoyed by entropic optimal couplings, our estimator relates Sinkhorn’s matrix-scaling algorithm to the optimal drift that arises in the Schrödinger bridge problem, and existing theory in the statistical optimal transport literature provide us with statistical guarantees. A novelty of our approach is the reduction of a “dynamic” estimation problem to a “static” one, where the latter is easy to analyze.

Several questions arise from our work, we highlight some here:

Further connections to other processes: Our arguments for the Schrödinger bridge used the particular form of the reversible Brownian motion. It would be interesting to develop this approach for other types of reference processes for the purposes of developing statistical guarantees. The Sinkhorn bridge estimator can also be implemented through an ordinary differential equation (ODE) and not necessarily through an SDE. This gives rise to the *probability flow* ODE in the generative modeling literature (Song et al., 2020). Chen et al. (2024a) showed that this approach can achieve results comparable to those obtained by diffusion models (Chen et al., 2022; Lee et al., 2023). We anticipate analogous results would hold in our setting.

Lower bounds: Entropic optimal transport suffers from a dearth of lower bounds

in the literature. It is unclear whether our approach is optimal in terms of its dependence on ε and τ . Developing estimators with better performance or nontrivial lower bounds would help establish how far our estimators are from optimality.

Computation in practice: On the computational side, one can ask if there are better estimators of the drift b_t^* than the plug-in estimator we outlined (possibly amenable to statistical analysis), and to consider using our estimator on non-synthetic problems. For example, it seems advisable to compute the Sinkhorn bridge in a latent space, and reverting the latent transformation later (Rombach et al., 2022).

Acknowledgements

AAP thanks NSF grant DMS-1922658 and Meta AI Research for financial support. JNW is supported by the Sloan Research Fellowship and NSF grant DMS-2339829.

References

- Albergo, M. S. and Vanden-Eijnden, E. (2022). Building normalizing flows with stochastic interpolants. *arXiv preprint arXiv:2209.15571*.
- Altschuler, J., Weed, J., and Rigollet, P. (2017). Near-linear time approximation algorithms for optimal transport via Sinkhorn iteration. In *Advances in Neural Information Processing Systems 30*.
- Altschuler, J. M., Niles-Weed, J., and Stromme, A. J. (2022). Asymptotics for semidiscrete entropic optimal transport. *SIAM Journal on Mathematical Analysis*, 54(2):1718–1741.
- Benamou, J.-D. and Brenier, Y. (2000). A computational fluid mechanics solution to the Monge–Kantorovich mass transfer problem. *Numerische Mathematik*, 84(3):375–393.
- Bernton, E., Heng, J., Doucet, A., and Jacob, P. E. (2019). Schrödinger bridge samplers. *arXiv preprint arXiv:1912.13170*.
- Brenier, Y. (1991). Polar factorization and monotone rearrangement of vector-valued functions. *Comm. Pure Appl. Math.*, 44(4):375–417.
- Bunne, C., Hsieh, Y.-P., Cuturi, M., and Krause, A. (2023a). The Schrödinger bridge between Gaussian measures has a closed form. In *International Conference on Artificial Intelligence and Statistics*, pages 5802–5833. PMLR.

- Bunne, C., Stark, S. G., Gut, G., Del Castillo, J. S., Levesque, M., Lehmann, K.-V., Pelkmans, L., Krause, A., and Rätsch, G. (2023b). Learning single-cell perturbation responses using neural optimal transport. *Nature methods*, 20(11):1759–1768.
- Bures, D. (1969). An extension of Kakutani’s theorem on infinite product measures to the tensor product of semifinite w^* -algebras. *Transactions of the American Mathematical Society*, 135:199–212.
- Carlier, G., Chernozhukov, V., and Galichon, A. (2016). Vector quantile regression: An optimal transport approach. *The Annals of Statistics*, 44(3):1165–1192.
- Chen, R. T., Rubanova, Y., Bettencourt, J., and Duvenaud, D. K. (2018). Neural ordinary differential equations. *Advances in neural information processing systems*, 31.
- Chen, S., Chewi, S., Lee, H., Li, Y., Lu, J., and Salim, A. (2024a). The probability flow ODE is provably fast. *Advances in Neural Information Processing Systems*, 36.
- Chen, S., Chewi, S., Li, J., Li, Y., Salim, A., and Zhang, A. R. (2022). Sampling is as easy as learning the score: Theory for diffusion models with minimal data assumptions. *arXiv preprint arXiv:2209.11215*.
- Chen, T., Liu, G.-H., and Theodorou, E. A. (2021a). Likelihood training of Schrödinger bridge using forward-backward SDEs theory. *arXiv preprint arXiv:2110.11291*.
- Chen, Y., Georgiou, T. T., and Pavon, M. (2016). On the relation between optimal transport and Schrödinger bridges: A stochastic control viewpoint. *Journal of Optimization Theory and Applications*, 169:671–691.
- Chen, Y., Georgiou, T. T., and Pavon, M. (2021b). Stochastic control liaisons: Richard Sinkhorn meets Gaspard Monge on a Schrödinger bridge. *Siam Review*, 63(2):249–313.
- Chen, Y., Goldstein, M., Hua, M., Albergo, M. S., Boffi, N. M., and Vanden-Eijnden, E. (2024b). Probabilistic forecasting with stochastic interpolants and Föllmer processes. *arXiv preprint arXiv:2403.13724*.
- Chernozhukov, V., Galichon, A., Hallin, M., and Henry, M. (2017). Monge–Kantorovich depth, quantiles, ranks and signs. *The Annals of Statistics*, 45(1):223–256.
- Chewi, S., Niles-Weed, J., and Rigollet, P. (2024). Statistical optimal transport.
- Chewi, S. and Pooladian, A.-A. (2023). An entropic generalization of Caffarelli’s contraction theorem via covariance inequalities. *Comptes Rendus. Mathématique*, 361(G9):1471–1482.

- Chiarini, A., Conforti, G., Greco, G., and Tamanini, L. (2022). Gradient estimates for the Schrödinger potentials: Convergence to the Brenier map and quantitative stability. *arXiv preprint arXiv:2207.14262*.
- Chizat, L., Roussillon, P., Léger, F., Vialard, F.-X., and Peyré, G. (2020). Faster Wasserstein distance estimation with the Sinkhorn divergence. *Advances in Neural Information Processing Systems*, 33:2257–2269.
- Chizat, L., Zhang, S., Heitz, M., and Schiebinger, G. (2022). Trajectory inference via mean-field Langevin in path space. *Advances in Neural Information Processing Systems*, 35:16731–16742.
- Conforti, G. (2022). Weak semiconvexity estimates for Schrödinger potentials and logarithmic Sobolev inequality for Schrödinger bridges. *arXiv preprint arXiv:2301.00083*.
- Conforti, G., Durmus, A., and Greco, G. (2023). Quantitative contraction rates for Sinkhorn algorithm: Beyond bounded costs and compact marginals. *arXiv preprint arXiv:2304.04451*.
- Conforti, G. and Tamanini, L. (2021). A formula for the time derivative of the entropic cost and applications. *Journal of Functional Analysis*, 280(11):108964.
- Csiszár, I. (1975). I -divergence geometry of probability distributions and minimization problems. *Ann. Probability*, 3:146–158.
- Cuturi, M. (2013). Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in Neural Information Processing Systems*, 26.
- Daniels, M., Maunu, T., and Hand, P. (2021). Score-based generative neural networks for large-scale optimal transport. *Advances in neural information processing systems*, 34:12955–12965.
- De Bortoli, V., Thornton, J., Heng, J., and Doucet, A. (2021). Diffusion Schrödinger bridge with applications to score-based generative modeling. *Advances in Neural Information Processing Systems*, 34:17695–17709.
- del Barrio, E., Gonzalez-Sanz, A., Loubes, J.-M., and Niles-Weed, J. (2022). An improved central limit theorem and fast convergence rates for entropic transportation costs. *arXiv preprint arXiv:2204.09105*.
- Divol, V., Niles-Weed, J., and Pooladian, A.-A. (2022). Optimal transport map estimation in general function spaces. *arXiv preprint arXiv:2212.03722*.
- Divol, V., Niles-Weed, J., and Pooladian, A.-A. (2024). Tight stability bounds for entropic Brenier maps. *arXiv preprint arXiv:2404.02855*.

- Eldan, R., Lehec, J., and Shenfeld, Y. (2020). Stability of the logarithmic Sobolev inequality via the Föllmer process.
- Finlay, C., Gerolin, A., Oberman, A. M., and Pooladian, A.-A. (2020). Learning normalizing flows from Entropy-Kantorovich potentials. *arXiv preprint arXiv:2006.06033*.
- Föllmer, H. (1985). An entropy approach to the time reversal of diffusion processes. In *Stochastic differential systems (Marseille-Luminy, 1984)*, volume 69 of *Lect. Notes Control Inf. Sci.*, pages 156–163. Springer, Berlin.
- Fortet, R. (1940). Résolution d’un système d’équations de m. Schrödinger. *Journal de Mathématiques Pures et Appliquées*, 19(1-4):83–105.
- Genevay, A. (2019). *Entropy-regularized optimal transport for machine learning*. PhD thesis, Paris Sciences et Lettres (ComUE).
- Gentil, I., Léonard, C., Ripani, L., and Tamanini, L. (2020). An entropic interpolation proof of the HWI inequality. *Stochastic Processes and their Applications*, 130(2):907–923.
- Ghosal, P., Nutz, M., and Bernton, E. (2022). Stability of entropic optimal transport and Schrödinger bridges. *Journal of Functional Analysis*, 283(9):109622.
- Ghosal, P. and Sen, B. (2022). Multivariate ranks and quantiles using optimal transport: consistency, rates and nonparametric testing. *Ann. Statist.*, 50(2):1012–1037.
- Goldfeld, Z., Kato, K., Rioux, G., and Sadhu, R. (2022a). Limit theorems for entropic optimal transport maps and the Sinkhorn divergence. *arXiv preprint arXiv:2207.08683*.
- Goldfeld, Z., Kato, K., Rioux, G., and Sadhu, R. (2022b). Statistical inference with regularized optimal transport. *arXiv preprint arXiv:2205.04283*.
- Gonzalez-Sanz, A., Loubes, J.-M., and Niles-Weed, J. (2022). Weak limits of entropy regularized optimal transport; potentials, plans and divergences. *arXiv preprint arXiv:2207.07427*.
- Groppe, M. and Hundrieser, S. (2023). Lower complexity adaptation for empirical entropic optimal transport. *arXiv preprint arXiv:2306.13580*.
- Gushchin, N., Kolesov, A., Mokrov, P., Karpikova, P., Spiridonov, A., Burnaev, E., and Korotin, A. (2023). Building the bridge of Schrödinger: A continuous entropic optimal transport benchmark. *Advances in Neural Information Processing Systems*, 36:18932–18963.
- Huang, H. (2024). One-step data-driven generative model via Schrödinger bridge. *arXiv preprint arXiv:2405.12453*.

- Huang, J., Jiao, Y., Kang, L., Liao, X., Liu, J., and Liu, Y. (2021). Schrödinger–Föllmer sampler: Sampling without ergodicity. *arXiv preprint arXiv:2106.10880*.
- Hütter, J.-C. and Rigollet, P. (2021). Minimax estimation of smooth optimal transport maps. *The Annals of Statistics*, 49(2):1166–1194.
- Kassraie, P., Pooladian, A.-A., Klein, M., Thornton, J., Niles-Weed, J., and Cuturi, M. (2024). Progressive entropic optimal transport solvers. *arXiv preprint arXiv:2406.05061*.
- Kato, K. (2024). Large deviations for dynamical Schrödinger problems. *arXiv preprint arXiv:2402.05100*.
- Kawakita, G., Kamiya, S., Sasai, S., Kitazono, J., and Oizumi, M. (2022). Quantifying brain state transition cost via Schrödinger bridge. *Network Neuroscience*, 6(1):118–134.
- Lavenant, H., Zhang, S., Kim, Y.-H., and Schiebinger, G. (2021). Towards a mathematical theory of trajectory inference. *arXiv preprint arXiv:2102.09204*.
- Lee, D., Lee, D., Bang, D., and Kim, S. (2024). Disco: Diffusion Schrödinger bridge for molecular conformer optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 13365–13373.
- Lee, H., Lu, J., and Tan, Y. (2023). Convergence of score-based generative modeling for general data distributions. In *International Conference on Algorithmic Learning Theory*, pages 946–985. PMLR.
- Léonard, C. (2012). From the Schrödinger problem to the Monge–Kantorovich problem. *Journal of Functional Analysis*, 262(4):1879–1920.
- Léonard, C. (2013). A survey of the Schrödinger problem and some of its connections with optimal transport. *arXiv preprint arXiv:1308.0215*.
- Lipman, Y., Chen, R. T., Ben-Hamu, H., Nickel, M., and Le, M. (2022). Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*.
- Liu, G.-H., Chen, T., So, O., and Theodorou, E. (2022a). Deep generalized Schrödinger bridge. *Advances in Neural Information Processing Systems*, 35:9374–9388.
- Liu, X., Gong, C., and Liu, Q. (2022b). Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*.
- Manole, T., Balakrishnan, S., Niles-Weed, J., and Wasserman, L. (2021). Plugin estimation of smooth optimal transport maps. *arXiv preprint arXiv:2107.12364*.

- Manole, T., Bryant, P., Alison, J., Kuusela, M., and Wasserman, L. (2022). Background modeling for double Higgs boson production: Density ratios and optimal transport. *arXiv preprint arXiv:2208.02807*.
- McCann, R. J. (1997). A convexity principle for interacting gases. *Advances in mathematics*, 128(1):153–179.
- Mena, G. and Niles-Weed, J. (2019). Statistical bounds for entropic optimal transport: sample complexity and the central limit theorem. *Advances in Neural Information Processing Systems*, 32.
- Mikulincer, D. and Shenfeld, Y. (2024). The Brownian transport map. *Probability Theory and Related Fields*, pages 1–66.
- Mokrov, P., Korotin, A., Kolesov, A., Gushchin, N., and Burnaev, E. (2023). Energy-guided entropic neural optimal transport. *arXiv preprint arXiv:2304.06094*.
- Nusken, N., Vargas, F., Ovsianas, A., Fernandes, D., Girolami, M., and Lawrence, N. (2022). Bayesian learning via neural Schrödinger–Föllmer flows. *STATISTICS AND COMPUTING*, 33.
- Nutz, M. and Wiesel, J. (2021). Entropic optimal transport: Convergence of potentials. *Probability Theory and Related Fields*, pages 1–24.
- Pavon, M., Trigila, G., and Tabak, E. G. (2021). The data-driven Schrödinger bridge. *Communications on Pure and Applied Mathematics*, 74(7):1545–1573.
- Peyré, G. and Cuturi, M. (2019). Computational optimal transport. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607.
- Pooladian, A.-A., Divol, V., and Niles-Weed, J. (2023). Minimax estimation of discontinuous optimal transport maps: The semi-discrete case. *arXiv preprint arXiv:2301.11302*.
- Pooladian, A.-A. and Niles-Weed, J. (2021). Entropic estimation of optimal transport maps. *arXiv preprint arXiv:2109.12004*.
- Rigollet, P. and Stromme, A. J. (2022). On the sample complexity of entropic optimal transport. *arXiv preprint arXiv:2206.13472*.
- Ripani, L. (2019). Convexity and regularity properties for entropic interpolations. *Journal of Functional Analysis*, 277(2):368–391.
- Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695.

- Salimans, T., Zhang, H., Radford, A., and Metaxas, D. (2018). Improving GANs using optimal transport. In *International Conference on Learning Representations*.
- Santambrogio, F. (2015). Optimal transport for applied mathematicians. *Birkäuser, NY*, 55(58-63):94.
- Schrödinger, E. (1932). Sur la théorie relativiste de l’électron et l’interprétation de la mécanique quantique. In *Annales de l’institut Henri Poincaré*, volume 2, pages 269–310.
- Shi, Y., De Bortoli, V., Campbell, A., and Doucet, A. (2024). Diffusion Schrödinger bridge matching. *Advances in Neural Information Processing Systems*, 36.
- Shi, Y., De Bortoli, V., Deligiannidis, G., and Doucet, A. (2022). Conditional simulation using diffusion Schrödinger bridges. In *Uncertainty in Artificial Intelligence*, pages 1792–1802. PMLR.
- Sinkhorn, R. (1964). A relationship between arbitrary positive matrices and doubly stochastic matrices. *The annals of mathematical statistics*, 35(2):876–879.
- Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. (2020). Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*.
- Stromme, A. (2023a). Sampling from a Schrödinger bridge. In *International Conference on Artificial Intelligence and Statistics*, pages 4058–4067. PMLR.
- Stromme, A. J. (2023b). Minimum intrinsic dimension scaling for entropic optimal transport. *arXiv preprint arXiv:2306.03398*.
- Thornton, J., Hutchinson, M., Mathieu, E., De Bortoli, V., Teh, Y. W., and Doucet, A. (2022). Riemannian diffusion Schrödinger bridge. *arXiv preprint arXiv:2207.03024*.
- Tong, A., Malkin, N., Fatras, K., Atanackovic, L., Zhang, Y., Huguet, G., Wolf, G., and Bengio, Y. (2023). Simulation-free Schrödinger bridges via score and flow matching. *arXiv preprint arXiv:2307.03672*.
- Vargas, F., Ovsianas, A., Fernandes, D., Girolami, M., Lawrence, N. D., and Nüsken, N. (2023). Bayesian learning via neural Schrödinger–Föllmer flows. *Statistics and Computing*, 33(1):3.
- Vargas, F., Thodoroff, P., Lamacraft, A., and Lawrence, N. (2021). Solving Schrödinger bridges via maximum likelihood. *Entropy*, 23(9):1134.
- Vempala, S. and Wibisono, A. (2019). Rapid convergence of the unadjusted langevin algorithm: Isoperimetry suffices. *Advances in neural information processing systems*, 32.

Villani, C. (2009). *Optimal transport: old and new*, volume 338. Springer.

Werenski, M., Murphy, J. M., and Aeron, S. (2023). Estimation of entropy-regularized optimal transport maps between non-compactly supported measures. *arXiv preprint arXiv:2311.11934*.

Wibisono, A., Wu, Y., and Yang, K. Y. (2024). Optimal score estimation via empirical Bayes smoothing. *arXiv preprint arXiv:2402.07747*.

Yim, J., Trippe, B. L., De Bortoli, V., Mathieu, E., Doucet, A., Barzilay, R., and Jaakkola, T. (2023). Se (3) diffusion model with application to protein backbone generation. *arXiv preprint arXiv:2302.02277*.

A Dynamic entropic optimal transport

A.1 Connecting the two formulations

In this section, we reconcile (at a formal level) two versions of the dynamic formulation for entropic optimal transport. We will start with (11) and show that this is equivalent to (10) by a reparameterization.

We begin by recognizing that $\Delta \mathbf{p}_t = \nabla \cdot (\mathbf{p}_t \nabla \log \mathbf{p}_t)$, which allows us to write the Fokker–Planck equation as

$$\partial_t \mathbf{p}_t + \nabla \cdot \left((v_t - \frac{\varepsilon}{2} \nabla \log \mathbf{p}_t) \mathbf{p}_t \right) = 0, \quad (28)$$

Inserting $b_t := v_t - \frac{\varepsilon}{2} \nabla \log \mathbf{p}_t$ into (11), we expand the square and arrive at

$$\inf_{(\mathbf{p}_t, b_t)} \int_0^1 \int \left(\frac{1}{2} \|b_t(x)\|^2 + \frac{\varepsilon^2}{8} \|\nabla \log \mathbf{p}_t(x)\|^2 + \frac{\varepsilon}{2} b_t^\top \nabla \log \mathbf{p}_t \right) \mathbf{p}_t(x) \, dx \, dt.$$

Up to the cross-term, this aligns with (10); it remains to eliminate the cross term. Using integration-by-parts and (28), we obtain

$$\int_0^1 \int (b_t \mathbf{p}_t)^\top \nabla \log \mathbf{p}_t \, dx \, dt = - \int_0^1 \int \nabla \cdot (b_t \mathbf{p}_t) \log \mathbf{p}_t \, dx \, dt = \int_0^1 \int (\partial_t \mathbf{p}_t) \log \mathbf{p}_t \, dx \, dt.$$

Though, we have (by product rule) the equivalence

$$\partial_t (\mathbf{p}_t \log \mathbf{p}_t) - \partial_t \mathbf{p}_t = (\partial_t \mathbf{p}_t) \log \mathbf{p}_t.$$

Exchanging partial derivatives under the integral, this yields the following simplification

$$\begin{aligned}
\int_0^1 \int (\partial_t \mathbf{p}_t) \log \mathbf{p}_t \, dx \, dt &= \int_0^1 \int \partial_t (\mathbf{p}_t \log \mathbf{p}_t) \, dx \, dt - \int_0^1 \int \partial_t \mathbf{p}_t \, dx \, dt \\
&= \int_0^1 \partial_t \int \mathbf{p}_t \log \mathbf{p}_t \, dx \, dt - \int_0^1 \partial_t \int \mathbf{p}_t \, dx \, dt \\
&= \int_0^1 \partial_t \mathcal{H}(\mathbf{p}_t) \, dt + 0 \\
&= \mathcal{H}(\mathbf{p}_1) - \mathcal{H}(\mathbf{p}_0),
\end{aligned}$$

where $\mathbf{p}_1 = \nu$ and $\mathbf{p}_0 = \mu$. We see that (11) is equivalent to

$$\frac{\varepsilon}{2}(\mathcal{H}(\nu) - \mathcal{H}(\mu)) + \inf_{(\mathbf{p}_t, b_t)} \int_0^1 \int \left(\frac{1}{2} \|b_t(x)\|^2 + \frac{\varepsilon^2}{8} \|\nabla \log \mathbf{p}_t(x)\|^2 \right) \mathbf{p}_t(x) \, dx \, dt.$$

A.2 Connecting Markov processes and entropic Brenier maps

Here we prove Proposition 3.1. To continue, we require the following lemma.

Lemma A.1. *Fix any $t \in [0, 1]$. Under $\mathbf{M}$, the random variables X_0 and X_1 are conditionally independent given X_t .*

Proof. A calculation shows that the joint density of X_0 , X_1 , and X_t with respect to $\mu_0 \otimes \mu_1 \otimes \text{Lebesgue}$ equals

$$\begin{aligned}
\Lambda_\varepsilon \Lambda_{t(1-t)\varepsilon} e^{-\frac{1}{2\varepsilon t(1-t)} \|x_t - ((1-t)x_0 + tx_1)\|^2} e^{(f(x_0) + g(x_1) - \frac{1}{2} \|x_0 - x_1\|^2)/\varepsilon} \\
= F_t(x_t, x_0) G_t(x_t, x_1),
\end{aligned}$$

where

$$\begin{aligned}
F_t(x_t, x_0) &= \Lambda_{\varepsilon t} e^{f(x_0)/\varepsilon} e^{-\frac{1}{2\varepsilon t} \|x_t - x_0\|^2} \\
G_t(x_t, x_1) &= \Lambda_{(1-t)\varepsilon} e^{g(x_1)/\varepsilon} e^{-\frac{1}{2\varepsilon(1-t)} \|x_t - x_1\|^2}.
\end{aligned}$$

Since this density factors, the law of X_0 and X_1 given X_t is a product measure, proving the claim. $\square$

Proof of Proposition 3.1. First, we prove that $\mathbf{M}$ is Markov. Let $(X_t)_{t \in [0, 1]}$ be distributed according to $\mathbf{M}$. It suffices to show that for any integrable $a \in \sigma(X_{[0, t]})$, $b \in \sigma(X_{[t, 1]})$, we have the identity

$$\mathbb{E}[ab|X_t] = \mathbb{E}[a|X_t] \mathbb{E}[b|X_t] \quad \text{a.s.}$$

Using the tower property and the fact that, conditioned on X_0 and X_1 , the law of the path is a Brownian bridge between X_0 and X_1 , and hence is Markov, we have

$$\begin{aligned}\mathbb{E}_{\mathbf{M}}[ab|X_t] &= \mathbb{E}[\mathbb{E}[ab|X_0, X_t, X_1]|X_t] \\ &= \mathbb{E}[\mathbb{E}[a|X_0, X_t]\mathbb{E}[b|X_t, X_1]|X_t].\end{aligned}$$

By Lemma A.1, the sigma-algebras $\sigma(X_0, X_t)$ and $\sigma(X_t, X_1)$ are conditionally independent given X_t , hence

$$\mathbb{E}[\mathbb{E}[a|X_0, X_t]\mathbb{E}[b|X_t, X_1]|X_t] = \mathbb{E}[\mathbb{E}[a|X_0, X_t]|X_t]\mathbb{E}[\mathbb{E}[b|X_t, X_1]|X_t] = \mathbb{E}[a|X_t]\mathbb{E}[b|X_t],$$

as claimed.

The proof of the second statement follows directly from the computations presented below (16), which hold under no additional assumptions.

We now prove the third statement. Following the approach of Föllmer (1985), the representation of $\mathbf{M}$ as a mixture of Brownian bridges shows that the law of $X_{[0,t]}$ for any $t < 1$ has finite entropy with respect to the law of $X_0 + \sqrt{\varepsilon}B_t$, for $X \sim \mu_0$. Hence, to verify the representation in terms of the SDE, it suffices to compute the stochastic derivative:

$$\lim_{h \rightarrow 0} \frac{1}{h} \mathbb{E}[X_{t+h} - X_t | X_{[0,t]}],$$

where the limit is taken in L^2 . Using the fact that the process is Markov and, conditioned on X_0 and X_1 , the path is a Brownian bridge, we obtain

$$\begin{aligned}\lim_{h \rightarrow 0} \frac{1}{h} \mathbb{E}[X_{t+h} - X_t | X_{[0,t]}] &= \lim_{h \rightarrow 0} \frac{1}{h} \mathbb{E}[\mathbb{E}[X_{t+h} - X_t | X_0, X_t, X_1] | X_t] \\ &= \frac{1}{1-t} \mathbb{E}[X_1 - X_t | X_t].\end{aligned}$$

Recalling the computations in Lemma A.1, we observe that, conditioned on $X_t = x_t$, the variable X_1 has μ_1 density proportional to $\mathbf{G}_t(x_t, x_1)$. Since π is a probability measure, in particular we have that e^g lies in $L^1(\mu_1)$. We can therefore apply dominated convergence to obtain

$$\frac{1}{1-t} \mathbb{E}[X_1 - X_t | X_t = x_t] = \frac{\int \frac{x_1 - x_t}{1-t} \mathbf{G}_t(x_t, x_1) \mu_1(dx_1)}{\int \mathbf{G}_t(x_t, x_1) \mu_1(dx_1)} = \varepsilon \nabla \log \mathcal{H}_{(1-t)\varepsilon}[\exp(g/\varepsilon) \mu_1](x_t),$$

as desired.

For the fourth statement, we require the following claim.

Claim: The joint probability measure $\pi_t(z, x_1)$, defined as

$$\exp((-(1-t)f_{1-t}(z) + (1-t)g(x_1) - \frac{1}{2}\|z - x_1\|^2))/((1-t)\varepsilon)) \mathbf{m}_t(dz) \mu_1(dx_1),$$

is the optimal entropic coupling from $\mathbf{m}_t$ to ρ with regularization parameter $(1-t)\varepsilon$, where $f_{1-t}(z) := \varepsilon \log \mathcal{H}_{(1-t)\varepsilon}[e^{g/\varepsilon} \mu_1](z)$. Under this claim, it is easy to verify that

the definition of $\nabla\varphi_{1-t}$ is precisely this conditional expectation, which concludes the proof.

To prove the claim, we notice that π_t is already in the correct form of an optimal entropic coupling, and $\pi_t \in \Gamma(\mathbf{m}_t, ?)$ by construction. Thus, it suffices to only check the second marginal. By the second part, above, we have that

$$\mathbf{m}_t(z) = \mathcal{H}_{(1-t)\varepsilon}[\exp(g/\varepsilon)\mu_1](z)\mathcal{H}_{t\varepsilon}[\exp(f/\varepsilon)\mu_0](z).$$

Integrating, performing the appropriate cancellations, and applying the semigroup property, we have

$$\begin{aligned} \int \pi_t(z, dx_1) dz &= e^{g(x_1)/\varepsilon} \mu_1(dx_1) \mathcal{H}_{(1-t)\varepsilon}[\mathcal{H}_{t\varepsilon}[e^{f/\varepsilon}\mu_0]](x_1) \\ &= e^{g(x_1)/\varepsilon} \mu_1(dx_1) \mathcal{H}_\varepsilon[e^{f/\varepsilon}\mu_0](x_1), \end{aligned}$$

which proves the claim. $\square$

B Proofs for Section 4

B.1 One-sample analysis

Proof of Proposition 4.2. First, we recognize that a path with law $\tilde{\mathbf{P}}$ (resp. $\bar{\mathbf{P}}$) can be obtained by sampling a Brownian bridge between $(X_0, X_1) \sim \pi_n$ (resp. $\bar{\pi}_n$), by Proposition 3.1. Thus, by the data processing inequality,

$$\begin{aligned} \mathbb{E}[\text{KL}(\tilde{\mathbf{P}}_{[0,\tau]} \parallel \bar{\mathbf{P}}_{[0,\tau]})] &\leq \mathbb{E}[\text{KL}(\tilde{\mathbf{P}} \parallel \bar{\mathbf{P}})] \\ &\leq \mathbb{E}[\text{KL}(\pi_n \parallel \bar{\pi}_n)] \\ &= \mathbb{E}\left[\int \log(\pi_n/\bar{\pi}_n) d\pi_n\right], \end{aligned}$$

where the above manipulations are valid as both π_n and $\bar{\pi}_n$ have densities with respect to $\mu \otimes \nu_n$. Completing the expansion by explicitly writing out the densities, we obtain

$$\begin{aligned} \mathbb{E}[\text{KL}(\tilde{\mathbf{P}}_{[0,\tau]} \parallel \bar{\mathbf{P}}_{[0,\tau]})] &\leq \frac{1}{\varepsilon} \mathbb{E}\left[\int (\hat{f} + \hat{g} - \bar{f} - g^*) d\pi_n\right] \\ &= \frac{1}{\varepsilon} \mathbb{E}[\text{OT}_\varepsilon(\mu, \nu_n) - \int \bar{f} d\mu - \int g^* d\nu_n]. \end{aligned}$$

We now employ the rounding trick of [Stromme \(2023b\)](#): the rounded potential $\bar{f}$ satisfies

$$\bar{f} = \operatorname{argmax}_{f \in L^1(\mu)} \Phi^{\mu\nu_n}(f, g^*);$$

Therefore, in particular, $\Phi^{\mu\nu_n}(\bar{f}, g^*) \geq \Phi^{\mu\nu_n}(f^*, g^*)$. Continuing from above, we obtain

$$\begin{aligned} \mathbb{E}[\text{KL}(\tilde{\mathbf{P}}_{[0,\tau]} \parallel \bar{\mathbf{P}}_{[0,\tau]})] &\leq \frac{1}{\varepsilon} \mathbb{E}[\text{OT}_\varepsilon(\mu, \nu_n) - \int f^* d\mu - \int g^* d\nu_n] \\ &= \frac{1}{\varepsilon} \mathbb{E}[\text{OT}_\varepsilon(\mu, \nu_n) - \int f^* d\mu - \int g^* d\nu] \\ &= \frac{1}{\varepsilon} \mathbb{E}[\text{OT}_\varepsilon(\mu, \nu_n) - \text{OT}_\varepsilon(\mu, \nu)], \end{aligned}$$

where in the penultimate equality we observed that g is independent of the data $Y_1, \dots, Y_n$. Combined with Theorem 2.6 of [Groppe and Hundrieser \(2023\)](#), the proof is complete. $\square$

Proof of Proposition 4.3. We start by applying Girsanov's theorem to obtain a difference in the drifts, which can be re-written as differences in entropic Brenier maps:

$$\begin{aligned} \mathbb{E}[\text{KL}(\mathbf{P}_{[0,\tau]}^* \parallel \bar{\mathbf{P}}_{[0,\tau]})] &\leq \int_0^\tau \mathbb{E} \|\bar{b}_t - b_t^*\|_{L^2(\mathbf{p}_t)}^2 dt \\ &= \int_0^\tau (1-t)^{-2} \mathbb{E} \|\nabla \bar{\varphi}_{1-t} - \nabla \varphi_{1-t}^*\|_{L^2(\mathbf{p}_t)}^2 dt. \end{aligned} \tag{29}$$

The result then follows from Lemma B.1, where we lazily bound the resulting integral:

$$\begin{aligned} \mathbb{E}[\text{KL}(\mathbf{P}_{[0,\tau]}^* \parallel \bar{\mathbf{P}}_{[0,\tau]})] &\leq \frac{R^2 \varepsilon^{-k}}{n} \int_0^\tau (1-t)^{-k-2} dt \\ &\leq \frac{R^2 \varepsilon^{-k}}{n} (1-\tau)^{-k-2}. \end{aligned}$$

$\square$

Lemma B.1 (Point-wise drift bound). *Under the assumptions of Proposition 4.3, let $\bar{\varphi}_{1-t}$ be the entropic Brenier map between $\bar{\mathbf{p}}_t$ and $\bar{\nu}_n$ and φ_{1-t}^* be the entropic Brenier map between $\mathbf{p}_t^*$ and ν , both with regularization parameter $(1-t)\varepsilon$. Then*

$$\mathbb{E} \|\nabla \bar{\varphi}_{1-t} - \nabla \varphi_{1-t}^*\|_{L^2(\mathbf{p}_t)}^2 \lesssim \frac{R^2}{n} ((1-t)\varepsilon)^{-k}.$$

Proof. Setting some notation, we express $\nabla \varphi_{1-t}^*$ as the conditional expectation of the optimal entropic coupling π_t^* between $\mathbf{p}_t^*$ and ν (recall Proposition 3.1), where we write $\pi_t^*(z, y) = \gamma_t^*(z, y) \mathbf{p}_t^*(dz) \nu(dy)$.

The rest of our proof follows a technique due to [Stromme \(2023b\)](#): by triangle inequality, we can add and subtract the following term

$$\frac{1}{n} \sum_{j=1}^n Y_j \gamma_t^*(z, Y_j),$$

into the integrand in (29), resulting in

$$\begin{aligned} \mathbb{E} \|\nabla \bar{\varphi}_{1-t} - \nabla \varphi_{1-t}^*\|_{L^2(\mathbf{p}_t^*)}^2 &\lesssim \mathbb{E} \|\nabla \bar{\varphi}_{1-t} - n^{-1} \sum_{j=1}^n Y_j \gamma_t^*(\cdot, Y_j)\|_{L^2(\mathbf{p}_t^*)}^2 \\ &\quad + \mathbb{E} \|n^{-1} \sum_{j=1}^n Y_j \gamma_t^*(\cdot, Y_j) - \nabla \varphi_{1-t}^*\|_{L^2(\mathbf{p}_t^*)}^2. \end{aligned} \quad (30)$$

For the second term, with the same manipulations as [Stromme \(2023b, Lemma 20\)](#), we obtain a final bound of

$$\mathbb{E} \|n^{-1} \sum_{j=1}^n Y_j \gamma_t^*(\cdot, Y_j) - \nabla \varphi_{1-t}^*\|_{L^2(\mathbf{p}_t^*)}^2 = \frac{R^2}{n} \|\gamma_t^*\|_{L^2(\mathbf{p}_t^* \otimes \nu)}^2 \leq \frac{R^2}{n} ((1-t)\varepsilon)^{-k},$$

where the final inequality is also due to [Stromme \(2023b, Lemma 16\)](#). To control the first term in (30), we also appeal to his calculations of the same theorem: observing that, from (26)

$$\begin{aligned} \nabla \bar{\varphi}_{1-t}(z) &= \frac{1}{n} \sum_{j=1}^n Y_j \frac{\exp((g^*(Y_j) - \frac{1}{2(1-t)} \|z - Y_j\|^2)/\varepsilon)}{\frac{1}{n} \sum_{j=1}^n \exp((g^*(Y_j) - \frac{1}{2(1-t)} \|z - Y_j\|^2)/\varepsilon)} \\ &= \frac{1}{n} \sum_{j=1}^n Y_j \bar{\gamma}_t(z, Y_j). \end{aligned}$$

Since the following equality is true

$$\bar{\gamma}_t(z, Y_j) = \frac{\gamma_t^*(z, Y_j)}{\frac{1}{n} \sum_{k=1}^n \gamma_t^*(z, Y_k)},$$

we can verbatim apply the remaining arguments of [Stromme \(2023b, Lemma 20\)](#). Indeed, for fixed $x \in \mathbb{R}^d$, we have

$$\|n^{-1} \sum_{j=1}^n Y_j (\gamma_t^*(x, Y_j) - \bar{\gamma}_t(x, Y_j))\|^2 \leq R^2 \left| \sum_{j=1}^n \gamma_t^*(x, Y_j) - 1 \right|^2.$$

Taking the $L^2(\mathbf{p}_t^*)$ norm and the outer expectation, we see that the remaining term is nothing but the first component of the gradient of the dual entropic objective function (see [Lemma C.3](#)), which can be bounded via [Lemma C.4](#), resulting in the chain of inequalities

$$\mathbb{E} \|n^{-1} \sum_{j=1}^n Y_j (\gamma_t^*(\cdot, Y_j) - \bar{\gamma}_t(\cdot, Y_j))\|_{L^2(\mathbf{p}_t^*)}^2 \lesssim \frac{R^2}{n} \|\gamma_t^*\|_{L^2(\mathbf{p}_t^* \otimes \nu)}^2 \leq \frac{R^2}{n} ((1-t)\varepsilon)^{-k},$$

where the last inequality again holds via [Stromme \(2023b, Lemma 16\)](#). □

B.2 Completing the results

Proof of Proposition 4.4. This proof closely follows the ideas of [Chen et al. \(2022\)](#). Applying Girsanov's theorem, we obtain

$$\text{TV}^2(\hat{\mathbf{P}}_{[0,\tau]}, \tilde{\mathbf{P}}_{[0,\tau]}) \lesssim \text{KL}(\tilde{\mathbf{P}}_{[0,\tau]} \| \hat{\mathbf{P}}_{[0,\tau]}) = \sum_{k=0}^{N-1} \int_{k\eta}^{(k+1)\eta} \mathbb{E}_{\tilde{\mathbf{P}}_{[0,\tau]}} \|\hat{b}_{k\eta}(X_{k\eta}) - \hat{b}_t(X_t)\|^2 dt.$$

Recall that $\eta \in (0, 1)$ is a chosen step-size based on N , the number of steps to be taken. As in prior analyses, we hope to uniformly bound the integrand above for any $t \in [k\eta, (k+1)\eta]$. Adding and subtracting the appropriate terms, we have

$$\begin{aligned} \mathbb{E}_{\tilde{\mathbf{P}}_{[0,\tau]}} \|\hat{b}_{k\eta}(X_{k\eta}) - \hat{b}_t(X_t)\|^2 &\lesssim \mathbb{E}_{\tilde{\mathbf{P}}_{[0,\tau]}} \|\hat{b}_{k\eta}(X_{k\eta}) - \hat{b}_t(X_{k\eta})\|^2 \\ &\quad + \mathbb{E}_{\tilde{\mathbf{P}}_{[0,\tau]}} \|\hat{b}_t(X_{k\eta}) - \hat{b}_t(X_t)\|^2. \end{aligned} \tag{31}$$

By the semigroup property, we first notice that

$$\mathcal{H}_{1-k\eta}[e^{\hat{g}/\varepsilon}\nu_n] = \mathcal{H}_{t-k\eta}[\mathcal{H}_{1-t}[e^{\hat{g}/\varepsilon}\nu_n]].$$

We can verbatim apply Lemma 16 of [Chen et al. \(2022\)](#) with $\mathbf{q} := \mathcal{H}_{1-t}[e^{\hat{g}/\varepsilon}\nu_n]$, $\mathbf{M}_0 = \text{id}$ and $\mathbf{M}_1 = (t - k\eta)I$, since $\mathcal{H}_{1-k\eta}[e^{\hat{g}/\varepsilon}\nu_n] = \mathbf{q} * \mathcal{N}(0, (t - k\eta)I)$. This gives

$$\begin{aligned} \|\hat{b}_{k\eta}(X_{k\eta}) - \hat{b}_t(X_{k\eta})\|^2 &= \left\| \varepsilon \nabla \log \frac{\mathbf{q} * \mathcal{N}(0, (t - k\eta)I)}{\mathbf{q}}(X_{k\eta}) \right\|^2 \\ &\lesssim L_t^2 \eta d + L_t^2 \eta^2 \|\varepsilon \nabla \log \mathbf{q}(X_{k\eta})\|^2. \end{aligned}$$

Since $\varepsilon \log \mathbf{q}$ is L_t -smooth, we obtain the bounds

$$\begin{aligned} \mathbb{E}_{\tilde{\mathbf{P}}_{[0,\tau]}} \|\varepsilon \nabla \log \mathbf{q}(X_{k\eta})\|^2 &\lesssim \mathbb{E}_{\tilde{\mathbf{P}}_{[0,\tau]}} \|\varepsilon \nabla \log \mathbf{q}(X_t)\|^2 + L_t^2 \|X_t - X_{k\eta}\|^2 \\ &\leq \varepsilon L_t d + L_t^2 \mathbb{E}_{\tilde{\mathbf{P}}_{[0,\tau]}} \|X_t - X_{k\eta}\|^2. \end{aligned}$$

where the final inequality is a standard smoothness inequality (see Lemma C.2). Similarly, the second term on the right-hand side of (31) can be bounded by

$$\mathbb{E}_{\tilde{\mathbf{P}}_{[0,\tau]}} \|\hat{b}_t(X_{k\eta}) - \hat{b}_t(X_t)\|^2 \leq L_t^2 \mathbb{E}_{\tilde{\mathbf{P}}_{[0,\tau]}} \|X_{k\eta} - X_t\|^2.$$

Combining the terms, we obtain

$$\mathbb{E}_{\tilde{\mathbf{P}}_{[0,\tau]}} \|\hat{b}_{k\eta}(X_{k\eta}) - \hat{b}_t(X_t)\|^2 \lesssim \varepsilon L_t^2 \eta d + L_t^2 \mathbb{E}_{\tilde{\mathbf{P}}_{[0,\tau]}} \|X_{k\eta} - X_t\|^2,$$

where, to simplify, we use the fact that $\eta \leq 1/L_t$ (with $L_t \geq 1$), and that $\eta^2 \leq \eta$ for $\eta \in [0, 1]$. We now bound the remaining expectation. Under $\tilde{\mathbf{P}}_{[0,\tau]}$, we can write

$$X_t = \int_0^t \hat{b}_s(X_s) ds + \sqrt{\varepsilon} B_t, \quad X_{k\eta} = \int_0^{k\eta} \hat{b}_s(X_s) ds + \sqrt{\varepsilon} B_{k\eta},$$

and thus

$$X_t - X_{k\eta} = \int_{k\eta}^t \hat{b}_s(X_s) \, ds + \sqrt{\varepsilon}(B_t - B_{k\eta}).$$

Taking squared expectations, writing $\delta := t - k\eta \leq \eta$ (recall that $t \in [k\eta, (k+1)\eta)$), we obtain (through an application of the triangle inequality and Jensen's inequality)

$$\begin{aligned} \mathbb{E}_{\tilde{\mathbf{p}}_{[0,\tau]}} \|X_t - X_{k\eta}\|^2 &\lesssim \varepsilon \mathbb{E}_{\tilde{\mathbf{p}}_{[0,\tau]}} \|B_t - B_{k\eta}\|^2 + \delta \int_{k\eta}^t \mathbb{E}_{\tilde{\mathbf{p}}_{[0,\tau]}} \|\hat{b}_s(X_s)\|^2 \, ds \\ &\lesssim \varepsilon \eta d + \delta^2 L_t d \\ &\leq (\varepsilon + 1) \eta d \end{aligned}$$

where we again used Lemma C.2. Combining all like terms, we obtain the final result.

The estimates for the Lipschitz constant follow from Lemma C.1. $\square$

B.3 Proofs for Section 4.3

B.3.1 Computing Equation 27

The Föllmer drift is a special case of the Schrödinger bridge, where $\mu = \delta_a$ for any $a \in \mathbb{R}^d$. Let (f^F, g^F) denote the optimal entropic potentials in this setting. Note that these potentials are defined up to translation (i.e., the solution is the same if we take $f^F + c$ and $g^F - c$ for any $c \in \mathbb{R}$). So, we further impose the condition that $f^F(a) = 0 = c$. Then the optimality conditions yield

$$g^F(y) = \frac{1}{2\varepsilon} \|y\|^2. \quad (32)$$

Plugging this into the expression for the Schrödinger bridge drift, we obtain

$$\begin{aligned} b_t^F(z) &= \varepsilon \nabla \log \mathcal{H}_{(1-t)\varepsilon} [e^{\frac{1}{2\varepsilon} \|\cdot\|^2} \nu](z) \\ &= (1-t)^{-1} \left(-z + \frac{\int y e^{\frac{1}{2\varepsilon} \|y\|^2 - \frac{1}{2(1-t)\varepsilon} \|z-y\|^2} \nu(dy)}{\int e^{\frac{1}{2\varepsilon} \|y\|^2 - \frac{1}{2(1-t)\varepsilon} \|z-y\|^2} \nu(dy)} \right). \end{aligned}$$

Replacing the integrals with respect to ν with their empirical counterparts yields the estimator.

B.3.2 Proof of Proposition 4.6

Our goal is to prove the following lemma.

Lemma B.2. *Let $\mathbf{p}_\tau$ be the Föllmer bridge at time $\tau \in [0, 1)$ between $\mu = \delta_0$ and $\nu \in \mathcal{P}_2(\mathbb{R}^d)$ with $\varepsilon = 1$ and suppose the squared second moment of ν is bounded above by d . Then*

$$W_2^2(\mathbf{p}_\tau, \nu) \leq d(1 - \tau).$$

Proof. Note that $\mathbf{p}_\tau = \mathbf{P}_{1-\tau}$, where $\mathbf{P}_{1-\tau}$ is the *reverse* bridge, which starts at ν and ends at $\mu = \delta_0$. This reverse bridge is well known to satisfy a simple SDE [Föllmer \(1985\)](#): the measure $\mathbf{P}_{1-\tau}$ is the law of $Y_{1-\tau}$, where Y_s solves

$$dY_s = -\frac{Y_s}{1-s}ds + dB_s, \quad Y_0 \sim \nu,$$

which has the explicit solution

$$Y_s = (1-s)Y_0 + (1-s) \int_0^s \frac{1}{1-r} dB_r.$$

In particular, we obtain

$$\begin{aligned} W_2^2(\mathbf{P}_s, \nu) &\leq \mathbb{E} \|Y_s - Y_0\|^2 \\ &= \mathbb{E} \left\| -sY_0 + (1-s) \int_0^s \frac{1}{1-r} dB_r \right\|^2 \\ &= s^2 \mathbb{E} \|Y_0\|^2 + ds(1-s) \\ &\leq ds, \end{aligned}$$

which proves the claim. □

C Technical lemmas

Lemma C.1 (Hessian calculation and bounds). *Let $(\mathbf{p}_t, b_t)$ be the optimal density-drift pair satisfying the Fokker–Planck equation (11) between μ_0 and μ_1 . For $t \in [0, 1)$, b_t is Lipschitz with constant L_t given by*

$$L_t := \sup_x \|\nabla b_t(x)\|_{\text{op}} \leq \frac{1}{(1-t)} \left(1 \vee \|\nabla^2 \varphi_{1-t}(x)\|_{\text{op}} \right),$$

where $\nabla \varphi_{1-t}$ is the entropic Brenier map between $\mathbf{p}_t$ and μ_1 with regularization parameter $(1-t)\varepsilon$. Moreover, if the support of μ_1 is contained in $B(0, R)$, then

$$L_t \leq (1-t)^{-1} (1 \vee R^2((1-t)\varepsilon)^{-1}). \quad (33)$$

Proof. Taking the Jacobian of b_t , we arrive at

$$\nabla b_t(x) = (1-t)^{-1} (\nabla^2 \varphi_{1-t}(x) - I),$$

As entropic Brenier potentials are convex (recall that their Hessians are covariance matrices; see (8)), we have the bounds

$$-(1-t)^{-1} I \preceq \nabla b_t(x) \preceq (1-t)^{-1} \nabla^2 \varphi_{1-t}(x).$$

The first claim follows by considering the larger of the two operator norms of both sides.

The second claim follows from the fact that since φ_{1-t} is an optimal entropic Brenier potential, its Hessian is the conditional covariance of an optimal entropic coupling $\pi_t \in \Gamma(\mathbf{p}_t, \mu_1)$, so

$$\|\nabla^2 \varphi_{1-t}(z)\|_{\text{op}} = \frac{1}{(1-t)\varepsilon} \|\text{Cov}_{\pi_t}[Y|X_t = z]\|_{\text{op}} \leq \frac{R^2}{(1-t)\varepsilon},$$

since $\text{supp}(\mu_1) \subseteq B(0, R)$. $\square$

Lemma C.2. *Let $(\mathbf{p}_t, b_t)$ be the optimal density-drift pair satisfying the Fokker–Planck equation (11) between μ_0 and μ_1 . Then for any $t \in [0, 1]$*

$$\mathbb{E}_{\mathbf{p}_t} \|b_t\|^2 \leq \frac{\varepsilon}{2} L_t d.$$

Proof. This proof follows the ideas of [Vempala and Wibisono \(2019, Lemma 9\)](#). We note that the generator given by the forward Schrödinger bridge with volatility ε is

$$\mathcal{L}_t f = \frac{\varepsilon}{2} \Delta f - \langle b_t, \nabla f \rangle,$$

for a smooth function f . Writing $b_t = \nabla(\varepsilon \log \mathcal{H}_{1-t}[e^{g/\varepsilon} \mu_1])$, we obtain

$$0 = \mathbb{E}_{\mathbf{p}_t} \mathcal{L}_t(\varepsilon \log \mathcal{H}_{1-t}[e^{g/\varepsilon} \mu_1]) \implies \mathbb{E}_{\mathbf{p}_t} \|b_t(X_t)\|^2 = \frac{\varepsilon}{2} \mathbb{E}_{\mathbf{p}_t} [\nabla \cdot b_t] \leq \frac{\varepsilon}{2} L_t d.$$

$\square$

Lemma C.3. ([Stromme, 2023b, Proposition 3.1](#)) *Let P, Q be probability measures on $\mathbb{R}^d$. For every pair $h_1 = (f_1, g_1) \in L^\infty(P) \times L^\infty(Q)$, there exists an element of $L^\infty(P) \times L^\infty(Q)$ which we denote by $\nabla \Phi_\varepsilon^{PQ}(f_1, g_1)$ such that for all $h_0 = (f_0, g_0) \in L^\infty(P) \times L^\infty(Q)$,*

$$\begin{aligned} \langle \nabla \Phi_\varepsilon^{PQ}(h_1), h_0 \rangle_{L^2(P) \times L^2(Q)} &= \int f_0(x) \left(1 - \int e^{-\varepsilon^{-1}(c(x,y) - f_1(x) - g_1(y))} dQ(y) \right) dP(x) \\ &\quad + \int g_0(y) \left(1 - \int e^{-\varepsilon^{-1}(c(x,y) - f_1(x) - g_1(y))} dP(x) \right) dQ(y). \end{aligned}$$

In other words, the gradient of Φ_ε^{PQ} at (f_1, g_1) is the marginal error corresponding to (f_1, g_1) .

Lemma C.4. *Following Lemma C.3, suppose $P = \mu$ and $Q = \nu_n$, where ν_n is the empirical measure of some measure ν on the basis of n i.i.d. samples. Let (f, g) be the optimal entropic potentials between μ and ν , which induce an optimal entropic coupling π (recall (6)). Then*

$$\mathbb{E} \|\nabla \Phi^{\mu\nu_n}(f, g)\|_{L^2(\mu) \times L^2(\nu_n)}^2 \lesssim \frac{\|\gamma\|_{L^2(\mu \otimes \nu)}^2}{n},$$

where the expectation is with respect to the data, and $\gamma = \frac{d\pi}{d(\mu \otimes \nu)}$.

Proof. Writing out the squared-norm of the gradient explicitly in the norm $L^2(\mu) \times L^2(\nu_n)$, we obtain

$$\begin{aligned} \mathbb{E} \|\nabla \Phi^{\mu\nu_n}(f, g)\|_{L^2(\mu) \times L^2(\nu_n)}^2 &= \mathbb{E} \int \left(\frac{1}{n} \sum_{j=1}^n \gamma(x, Y_j) - 1 \right)^2 \mu(dx) \\ &\quad + \mathbb{E} \frac{1}{n} \sum_{j=1}^n \left(\int \gamma(x, Y_j) \mu(dx) - 1 \right)^2. \end{aligned}$$

Note that by the optimality conditions, $\int \gamma(x, Y_j) \mu(dx) = 1$ for all Y_j . Thus, writing $Z_j := \gamma(x, Y_j)$ which are i.i.d., we see that

$$\begin{aligned} \mathbb{E} \int \left(\frac{1}{n} \sum_{j=1}^n \gamma(x, Y_j) - 1 \right)^2 \mu(dx) &= \int \mathbb{E} \left(\frac{1}{n} \sum_{j=1}^n (Z_j - \mathbb{E}[Z_j]) \right)^2 \\ &= \text{Var}_{\mu \otimes \nu} \left(\frac{1}{n} \sum_{j=1}^n Z_j \right) \\ &= \frac{1}{n} \text{Var}_{\mu \otimes \nu}(Z_1). \end{aligned}$$

The remaining component of the squared gradient vanishes, and we obtain

$$\mathbb{E} \|\nabla \Phi^{\mu\nu_n}(f, g)\|_{L^2(\mu) \times L^2(\nu_n)}^2 = \frac{1}{n} \text{Var}_{\mu \otimes \nu}(\gamma) \leq \frac{\|\gamma\|_{L^2(\mu \otimes \nu)}^2}{n}.$$

□